Vorwort

Die Gesellschaft für Informatik e.V. (GI) vergibt gemeinsam mit der Schweizer Informatik Gesellschaft (SI) und der Österreichischen Computergesellschaft (OCG) jährlich einen Preis für eine hervorragende Dissertation im Bereich der Informatik. Deutsche, österreichische und schweizer Universitäten und Hochschulen schlagen jeweils eine ausgezeichnet bewertete Dissertation vor, die zur Weiterentwicklung im Bereich Informatik und / oder in den Anwendungsgebieten beiträgt oder die Wechselwirkung zwischen Informatik und Gesellschaft untersucht. Somit sind die im Auswahlverfahren vorgeschlagenen Kandidatinnen und Kandidaten bereits „Preisträger“ ihrer Hochschule.

Die 36 Einreichungen für das Jahr 2020 waren die höchste Zahl an Anreichungen, die es bisher gab. Sie belegen die Bedeutung des Dissertationspreises. Leider konnten wir in diesem Jahr aufgrund der Corona-Pandemie das Kolloquium zum Dissertationspreis erneut nicht im Leibniz-Zentrum für Informatik Schloss Dagstuhl durchführen. Vielmehr haben wir online-Vorträge der Nominierten organisiert, die verteilt über mehrere Tage abliefen. Das wissenschaftliche Niveau der Vorträge war sehr hoch, die sich daran anschließenden Diskussionen aber leider doch zeitlich sehr beschränkt.

Wie in jedem Jahr fiel es dem Nominierungsausschuss sehr schwer, eine Dissertation auszuwählen, die durch den Preis besonders gewürdigt wird. Mit der Präsentation aller vorgeschlagenen Arbeiten in diesem Band wird die Ungerechtigkeit, eine aus mehreren ebenbürtigen Dissertationen hervorzuheben, etwas ausgeglichen. Der Band soll zudem einen Beitrag zum Wissenstransfer innerhalb der Informatik und von den Universitäten und Hochschulen in die Bereiche Technik, Wirtschaft und Gesellschaft leisten.

Die genannten Gesellschaften zeichnen Frau Dr. Daniela Kaufmann für ihre Dissertation „Formal Verification of Multiplier Circuits using Computer Algebra“ mit dem Dissertationspreis 2020 aus. Frau Dr. Kaufmann hat aktuelle Verifikationsmethoden basierend auf Computeralgebra verbessert und neue Methoden entwickelt, die für einen gegebenen Integer-Multiplizierer auf Gatterebene vollautomatisch über dessen Korrektheit entscheiden, ohne dass der Entwickler manuell in den Verifikationsprozess eingreifen muss.

Mit dieser Preisverleihung wird eine herausragende Arbeit gewürdigt, die eine überraschende Lösung für ein lange bekanntes Problem im Spektrum Theorie, Software und industrielle Anwendung liefert.

Ein großer Dank gilt dem Nominierungsausschuss für die sehr effiziente und konstruktive Arbeit. Desweiteren möchte ich mich besonders bedanken bei Frau Sylvia Wünsch für die Organisation der online-Vorträge, Frau Lena Reinfelder und Herrn Stefan Sobernig für die Zusammenstellung des Bandes und dem Team der Gesellschaft für Informatik e.V. für die technische Unterstützung des Auswahlverfahrens.

Steffen Hölldobler
Dresden im November 2021

Kandidat*innen für den GI-Dissertationspreis 2020

Dr. Alperovich, Anna	Universität Konstanz
Dr. Barz, Björn	Friedrich-Schiller-Universität Jena
Dr. Bauer, Andre	Universität Würzburg
Dr. Bieshaar, Maarten	Universität Kassel
Dr. rer. tech. Borkowski, Michael	Deutsches Zentrum für Luft- und Raumfahrt Braunschweig
Dr.-Ing. Chrszon, Philipp	Technische Universität Dresden
Dr.-Ing. Dockhorn, Alexander	Otto-von-Guericke-Universität Magdeburg
Dr. George, Ceenu	Ludwig-Maximilians-Universität München
Dr.-Ing. Gnad, Dennis	Karlsruher Institut für Technologie
Dr. Grotherr, Christian	Universität Hamburg
Dr. Hassan, Teena	Otto-Friedrich Universität Bamberg
Dr. Herdt, Vladimir	Universität Bremen
Dr. rer. nat. Iosifidis, Vasileios	Leibniz Universität Hannover
Dr. Juhnke, Katharina	Universität Ulm
Dr. Jung, Ralf	Universität des Saarlandes
Dr. Junges, Sebastian	RWTH Aachen
Dr.-Ing. Kaufmann, Daniela	Johannis Kepler Universität Linz
Dr. Keldenich, Phillip	Technische Universität Braunschweig
Dr. Kragl, Bernhard	Institute of Science and Technology Austria
Dr. Krüger, Stefan	Universität Paderborn
Dr. rer. nat. Laina, Iro	Technische Universität München
Dr. Marky, Karola	Technische Universität Darmstadt
Dr. Mescheder, Lars	Universität Tübingen
Dr. Neumann, Stefan	Universität Wien
Dr. Paul, Erik	Universität Leipzig
Dr. Piovarci, Michal	Universita della Svizzera italiana
Dr. Reuter, Britta	Zeppelin Universität Friedrichshafen
Dr. Riedelbauch, Dominik	Universität Bayreuth
Dr.-Ing. Scherr, Franz	Technische Universität Graz
Dr. Schwammberger, Maike	Carl von Ossietzky Universität Oldenburg
Dr.-Ing. Then, Matthias	Fern Universität Hagen
Dr. Van der Zander, Benito	Universität zu Lübeck
Dr. Wägemann, Peter	Friedrich-Alexander-Universität Erlangen-Nürnberg
Dr. rer. nat. Walzer, Stefan	Universität zu Köln
Dr. Wang, Xi	Technische Universität Berlin
Dr. Zieris, Franz	Freie Universität Berlin

Mitglieder des Nominierungsausschusses für den GI-Dissertationspreis 2020

Prof. Dr. Steffen Hölldobler (Vorsitzender)	Technische Universität Dresden
Prof. Dr. Sven Apel	Universität des Saarlandes
Prof. Dr. Abraham Bernstein	Universität Zürich
Prof. Dr.-Ing. Felix Freiling	Universität Erlangen-Nürnberg
Prof. Dr. Hans-Peter Lenhof	Universität des Saarlandes
Prof. Dr. Gustaf Neumann	Wirtschaftsuniversität Wien
Prof. Dr. Rüdiger Reischuk	Universität zu Lübeck
Prof. Dr. Kay Uwe Römer	TU Graz
Prof. Dr. Björn Scheuermann	Humbold-Universität zu Berlin
Prof. Dr. Nicole Schweikardt	Humbold-Universität zu Berlin
Prof. Dr. Myra Spiliopoulou	Otto-von-Guericke-Universität Magdeburg
Prof. Dr. Sabine Süsstrunk	École Polytechnique Fédérale de Lausanne
Prof. Dr. Klaus Wehrle	RWTH Aachen

Inhaltsverzeichnis

Variations- und Deep-Learning-Ansätze bei der intrinsischen Zerlegung von Lichtfeldern [1]

Anna Alperovich[2]

Abstract: Bei der intrinsischen Bildzerlegung geht es darum, ein beleuchtungsinvariantes Reflexionsbild von einem Eingangsfarbbild zu trennen, was nach wie vor noch eines der grundlegenden Probleme im Bereich der Computer Vision darstellt. Diese Zerlegungsart wird häufig bei der Bearbeitung von Fotos und Materialien, der Bildsegmentierung sowie der Formschätzung eingesetzt. Im Fokus dieser Arbeit liegt die intrinsische 4D-Zerlegung eines Lichtfelds. Im Rahmen dessen soll das Problem in Bezug auf die folgenden drei Variablen formuliert und gelöst werden soll: Albedo, Schattierung und Spekularität. Dadurch wird es wiederum möglich, sich mit nicht-Lambertschen Szenen auseinanderzusetzen. Dem Problem soll sich mit Variations- und Deep-Learning-Ansätzen angenähert, ihre Leistung verglichen und die Stärken und Schwächen beider Techniken diskutiert werden. Es soll nachgewiesen werden, dass der in dieser Arbeit vorgestellte Deep-Learning-Ansatz eine generische Lösung für Lichtfelder darstellt und bei vier zeitgenössischen Computer-Vision-Problemstellungen eingesetzt werden kann: Disparitätsschätzung, Reflexionstrennung, intrinsische Bildgebung und bildverarbeitende Ultrahochauflösung. Umfangreiche Auswertungen auf der Grundlage mehrerer öffentlich zugänglicher, synthetischer und realer Datensätze belegen die Fruchtbarkeit der im Rahmen dieser Arbeit vorgestellten Methodik. Im Ergebnis werden die Vorteile der Verwendung von Lichtfeldern gegenüber anderen Datenstrukturen aufgezeigt.

1 Einführung

Der Fokus liegt dabei auf inversen Problemen, bei denen bestimmte Komponenten einer Szene aufgrund ihrer fotografischen Bilder wiederhergestellt werden sollen. Es werden zwei Ansätze untersucht: Bei dem ersten Ansatz wird die Modellierung des Problems auf der Grundlage physikalischer Prinzipien der Bilderzeugung untersucht. Bei dem zweiten Ansatz kommt Deep Learning zum Einsatz.

Die vorliegende Arbeit unterscheidet sich in vielerlei Hinsicht von vergleichbaren Forschungsvorhaben auf diesem Gebiet. Erstens wird in dieser Untersuchung anstelle eines Einzelbildes einer Szene eine Serie von Bildern, die aus einer etwas anderen Perspektive (Lichtfelder, siehe Abb. 1) aufgenommen wurden, als Eingabe für die verwendeten Algorithmen verwendet. In früheren Untersuchungen wurde aufgezeigt, dass vom Lichtfeld übernommene umfangreiche Informationen verwendet werden können, um die Geometrie zuverlässig abzuschätzen und verschiedene physikalische Eigenschaften einer Szene wiederherstellen zu können. Zweitens liegt der Fokus in dieser Arbeit auf den nicht-Lambertschen Phänomenen, die dank der Lichtfelddaten von der diffusen Reflexion getrennt betrachtet werden können.

[1] Englischer Titel der Dissertation: "Variational and Deep Learning Approaches for Intrinsic Light Field Decomposition"

[2] Universität Konstanz, ann.alperovich@gmail.com

Diese Arbeit besteht aus zwei Teilen, wobei im ersten Teil Variationsmethoden für intrinsische Lichtfelder diskutiert werden, bei denen das Eingangslichtfeld in drei Rendering-Komponenten zerlegt wird: Albedo, Schattierung und Spekularität. Um diese Komponenten zu modellieren, ist die exakte Geometrie einer Szene erforderlich. Zunächst jedoch soll in dieser Untersuchung die Methode zur Disparitätsschätzung vorgestellt werden, wodurch nicht ausschließlich Disparitätsmarker wiederhergestellt, sondern auch stückweise glatte Normalen-Maps ausgegeben werden können. Im Ergebnis kann die 3D-Darstellung einer Szene berechnet und zur Modellierung der intrinsischen Komponenten verwendet werden. Der größte Nachteil dieser Methode besteht darin, dass zur Berechnung der Disparität davon ausgegangen wird, dass die Szene rein Lambertsch ist, was in der Realität so gut wie nie der Fall ist. Somit kann der vorgeschlagene Algorithmus keine genauen Disparitätsmarker in spiegelnden Bereichen wiederherstellen. Durch die Regularisierung der Normalen kann dieses Problem teilweise für kleine hervorgehobene Bereiche gelöst werden, die von nahezu diffusen Bereichen umgeben sind, aber das Problem bleibt für große spiegelnde Oberflächen bestehen.

Im zweiten Teil wird ein Deep-Learning-Ansatz beschrieben, bei dem die Möglichkeit untersucht wird, ein physikalisches Modell zu beschreiben, indem das neuronale Netz ausschließlich mit Trainingsbeispielen gespeist wird. Hierzu wurde ein 3D-Convolutional-Neural-Network (CNN) entwickelt, um die winkelabhängigen Daten im Lichtfeld zu nutzen und den Mehrwert dieses Netzes im Zusammenhang mit den zeitgenössischen Computer-Vision-Problemen aufzuzeigen. Da im Rahmen dieser Untersuchung das Netz auf die nicht-Lambertschen Daten trainiert wird, wird es dazu gezwungen, spiegelnde Bereiche zu verarbeiten und als Ergebnis die Einschränkung des Modellierungsansatzes zu überwinden. In zahlreichen Versuchen wird die Leistung beider Ansätze verglichen und der Vorteil der

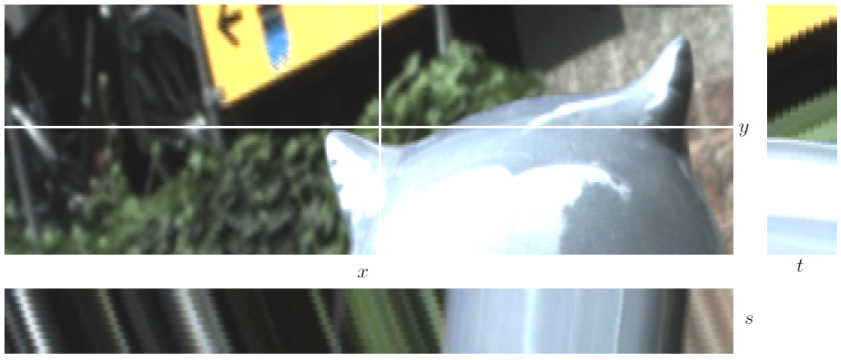


Abb. 1: In der Abbildung oben links ist die Mittelansicht eines Lichtfelds zu sehen, das durch die Bildkoordinaten x und y parametrisiert ist. Unten und rechts sind die Abbildungen der Epipolarebenen (EPIs) für die weißen Linien in der Mittelansicht zu sehen, wobei s und t die Koordinaten der Ansichtspunkte beschreiben. Während sich die Kamera bewegt, zeichnen 3D-Szenenpunkte gerade Linien auf den EPIs, deren Neigung der Disparität entspricht. Jede Zuweisung einer Eigenschaft eines Szenenpunktes zu den entsprechenden Strahlen sollte entlang dieser Linien konstant sein, was zur kohärenten Regulierung genutzt werden kann.

Verwendung eines Deep Networks für die Disparitätsschätzung und die intrinsische Bildgebung veranschaulicht.

Dadurch soll aufgezeigt werden, dass mit dem Lichtfeld als Eingabe sowohl Variations- als auch Deep-Learning-Methoden den Einzelbildmethoden und Bild + tiefen Methoden überlegen sind. Insbesondere soll der Vorteil der Verwendung von Lichtfeldern für spiegelnde Szenen aufgezeigt werden, bei denen die meisten Algorithmen aufgrund der Lambertschen Annahme versagen. Dies würde ein breites Spektrum von Anwendungen und Impulsen für zukünftige Forschungen zum Verständnis nicht-Lambertscher Szenen mit Lichtfeldern eröffnen.

2 Variationsmethoden

2.1 Disparitätsschätzung

Eine häufige Anwendung des Lichtfelds besteht darin, die Tiefe der aufgenommenen Szene zu schätzen, um darüber hinaus ihre 3D-Struktur rekonstruieren zu können. Aktuelle Algorithmen zur Disparitätsschätzung funktionieren außerordentlich gut, wenn die Tiefe der Lichtfeldbilder geschätzt wird. Diese Methoden sind jedoch normalerweise nicht auf die Normalenschätzung ausgelegt. Daher sind die Tiefenschätzungen von Algorithmen, die auf Kostenvolumina basieren, selbst wenn sie mit Sublabel-Genauigkeit [Mo16] optimiert wurden, häufig stückweise flach und können daher keine genauen Normalen-Maps bestimmen. Häufig ist ihre Genauigkeit naturgemäß auch um die Okklusionsgrenzen herum begrenzt. Das Ziel des in dieser Arbeit beschriebenen Disparitätsschätzungsalgorithmus besteht darin, zu einer Beseitigung dieser Nachteile beizutragen, siehe Abb. 2 für eine Beispielausgabe der vorgeschlagenen Methode.

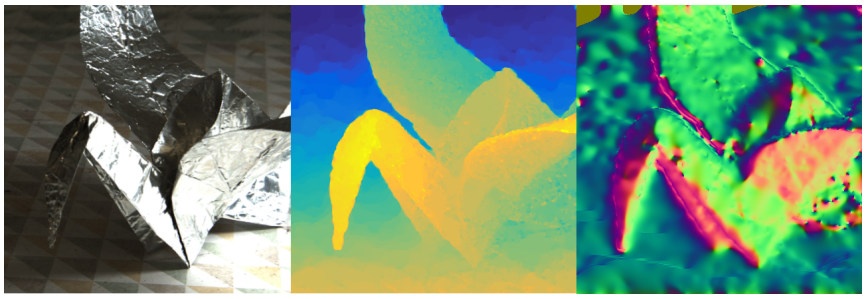


Abb. 2: Es soll ein neuartiger Ansatz zur Berechnung von Disparität-Kostenvolumina, der auf dem Konzept der okklusionsbewussten Fokalstapelsymmetrie basiert, vorgestellt. Mit dem vorgeschlagenen Konzept können die Tiefe als auch die Normalen gemeinsam optimiert werden, um herausfordernde reale Szenen rekonstruieren zu können, die mit einer plenoptischen Kamera (Lytro Illum) aufgenommen werden. In der Abbildung links ist die Mittelansicht des Lichtfelds dargestellt, die obere rechte Abbildung zeigt die Disparitätsmarker und in der unteren rechten Abbildung ist die Normalen-Map zu sehen

Zunächst soll eine neuartige Methode vorgestellt werden, wie mit Okklusionen umgegangen werden kann, wenn Kostenvolumina auf der Idee der Fokalstapelsymmetrie beruhen [Li15]. Mit diesem neuartigen Datenanteil können wesentlich genauere Ergebnisse als mit der vorherigen Methode [Li15] erzielt werden, wenn ein globales Optimum mithilfe der Sublabel-genauen Relaxierung berechnet wird [Mo16]. Zweitens soll eine Nachbearbeitung unter Verwendung einer gemeinsamen Regulierung von Tiefe und Normalen vorgeschlagen werden, um eine glatte Normalen-Map zu erhalten, die mit der Tiefenschätzung übereinstimmt. Hierfür sollen die Ansätze nach Graber et al. [Gr15] zur linearen Verbindung von der Tiefe und den Normalen sowie das Konzept der Relaxation nach Zeisl et al. [ZZP14] Anwendung finden, um sich der Nichtkonvexität der Einheitslängenbeschränkung auf der Normalen-Map anzunähern. Die daraus resultierenden Teilprobleme in Bezug auf Regulierung der Tiefe und den Normalen können mit dem Primal-Dual-Algorithmus effizient gelöst werden. Zum Zeitpunkt der Einreichung dieser Arbeit konnten wesentlich bessere Ergebnisse im Vergleich zu früheren Abhandlungen erzielt werden, in denen neuere Benchmark-Verfahren zur Disparitätsschätzung von Lichtfeldern in Bezug auf die Genauigkeit von Disparitätenkarten und Normal-Maps und mehrere andere Metriken behandelt wurden.

2.2 Intrinsische Lichtfelder

Es soll ein neuartiges Variationsmodell für die intrinsische Zerlegung eines Lichtfelds vorgestellt werden, das anstatt auf herkömmlichen Bildern auf dem 4D-Strahenraum definiert wird. Hierzu wird ein intrinsisches Lichtfeld als Funktion modelliert $L(\boldsymbol{r}) = A(\boldsymbol{r})S(\boldsymbol{r}) +$

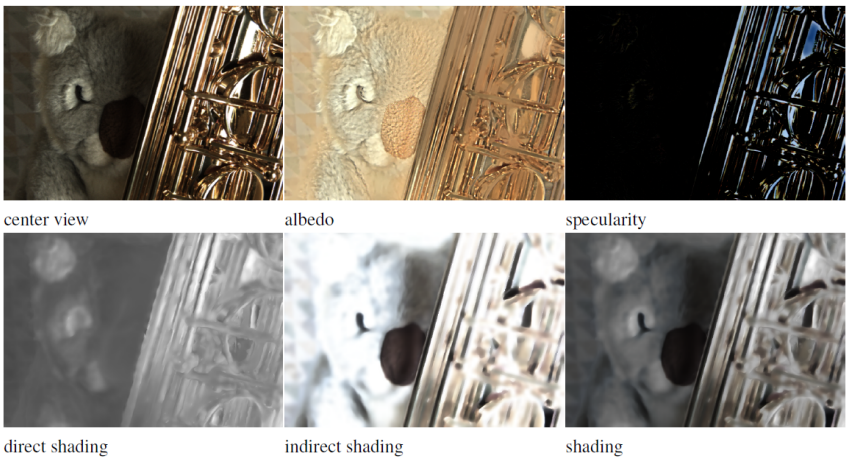


Abb. 3: Ein variationsbasierter Ansatz zur Zerlegung der Lichtfeldstrahlung in die intrinsischen Komponenten Albedo, Schattierung und Spekularität. Durch die weitere Aufteilung der Schattierung in direkte und indirekte Anteile wird im Vergleich zu früheren Arbeiten eine herausragende Leistung erzielt.

$H(\boldsymbol{r})$, Wo die Strahlung L jedes Strahls $\boldsymbol{r}$ in Albedo A, Schattierung S und Glanzfarben-Komponente H zerlegt wird. Die Funktionen $L,A,S,H : \mathscr{R} \rightarrow \mathbb{R}^3$ ordnen dem Strahlenraum entsprechende RGB-Farbwerte zu. Albedo steht für die Farbe eines Objekts, unabhängig von der Beleuchtung und der Kameraposition. Die Schattierung beschreibt Intensitätsänderungen aufgrund von Beleuchtung, Interreflexionen und der Objektgeometrie. Schließlich stellt die Spekularitätskomponente Glanzfarben dar, die an nicht-Lambertschen Oberflächen auftreten, was von der Beleuchtung, der Objektgeometrie sowie der Kameraposition abhängt.

Da die meisten vorhandenen intrinsischen Bildalgorithmen für Lambertsche Szenen entwickelt wurden, wird ihre Leistung dann beeinträchtigt, wenn Szenen betrachtet werden, die glänzende Oberflächen aufweisen. Im Gegensatz dazu wird es durch die reichhaltige Struktur des Lichtfelds mit vielen dicht abgetasteten Ansichten möglich, sich nicht-Lambertschen Oberflächen zu widmen, indem ein zusätzlicher Zerlegungsterm eingeführt wird, der die Spekularität modelliert. Die Regulierung entlang der Bilder in der Epipolarebene fördert die Albedo- und Schattierungsbeständigkeit in allen Ansichten.

Das ursprüngliche Modell soll erweitert werden, indem an diesem gewisse Neurungen vorgenommen werden, bevor sich der Schattenbildung und Zwischenreflexionen gewidmet wird. Somit besteht das neue Zerlegungsmodell $L(\boldsymbol{r}) = A(\boldsymbol{r})s_d(\boldsymbol{r})S_i(\boldsymbol{r}) + H(\boldsymbol{r})$ aus den Komponenten s_d und S_i, die die direkte und indirekte Schattierung beschreiben, siehe Abb. 3. Unter direkter Schattierung kann die Schattierung verstanden werden, die ein Objekt hätte, wenn es sich allein in einer Szene befinden würde, d. h. ohne das Vorhandensein von Schatten oder reflektiertem Licht. Durch die zweite Komponente S_i werden Reflexionen und Schattenbildungen modelliert.

Im Gegensatz zum ersten Ansatz, bei dem die Interreflexion nur auf der Grundlage der Geometrie modelliert wird, wird in dieser Abhandlung die indirekte Schattierung mittels einer Kombination aus Geometrie- und Farbinformationen modelliert. Es wird ein Konfidenzmaß zur Beschattung für das Lichtfeld berechnet, welches zur Regulierung herangezogen wird. In umfangreichen Experimenten konnte nachgewiesen werden, dass das hier ausgewählte Forschungsdesign die Schätzung der Schattierungskomponente signifikant verbessert. Eine weitere Verbesserung besteht in dem Hinzufügen einer Frequenzanalyse bei der Spekularitätsschätzung, die auf der Annahme basiert, dass hohe Variationen der Pixelintensitäten eines 3D-Punkts in Ansichten mit Subapertur durch Spekularität verursacht werden. Die neuen Schwerpunkte münden somit in einen neueren Ansatz, um das Eingangslichtfeld in die Komponenten Albedo, Schattierung und Spekularität zerlegen zu können.

3 Deep-CNN für Lichtfelder

Lichtfelder weisen eine komplexe, stark redundante Struktur auf, siehe Abb. 1 Für Szenen mit rein diffuser Reflexion weisen EPIs Muster orientierter Linien mit konstanter Farbe auf. Die Linien entsprechen der Projektion eines einzelnen 3D-Punkts im Raum, und ihre Neigung oder Disparität ist umgekehrt proportional zur Entfernung des Punkts zum Beob-

achter. Diskontinuitäten im Muster werden durch Okklusionen verursacht, die Übergänge zwischen mehreren Orientierungspunkten an der Okklusionskante erzeugen.

Die Situation wird zudem weniger eindeutig, wenn Reflexion oder glänzende, nicht - Lambertsche Oberflächen ins Spiel kommen, weil die EPIs dann überlagerte Muster aufweisen [JSG16]. Die Ausrichtung der Muster, die der Spiegelreflexion entsprechen, entspricht nicht der Disparität, sondern der Bewegungsrichtung der Spekularität, die von der intrinsischen Oberflächengeometrie abhängt. Um zwischen diesen beiden Fällen zu unterscheiden, ist zu eruieren, ob ein Punkt eine diffuse oder spiegelnde Reflexion aufweist. Mit einer bekannten Geometrie kann die Bewegungsrichtung der Spekularität direkt geschätzt und Reflexionskomponenten voneinander getrennt werden [Su16]. Wenn sowohl Form als auch Reflexionsvermögen unbekannt sind, ist es kaum möglich zu bestimmen, welche Phänomene zu einem bestimmten EPI geführt haben.

Dennoch weisen EPIs aus natürlichen Lichtfeldern eine insgesamt regelmäßige Struktur auf. Zudem ist es wahrscheinlich, dass sie im gesamten Bildraum der Epipolarebene eine vergleichsweise niedrigdimensionale Mannigfaltigkeit bilden. Darüber hinaus hängt eine entsprechende Codierung eines EPIs mit nur wenigen Parametern mit den komplexen miteinander verzahnten Aufgaben zusammen, wie z. B. der Disparitätsschätzung oder der Trennung von Albedo-, Schattierungs- und Spekularitätkomponenten. Die Vermutung liegt nahe, dass Sie die anderen Aufgaben besser meistern können, wenn Sie eruieren, wie man eine Komprimierung erfolgreich ausführt. Gegenstand dieses Teils der Arbeit ist es daher, eine niedrigdimensionale Darstellung von EPIs aus beliebigen Beispiellichtfeldern zu

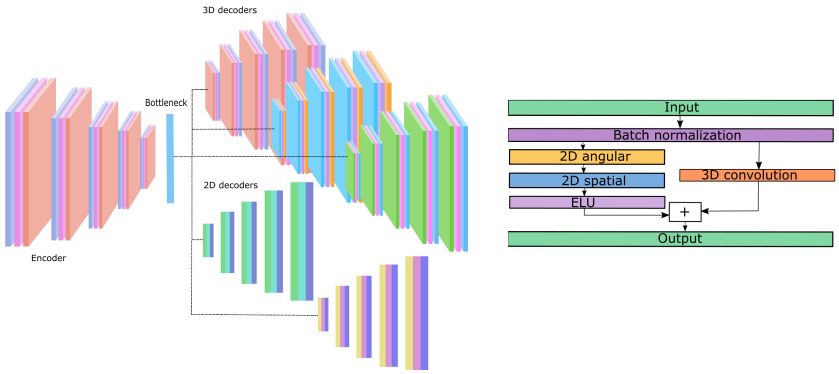


Abb. 4: **Links**: Übersicht über die vorgeschlagene Architektur. Abhängig von der verfügbaren Aufgabe und dem verfügbaren GPU-Speicher können Anzahl und Inhalt der Gruppen geändert werden. Die Decoderpfade stellen exakte Spiegelbilder dieser Kette dar. Die Abmessung der Ansichtspunkte der Form in den 2D-Decodern wird entfernt. Von dem Netzwerk wird eine beliebige Anzahl von 2D- und 3D-Decodern unterstützt. **Recht**: Beispiel eines residualer Block des Netzes. Nach der Batch-Normalisierung wird der Ausgangstensor über zwei Wege übergeben. Der rechte behält den Eingangstensor bei oder führt eine 3D-Faltungsberechnung durch, wenn er erneut abgetastet werden muss. Die linke führt winkelige und räumliche 2D- Faltungsberechnung durch, gefolgt von der ELU-Schicht. Der Ausgangstensor ergibt sich aus der Summe dieser beiden Pfade.

entwerfen, jedoch in einer Art und Weise, dass die latenten Variablen gemeinsam verwendet werden können, um verschiedene angeleitete Aufgaben in der Lichtfeldanalyse genau lösen zu können. Zu diesem Zweck wird in dieser Abhandlung ein neuronales Encoder-Decoder-Netz vorgeschlagen, das auf dem Konzept des Deep Autoencoders [HS06] basiert, welches in jüngster Zeit sehr erfolgreich aussagekräftige mannigfaltige Darstellungen erzeugen konnte.

Es soll die erste Netzwerkarchitektur für das Multitasking-Lernen in Lichtfeldern vorgestellt werden, siehe Abb. 4. Der Grundgedanke ist es, das Netzwerk um einen Autoencoder herum aufzubauen, damit es ohne Anleitung nur mit rohen Lichtfelddaten trainiert werden kann. Hierbei werden jedoch mehrere Pfade hinzugefügt, um die latente Darstellung zu dekodieren, die gemeinsam mit dem Autoencoder unter Anleitung trainiert werden können, je nachdem, welche Daten im aktuellen Trainingsbeispiel verfügbar sind. Durch die Kombination von angeleitetem und unbeaufsichtigtem Training kann sichergestellt werden, dass die latente Darstellung nicht ausschließlich auf die gewünschten Aufgaben wie Tiefenrekonstruktion oder die intrinsische Komponentendarstellung angewendet werden kann, sondern sich auch gut auf Datensätze übertragen lässt, bei denen für die auszuführenden Aufgaben keine Trainingsinformationen verfügbar sind. Sobald das Netz bereitgestellt wird, können alle Decoderketten nur anhand der Lichtfelddaten ausgewertet werden. In dieser Abhandlung liegt der Fokus auf den vier zeitgenössischen Computer-Vision-Problemen, Disparitätsschätzung, intrinsische Bildgebung, Reflexionstrennung und bildverarbeitende Ultrahochauflösung.

CNN zur Reflexionstrennung und Disparitätsschätzung. Gemäß dem dichromatischen Reflexionsmodell verfügt das von einem Szenenpunkt reflektierte Licht über zwei unabhängige Komponenten: Licht, das vom Oberflächenkörper und an der Grenzfläche reflektiert wird. Die Objektreflexion ist als diffuse Komponente bekannt und unabhängig von der Blickrichtung, während es sich bei der Grenzflächenreflexion um Folgendes handelt

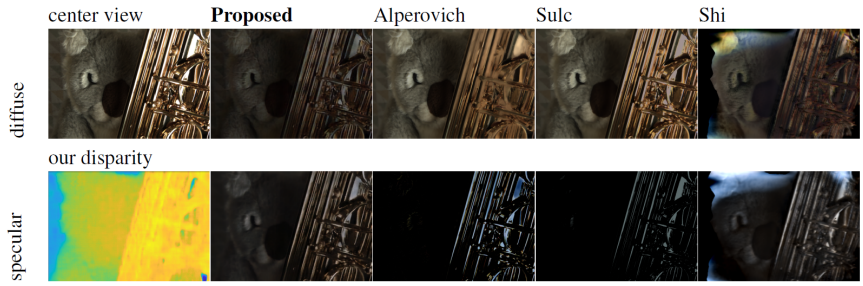


Abb. 5: Mit der plenoptischen Kamera (Lytro Illum) aufgenommene Lichtfelder. Die Szene besteht aus einem hochglänzenden Saxofon und einem fast lambertianischen Koala. Das in dieser Arbeit vorgestellte Netz erkennt im Vergleich zu anderen Methoden mit großem Erfolg mehr spiegelnde Oberflächen des Saxofons. Während der Koala als ein spiegelndes Objekt ähnlich zu [Sh17] fehlinterpretiert wird, ist die hier vorgestellte Methode die einzige, bei der der diffuse Teil hinter dem großen spiegelnden Fleck auf dem Saxofon nicht unscharf dargestellt wird.

die spiegelnde Komponente, die ansichtsabhängig ist. Bei der Trennung der spiegelnden und diffusen Reflexionsanteile handelt es sich um ein inverses Problem, das innerhalb der Computer-Vision-Community nach wie vor ein reges Forschungsgebiet darstellt. Unter Bezugnahme der Arbeit von Sulc et al. [Su16] wird dargestellt, dass das Auftreten von Spekularitäten in einem Lichtfeld von der Szenengeometrie abhängt und entlang der Strömungsrichtungen der Spekularität konstant bleibt, während die Lambertsche Komponente lediglich in der Disparitätsrichtung konstant bleibt. Daher bietet es sich an, eine im Lichtfeld codierte geometrische Information zu verwenden, um eine Reflexionstrennung durchzuführen. Es wird die in Abb. 4 beschriebene vorgeschlagene Architektur übernommen, um gemeinsam Aufgaben in den Bereichen Disparitätsschätzung und Reflexionstrennung auszuzuführen. In der Abb. 5 werden die im Rahmen dieser Untersuchung erzielten Ergebnisse und Vergleiche zum konkreten Beispiel dargestellt.

Intrinsische Zerlegung eines Lichtfelds auf der Grundlage von Deep-CNN. Im Gegensatz zum vorherigen Netz führen wird im Rahmen dieser Arbeit eine vollständige intrinsische Zerlegung durchgeführt, eine Architektur entworfen, mit der es ermöglicht wird, doppelt größere Patches als Eingabe zu verarbeiten und Sprungverbindungen vom Codierer zu entsprechenden Decoderteilen eingeführt, durch die die Rekonstruktionsqualität von Decodern verbessert wird. 3D-Falten werden durch eine 2D-Sequenz ersetzt, die auf die Orts- und Winkelverteilung einwirken. Eine schematische Beschreibung des Restblocks finden Sie in Abb. 4. Diese Entwurfsauswahl verringert die Anzahl der Parameter im Netz und beschleunigt den Trainingsprozess. Wie im vorherigen Fall kann dieses Netz Lichtfelder verarbeiten, für die keine Ground-Truth-Informationen verfügbar sind.

Generative Adversarial Networks für hochauflösende Lichtfelder. Schließlich soll aufgezeigt werden, dass die vorgeschlagene Encoder-Decoder-Architektur (Abb. 4) für eine völlig andere Aufgabe verwendet werden kann. Mit einigen notwendigen Modifikationen und gemäß den jüngsten Forschungstrends soll ein 3D-Generative-Adverserial-

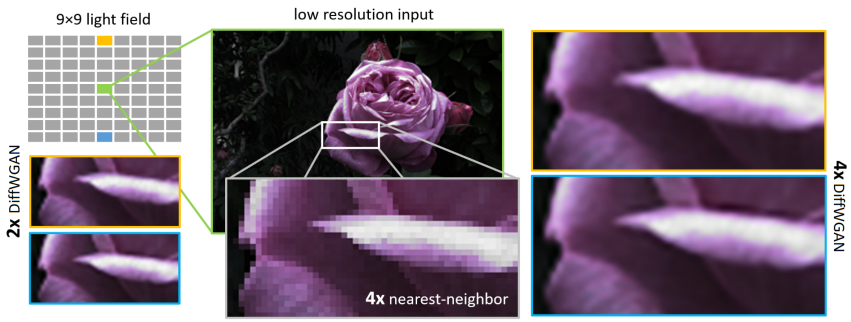


Abb. 6: Ausgabe des ultrahochauflösenden Netzwerks. Bei einer relativ bescheidenen Menge von Eingabeansichten mit Subapertur erzeugt das Netz dennoch ein ultrahochauflösendes Lichtfeld mit den Vergrößerungsfaktoren x2 und x4. Es wurde das Interpolations-Verfahren für das computererzeugte Bild mittels des Nächster-Nachbar-Interpolations-Verfahrens durchgeführt und das Ergebnis für den visuellen Vergleich präsentiert.

Autoencoder-Network vorgestellt werden, um das hochauflösende Lichtfeld aus einem niedrigauflösenden Lichtfeld mit einer geringen Anzahl von Ansichtspunkten wiederherzustellen.

Es werden lediglich drei Ansichten entlang der horizontalen und vertikalen Achse benötigt, um die Winkelauflösung um den Faktor drei zu erhöhen, während gleichzeitig die räumliche Auflösung in jeder Richtung um den Faktor zwei oder vier erhöht wird. In der Abb. 6 sind die Ein- und Ausgänge des Netzes dargestellt.

4 Zusammenfassung

In dieser Arbeit [Al20] werden verschiedene Forschungsfragen innerhalb der Computer Vision, insbesondere im Bereich der Lichtfelder, diskutiert, die allesamt eng miteinander verzahnt sind. Zunächst wäre da die Disparitätsschätzung, bei der es sich um ein bewährtes Verfahren im Bereich von Lichtfeldern handelt. Es wurden zahlreiche Algorithmen vorgeschlagen, um Disparitätsmarker zu berechnen, aber nur wenige von ihnen bewerten ihre Qualität für einige Anwendungen, bei denen geometrische Informationen benötigt werden. Es wurde ein Disparitätsschätzungsalgorithmus entworfen, der sich besonders zur Modellierung intrinsischer Komponenten eignet, indem Disparitätsmarker und Normalen-Maps gemeinsam optimiert werden. Die generierte Disparitätskarte wurde dazu verwendet, um eine 3D-Darstellung einer Szene zu rekonstruieren, die wiederum als Grundlage für die Modellierung intrinsischer Komponenten dient.

Im Rahmen dieser Abhandlung wurde ein variierender intrinsischer Zerlegungsmodus für Lichtfelder vorgeschlagen. Um die Lichtfeldstruktur also nutzbar zu machen, wurden die Komponenten Albedo und Schattierung dazu gebracht, entlang der Disparitätsrichtungen konstant zu bleiben. Um spiegelnde Oberflächen zu identifizieren, wird auf die Glanzfarbeigenschaften zurückgegriffen, um ansichtsabhängig operieren zu können. Es wurden die Projektionen eines 3D-Punkts in Subapertur- Ansichten analysiert und die Beschaffenheit der Reflexion ausgewertet. Dabei wurde das Variationsmodell optimiert, indem eine bessere Variante für die Schattierungskomponente eingeführt wurde, die sich durch eine feinere Aufgliederung in direkte und indirekte Anteile auszeichnet. Die Vorteile des neuen Modells wurden in einer Vielzahl von natürlichen und synthetischen Lichtfeldern veranschaulicht.

Es wurde eine neuronale Netzarchitektur entwickelt, die speziell für Lichtfelder konzipiert ist. Um die reichhaltige, aber redundante Lichtfeldstruktur nutzbar zu machen, wurde ein Encoder-Decoder-Netz entworfen, in dem ein kleiner Satz von Funktionen aus den Winkel- und Raumdimensionen extrahiert wurde und anschließend Verwenden Sie diese Funktionen, um verschiedene Aufgaben gleichzeitig zu auszuführen. Unter der Annahme, dass diese Aufgaben eng miteinander verbunden sind, werden die extrahierten Funktionen in separate Decodierungspfade unterteilt. Da es komplex ist, eine intrinsische Komponente mit Ground-Truth-Informationen oder Disparitätswerte für die realen Szenen zu erhalten, wurde das Netz in dieser Arbeit so konzipiert, dass es sowohl angeleitet als auch unbeaufsichtigt trainiert werden kann. Wenn Ground-Truth-Informationen verfügbar sind (für

synthetische Szenen), bestraft die Standardverlustfunktion die Abweichung zwischen der Ausgabe des Netzes und dem Ground-Truth-Beispiel. Für die natürlichen Szenen ohne Ground-Truth-Informationen stellt das Netz sicher, dass die Ausgabekomponenten dem zugrunde liegenden physikalischen Modell der Bilderzeugung entsprechen.

Literaturverzeichnis

[Al20] Alperovich, Anna: Variational and Deep Learning Approaches for Intrinsic Light Field Decomposition. 2020.

[Gr15] Graber, G.; Balzer, J.; Soatto, S.; Pock, T.: Efficient Minimal-Surface Regularization of Perspective Depth Maps in Variational Stereo. In: Proc. CVPR. 2015.

[HS06] Hinton, G.; Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

[JSG16] Johannsen, O.; Sulc, A.; Goldluecke, B.: What sparse light field coding reveals about scene structure. In: Proc. CVPR. 2016.

[Li15] Lin, H.; Chen, C.; Kang, S.-B.; Yu, J.: Depth Recovery from Light Field Using Focal Stack Symmetry. In: Proc. ICCV. 2015.

[Mo16] Moellenhoff, T.; Laude, E.; Moeller, M.; Lellmann, J.; Cremers, D.: Sublabel-Accurate Relaxation of Nonconvex Energies. In: Proc. CVPR. 2016.

[Sh17] Shi, J.; Dong, Y.; Su, H.; Yu, S.: Learning Non-Lambertian Object Intrinsics across ShapeNet Categories. In: Proc. CVPR. 2017.

[Su16] Sulc, A.; Alperovich, A.; Marniok, N.; Goldluecke, B.: Reflection Separation in Light Fields based on Sparse Coding and Specular Flow. In: VMV. 2016.

[ZZP14] Zeisl, B.; Zach, C.; Pollefeys, M.: Variational Regularization and Fusion of Surface Normal Maps. In: Proc. 3DV. 2014.

Anna Alperovich wurde am 17. Juni 1989 in Nowosibirsk, Russland, geboren. Sie bekam einen Bachelor- und einen Masterabschluss von der der Staatlichen Universität Nowosibirsk, wo sie Angewandte Mathematik studierte. 2012 verteidigte sie ihre Masterarbeit zum Thema der inversen Probleme. Anschließend wechselte sie an die Technische Universität Eindhoven in den Niederlanden, wo sie das Postgraduiertenprogramm Mathematics for Industry (= Mathematik in der Industrie) absolvierte. 2015 promovierte sie dort in Ingenieurwissenschaften (PDEng). Anschließend zog sie nach Deutschland, Konstanz, wo sie in Computer Vision promovierte. Sie genoss ihren Aufenthalt in Konstanz; in der Nähe der wunderschönen Berge und einer grandiosen Aussicht auf den Bodensee. Gegenwärtig arbeitet sie als Forschungsingenieurin für maschinelles Lernen bei der Carl Zeiss AG, wo sie erfolgreich eine Deep-Learning-Lösung für verschiedene industrielle Problemstellungen entwickelt.

Semantische und Interaktive Inhaltsbasierte Bildersuche[1]

Björn Barz[2]

Abstract: Methoden für die inhaltsbasierte Suche nach Bildern anhand eines Beispielbildes haben in jüngster Zeit rasante Fortschritte gemacht, konzentrieren sich jedoch größtenteils auf die visuelle Ähnlichkeit von Bildern und lassen deren Semantik außer Acht. Die Dissertation stellt eine Methode vor, welche menschliches Vorwissen über die Semantik der Welt in Form von Taxonomien in Deep-Learning-Verfahren integriert. Die daraus entstehenden semantischen Bildmerkmale verbessern die semantische Konsistenz der Suchergebnisse im Vergleich zu herkömmlichen Repräsentationen und Merkmalen erheblich. Darüber hinaus werden drei interaktive Suchverfahren präsentiert, welche die den Anfragebildern inhärente semantische Ambiguität durch Einbezug von Benutzerfeedback auflösen. Die verschiedenen Methoden decken eine große Bandbreite hinsichtlich der Komplexität des Feedbacks und des damit für den Benutzer verbundenen Aufwands ab. Alle Techniken liefern bereits nach wenigen Feedbackrunden deutlich relevantere Ergebnisse, was die Gesamtmenge der abgerufenen Bilder reduziert, die der Benutzer betrachten muss.

1 Einführung: Content-based Image Retrieval

Die Art und Weise, in der wir Informationen über die Welt mit anderen teilen, hat in den letzten Jahrzehnten radikale Veränderungen durchlaufen. Mit der Allgegenwärtigkeit von Kameras in Smartphones und der zunehmenden Verfügbarkeit mobiler Internetverbindungen haben Bilder als Medium für den Informationsaustausch deutlich an Stellenwert gewonnen. Eine Fotografie kann innerhalb von Sekunden angefertigt und geteilt werden und speichert doch sämtliche visuellen Informationen, auf die auch der Fotograf Zugriff hat. Neben rein faktischen Informationen wie Ereignissen, Aktivitäten und den Beziehungen zwischen den abgebildeten Objekten erfassen Bilder auch Stimmungen und Emotionen, die sich einer textuellen Beschreibung häufig entziehen.

Das Internet bietet eine Fülle an solchen in Bildern kodierten Informationen, die sowohl für alltägliche als auch wissenschaftliche Zwecke von großem Nutzen sein können. Der Überfluss an Bildern ist jedoch Segen und Fluch zugleich: Die Suche nach den wirklich relevanten und hilfreichen Bildern wird durch deren schiere Zahl zunehmend erschwert und macht daher den Einsatz ausgeklügelter Suchverfahren unausweichlich.

Die textbasierte Suche nach Bildern, die von den meisten Suchmaschinenanbietern noch immer verfolgt wird, ist dabei nicht zielführend, sondern eher dem Umstand erwachsen, dass dabei bereits vorhandene Indizes von Webseiten für die Suche nach in der Nähe von Bildern platzierten Schlüsselwörtern ohne großen Zusatzaufwand wiederverwendet werden können. Allerdings entbehren viele Bilder im Internet einer umfassenden begleitenden

[1] Englischer Titel der Dissertation: “Semantic and Interactive Content-based Image Retrieval”

[2] Lehrstuhl für Digitale Bildverarbeitung, Friedrich-Schiller-Universität Jena, bjoern.barz@uni-jena.de

Beschreibung und sind dadurch mit einem solchen Ansatz nicht auffindbar. Darüber hinaus weisen Bilder üblicherweise eine Vielzahl möglicher Interpretationen auf, die sich aus unterschiedlichen Perspektiven hinsichtlich des semantischen Inhalts des Bildes, der vom Bild transportierten Stimmung und ausgelösten Emotionen sowie dem künstlerischen Stil ergeben. Diese Bedeutungsvielfalt ginge bei einer Versprachlichung verloren.

Idealerweise muss die maschinelle Bildersuche also den tatsächlichen Inhalt des Bildes betrachten, ohne sich dabei auf etwaige Begleittexte zu verlassen. Auch die Anfrage der Suche besteht dabei nicht aus einer Menge von Schlüsselwörtern, sondern Beispielbildern. Der Nutzer wird dadurch von der Bürde entbunden, das mentale Bild seines Suchinteresses in Sprache zu überführen, was durchaus eine nennenswerte Herausforderung darstellen kann, etwa bei der Suche nach Gemälden in einem bestimmten künstlerischen Stil. Ein Beispielbild stellt daher eine sowohl aussagekräftige als auch kompakte Form der Anfrageformulierung dar.

Diese inhaltsbasierte Herangehensweise an die Bildersuche ist als *Content-based Image Retrieval (CBIR)* bekannt und seit den Neunzigerjahren aktiver Gegenstand der Forschung [Sm00]. Insbesondere in den letzten zehn Jahren hat die CBIR-Forschung große Fortschritte gemacht [BD21]. Der Schwerpunkt lag dabei größtenteils auf der Suche nach Bildern, die exakt dasselbe Objekt zeigen wie das Anfragebild, z. B. ein bestimmtes Gebäude aus einer anderen Perspektive. Der Vorteil dieser spezifischen Aufgabenstellung liegt darin, dass die Bilder keinerlei Mehrdeutigkeit aufweisen, wenn das Suchkriterium derart beschränkt wird. Ferner ist ein tieferes Verständnis des Bildinhalts zur Lösung dieser Aufgabe nicht notwendig, da ein Abgleich visueller Merkmale wie charakteristischer Formen oder Texturen häufig genügt.

In praktischen Nutzungsszenarien sind diese beiden Vereinfachungen jedoch in der Regel nicht haltbar, da Nutzer mit vollkommen verschiedenen Suchinteressen und Kriterien an das Suchsystem herantreten können. Die hier vorgestellte Dissertation [Ba20] präsentiert neuartige Methoden zur Lösung beider Probleme. Anstatt Bilder allein anhand ihrer oberflächlichen Erscheinung zu vergleichen, stellen wir eine Methode zur Integration menschlichen Weltwissens in Deep-Learning-Verfahren vor, wodurch das System die Semantik von Bildern besser zu verstehen lernt. Zur Auflösung der Ambiguität des Anfragebildes werden zudem drei interaktive Verfahren vorgestellt, welche den Benutzer kontinuierlich in den Suchprozess involvieren und die Möglichkeit zur Verfeinerung der Ergebnisse durch gezieltes Feedback ermöglichen.

2 Semantische Bildersuche

2.1 Visuelle und Semantische Ähnlichkeit

Moderne CBIR-Systeme bewerten die Ähnlichkeit von Bildern üblicherweise anhand von Bildrepräsentationen, die von einem künstlichen neuronalen Netzwerk erzeugt werden. Diese Netzwerke wurden entweder ursprünglich für eine Klassifikationsaufgabe trainiert [BL15] oder direkt für die inhaltsbasierte Bildersuche auf speziellen Datensätzen optimiert

[Br20], die eine Menge von Bildern charakteristischer Objekte und Gebäude enthalten und so einen sehr beschränkten Anwendungsfall abdecken. In beiden Fällen kodieren die so gelernten Repräsentationen detaillierte visuelle Merkmale der Bilder wie typische Formen und Texturen. Der Vergleich erfolgt daher maßgeblich anhand der *visuellen* Ähnlichkeit.

Menschen betrachten Bilder jedoch üblicherweise nicht primär oder gar ausschließlich hinsichtlich ihrer visuellen Eigenschaften, sondern nehmen vor allem deren *Semantik* wahr. So können wir beispielsweise Bilder von einer Raupe und einem Schmetterling vor dem Hintergrund unseres Weltwissens als ähnlich betrachten, weil sich die Raupe eines Tages in einen Schmetterling verwandeln wird. Visuell jedoch haben beide nur wenig gemein.

Umgekehrt impliziert auch die visuelle Ähnlichkeit zweier Bilder keineswegs einen semantischen Zusammenhang. So schrieb z. B. ein großer Technologiekonzern von Weltrang 2015 negative Schlagzeilen, da dessen Fotoverwaltungs-App Bilder dunkelhäutiger Menschen allzu leicht mit der Bezeichnung ‘Gorilla’ versah [Do15]. Die dadurch hervorgerufene öffentliche Empörung lässt den Schluss zu, dass die semantische Ähnlichkeit für das menschliche Verständnis von Bildern Vorrang vor der visuellen Ähnlichkeit hat.

Um sinnvolle Ergebnisse liefern zu können, muss also auch die inhaltsbasierte Bildersuche die Semantik der Bilder berücksichtigen. Zusammenhänge wie beispielsweise die Verbindung zwischen Raupe und Schmetterling oder zwischen Mensch und Gorilla können jedoch nicht aus Bildern allein erschlossen werden, sondern erfordern entweder zusätzliche Informationsquellen oder die explizite Integration solch menschlichen Wissens.

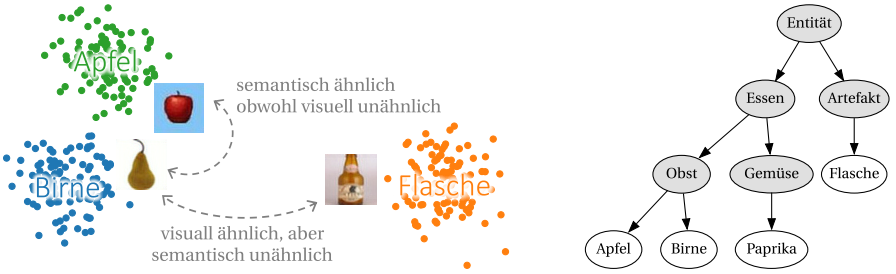
2.2 Taxonomien als Wissensquelle

Eine leicht erschließbare Quelle für solches Wissen über die Welt stellen Taxonomien dar. Sie organisieren Objektklassen in einer hierarchischen Struktur auf verschiedenen Abstraktionsebenen. So ist zum Beispiel ein Pudel nicht einfach nur ein Pudel, sondern auch ein Hund, ein Säugetier, ein Lebewesen, oder, noch allgemeiner, ein Organismus. Ein einfaches und reduziertes Beispiel für eine solche Taxonomie ist in Abb. 1b dargestellt.

Aus einer solchen Ordnung lässt sich nun die semantische Ähnlichkeit zweier beliebiger Klassen ablesen. Eine Birne ist zum Beispiel einem Apfel ähnlicher als eine Paprika, da Apfel und Birne beide Obst sind, ein gemeinsamer Vorfahre von Apfel und Paprika aber erst in der Klasse ‘Essen’ zu finden ist. ‘Essen’ aber ist generischer als ‘Obst’, was sich daran zeigt, dass diese Klasse in der Taxonomie ein tiefer verzweigtes Geflecht von Unterkonzepten, sogenannten *Hyponymen*, unter sich vereint.

Formal bezeichnen wir diesen nächsten gemeinsamen Vorfahren zweier Knoten $u, v \in V$ im die Taxonomie repräsentierenden Graphen als $\mathrm{lcs}(u,v)$. Die Höhe $\mathrm{height}(u)$ eines im Knoten u gründenden Teilgraphen der Taxonomie wird festgelegt durch die maximale Länge eines Pfads von u zu einem Blattknoten. Die semantische Ähnlichkeit $\mathrm{sim}_{\mathrm{sem}}$ zweier beliebiger Klassen auf Grundlage der Taxonomie definieren wir dann wie folgt:

$$\mathrm{sim}_{\mathrm{sem}}(u,v) = 1 - \frac{\mathrm{height}(\mathrm{lcs}(u,v))}{\max_{r \in V} \mathrm{height}(r)}\,.$$



(a) Merkmalsraum (b) Beispielhafte Taxonomie

Abb. 1: In einem semantischen Merkmalsraum sollten Bilder von semantisch ähnlichen Objekten näher beieinander liegen, obwohl sie sich nicht zwangsläufig visuell ähneln. Ein Maß für die semantische Ähnlichkeit von Klassen kann z. B. aus Taxonomien abgeleitet werden, welche menschliches Wissen über die Beziehungen zwischen Objektklassen kodieren.

2.3 Semantische Repräsentationen für Klassen und Bilder

Zum Zweck der semantischen Bildersuche benötigen wir einen Merkmalsraum, in dem der Abstand zwischen den Bildrepräsentationen nicht vornehmlich durch deren visuelle, sondern die semantische Ähnlichkeit bestimmt wird (siehe Abb. 1a). Zum Vergleich von Merkmalsvektoren hat sich im Umfeld von CBIR die Kosinus-Ähnlichkeit etabliert, da sie insbesondere in hochdimensionalen Räumen der euklidischen Distanz überlegen ist. Seien $x_i, x_j \in \mathbb{R}^D$ die L^2-normalisierten Merkmalsvektoren zweier Bilder aus den Klassen c_i, c_j. Die Anforderung, dass sich die Ähnlichkeit der Merkmale nach der semantischen Ähnlichkeit der Klassen richtet, lautet dann formal: $x_i^\top x_j = \mathrm{sim}_{\mathrm{sem}}(c_i, c_j)$.

Beschränken wir uns dabei zunächst auf einen Mittelpunkt oder Prototypen z_i für alle Beispiele einer bestimmten Klasse c_i, erlaubt uns das anhand der Taxonomie definierte semantische Ähnlichkeitsmaß die explizite Berechnung der Positionen dieser Klassenmittelpunkte, sodass obiges Kriterium erfüllt ist. Dazu sei $S \in \mathbb{R}^{C\times C}$ eine Matrix mit den semantischen Ähnlichkeiten zwischen allen Paaren der C bekannten Klassen und $\Gamma^\top \Lambda \Gamma = S$ die Eigenwertzerlegung von S. Dann gilt:

$$\mathrm{sim}_{\mathrm{sem}}(c_i, c_j) = S_{ij} = (\sqrt{\Lambda_{ii}}\Gamma_i)^\top (\sqrt{\Lambda_{jj}}\Gamma_j) = z_i^\top z_j \,.$$

Dabei bezeichnet Γ_i die i-te Spalte von Γ. Die Positionen der Klassenzentren im Merkmalsraum ergeben sich folglich unmittelbar aus der Eigenwertzerlegung von S.

Um nun zusätzlich zu den Klassen ebenfalls die Bilder in diesen semantischen Merkmalsraum einzubetten, kommt ein künstliches neuronales Netzwerk zum Einsatz. Dieses wird unter dem Kriterium trainiert, die Kosinus-Ähnlichkeit zwischen der gelernten Repräsentation jedes einzelnen Bildes und dem zuvor berechneten und fixierten Zentrum der jeweiligen Klasse zu maximieren. Die Kenntnis der Klassenzugehörigkeiten ist dabei nur für das Training des Netzwerkes erforderlich. Im Praxisbetrieb kann es dann genutzt werden, um allein aus den Anfrage- und Datenbankbildern zugehörige Repräsentationen zu erzeugen und durch einen Vergleich im Merkmalsraum diejenigen Bilder zu ermitteln, die gemäß ihrer Semantik der Anfrage am ähnlichsten sind.

Abb. 2: Ergebnisse der inhaltsbasierten Bildersuche mit klassischen Merkmalen (oben) und unseren hierarchiebasierten semantischen Merkmalen (unten). Das Bild links dient als Anfrage. Die Farbe der Rahmen kodiert die semantische Ähnlichkeit zwischen der Klasse der zurückgelieferten Bilder und des Anfragebildes (dunkelgrün = sehr ähnlich, dunkelrot = überhaupt nicht ähnlich).

2.4 Ergebnisse

Wir evaluierten den Nutzen dieser hierarchiebasierten semantischen Bildmerkmale im Kontext der Bildersuche auf zwei populären Datensätzen mit Bildern diversen Inhalts (*ImageNet* und *CIFAR-100*) sowie einem spezialisierten Datensatz, der Bilder von 555 verschiedenen Vogelarten enthält (*North American Birds*). Zur Bewertung der Qualität bestimmen wir für jedes zurückgelieferte Bild die semantische Ähnlichkeit seiner Klasse mit der Klasse des Anfragebildes und mitteln diese Werte über die ersten 250 Ergebnisse.

Es zeigte sich, dass unsere Methode die semantische Konsistenz der Suchergebnisse maßgeblich steigert. Im Vergleich zu existierenden Verfahren für das Lernen semantischer Bildrepräsentationen übertreffen wir die jeweils beste existierende Methode durchschnittlich um 20% bezüglich der besagten Metrik. Doch auch bei Verwendung traditioneller Bewertungskriterien wie der Average Precision, welche nur binäre Relevanzkriterien und keine Abstufung gemäß der semantischen Ähnlichkeit der Ergebnisse berücksichtigt, zeigt sich durch die Nutzung unserer semantischen Bildrepräsentationen eine Verbesserung.

Ein konkretes Beispiel ist in Abb. 2 dargestellt. Als Anfrage dient ein Bild eines Schimpansen, was bei der Verwendung herkömmlicher Merkmale häufig zu der unerwünschten Verwechselung mit Menschen führt. Dieser Fehler wird durch den Einbezug von Weltwissen eliminiert. Die Ergebnisse unserer Methode zeigen vorrangig andere Schimpansen, gefolgt von anderen Tieren. Menschen jedoch sind in diesem Fall in einem vollkommen anderen Zweig der Taxonomie verortet und werden daher nicht als ähnlich betrachtet.

Darüber hinaus stellten wir fest, dass die Integration von Wissen in Form von Hierarchien auch Vorteile für das Training künstlicher neuronaler Netzwerke zum Zweck der Klassifikation bietet, insbesondere auf kleinen Datensätzen. In der Praxis ist es in vielen Anwendungsgebieten illusorisch, Tausende von Beispielbildern pro Klasse vorliegen zu haben. In einem realistischeren Szenario von 20–100 Trainingsbeispielen pro Klasse erzielte unsere Methode eine um bis zu 30% höhere Klassifikationsgenauigkeit als das Standardverfahren.

Andere Wissenschaftler berichteten zudem, dass mit unserem Verfahren trainierte Klassifikatoren auch weniger gravierende Fehler begehen [Be20]. Statt zum Beispiel eine Eiche mit einem Menschen zu verwechseln, würde sie eher als ein anderer Baum klassifiziert.

3 Interaktive Bildersuche

Wie eingangs bereits erwähnt, kann ein Bild unter sehr unterschiedlichen Gesichtspunkten betrachtet und interpretiert werden. Im Allgemeinen ist es folglich unmöglich, die Intention des Nutzers anhand eines einzelnen Anfragebildes eindeutig zu bestimmen. Mechanismen zur gezielten Verfeinerung der Suchergebnisse sind daher nicht nur eine willkommene Zusatzfunktion, sondern obligatorisch.

3.1 Information-Theoretic Active Learning

Den am weitesten verbreiteten Ansatz für eine interaktive Gestaltung des Suchvorgangs stellt das sogenannte *Relevanzfeedback* dar. Der Benutzer erhält dabei die Möglichkeit, bestimmte Bilder in den Suchergebnissen als relevant oder irrelevant zu markieren, woraufhin das System unter Berücksichtigung dieser Präferenzen eine neue Suche durchführt und verfeinerte Ergebnisse zurückliefert. Dieser Vorgang lässt sich beliebig oft wiederholen. Allerdings können mitunter einige Iterationen erforderlich sein, bis der Benutzer mit den Ergebnissen zufrieden ist, da das System jeweils diejenigen Ergebnisse zuerst anzeigt, über deren Relevanz es sich bereits am sichersten ist. Dies beeinträchtigt den Informationsgehalt des Nutzerfeedbacks, da ein positives Feedback nur wenig Zusatzinformationen böte, ein negatives aber zu unspezifisch und wenig konstruktiv wäre.

Wir haben daher ein neuartiges Verfahren namens *Information-Theoretic Active Learning (ITAL)* entwickelt, welches aktiv Bilder aus dem Datensatz auswählt und den Nutzer um Rückmeldung zu diesen Bildern bittet. Die Auswahl erfolgt dabei anhand eines informationstheoretischen Kriteriums, um den durch das Feedback zu erwartenden Informationsgewinn bezüglich der Relevanz jedes Bildes im Datensatz zu maximieren. Auf diese Weise kann nach nur wenigen Runden nicht nur die Qualität, sondern auch die Diversität der Suchergebnisse deutlich verbessert werden.

Dazu modellieren wir die Relevanz aller Bilder in der Datenbank sowie das Nutzerfeedback als Zufallsvariablen und legen dem Nutzer jene Teilmenge an Bildern aus der Datenbank vor, die die sogenannte *Transinformation* zwischen der Relevanz aller Bilder und dem Feedback maximiert. Das heißt, der Informationsgewinn, den wir aus der Rückmeldung zu diesen Bildern über die Relevanz aller anderen Bilder erhalten, soll maximal sein.

Dieses Kriterium berücksichtigt nicht nur die zu erwartende Stärke der Modelländerung, die die Rückmeldung zur Folge hätte, sondern auch die Zuverlässigkeit des Nutzers, die in Form eines probabilistischen Nutzermodells explizit modelliert wird. Dadurch kann das Verhältnis zwischen Redundanz und Diversität der ausgewählten Bilder automatisch auf die Komplexität des Suchkriteriums sowie die Gründlichkeit des Nutzers bei der Begutachtung der Ergebnisse abgestimmt werden.

Jene Teilmenge des Datensatzes zu finden, die die exakte Transinformation maximiert, ist jedoch eine Aufgabe von exponentieller Komplexität in der Größe des Datensatzes. Wir verwenden daher diverse Approximationen, um dieses Problem traktabel zu gestalten. Details finden sich in Kapitel 6 der Dissertation [Ba20].

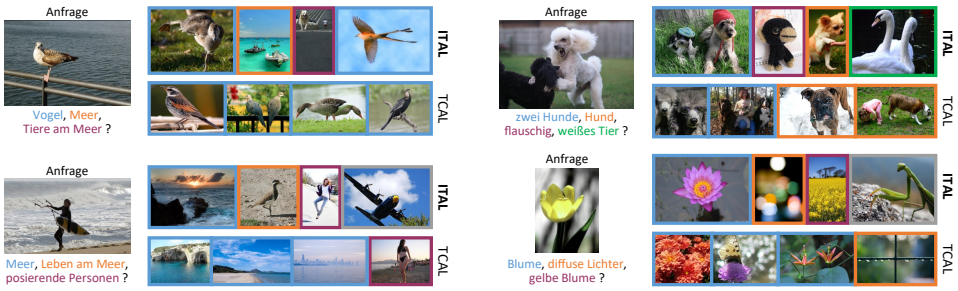


Abb. 3: Vergleich von Kandidaten für vier ausgewählte Anfragebilder, welche unsere Methode ITAL und der Wettbewerber TCAL dem Nutzer zwecks Rückmeldung vorlegen. Die Farbe der Rahmen entspricht verschiedenen Themen, die mit dem Anfragebild assoziiert werden können.

Im Vergleich mit zehn anderen Active-Learning-Verfahren auf fünf Datensätzen bewies ITAL über alle Datensätze hinweg entweder die beste oder zumindest eine vergleichbare Leistung, gemessen daran, wie schnell sich die Suchergebnisse verbessern. Die meisten anderen Methoden hingegen erbrachten nur auf manchen Datensätzen gute Ergebnisse.

Abb. 3 zeigt beispielhaft für vier ausgewählte Anfragebilder, welche Kandidaten ITAL im Vergleich zur nächstbesten Methode TCAL dem Nutzer in der ersten Feedbackrunde vorlegt. Es ist zu beobachten, dass die von ITAL ausgewählten Kandidaten eine höhere Diversität aufweisen und verschiedene mögliche Aspekte des Anfragebildes beleuchten. Dadurch kann der Anwender die Suche schneller in die gewünschte Richtung lenken. Nach weiteren Iterationen nimmt der Detailgrad der Unterscheidungen zwischen den verschiedenen Themen zu. Im Beispiel des Pudels werden im weiteren Verlauf etwa verschiedene Hunderassen und später dieselbe Hunderasse in verschiedenen Farben vorgeschlagen.

3.2 Automatische Disambiguierung des Anfragebildes

Active-Learning-Ansätze wie ITAL setzen einen kooperativen Nutzer voraus, der bereit ist, mehrere Feedback-Runden zu durchlaufen und somit für den Erhalt optimaler Ergebnisse ein wenig Zeit und Mühe zu investieren. Jedoch kann nicht in allen Situationen eine solch aktive Mitwirkung des Nutzers erwartet werden. Wir präsentieren daher auch eine alternative Herangehensweise, die den Aufwand des Benutzers zur Verfeinerung der Ergebnisse auf ein Minimum reduziert. Dazu wird das Bild automatisch disambiguiert und in seine verschiedenen Interpretationen zerlegt, welche durch Beispiele aus dem Datensatz repräsentiert werden. Der Benutzer muss dann lediglich eine dieser Interpretationen auswählen, um die Relevanz sämtlicher Bilder im Datensatz sofort neu bewerten zu lassen.

Um verschiedene Bildinterpretationen automatisch zu identifizieren, führen wir zunächst ein herkömmliches Retrieval durch, indem der Datensatz nach den 200 ähnlichsten Bildern durchsucht wird. Diese initialen Ergebnisse enthalten üblicherweise eine diverse Menge an Bildern, welche der Anfrage nicht alle hinsichtlich desselben, sondern ganz verschiedener Aspekte ähneln. Diese werden nun mithilfe des k-Means-Algorithmus in Gruppen unterteilt, welche verschiedene Facetten der Anfrage widerspiegeln.

Abb. 4: Ein beispielhaftes Anfragebild und die ersten Suchergebnisse, die unser Verfahren für jede der drei automatisch erkannten Bedeutungen der Anfrage zurückliefert.

Für jede so identifizierte Bedeutung werden nun eine Handvoll Bilder als Repräsentanten ausgewählt, welche der Anfrage am ähnlichsten sind. Auf diese Weise kann der Benutzer die verschiedenen Interpretationen, die von den Clustern repräsentiert werden, einfach und schnell visuell erfassen. Die Rückmeldung besteht dann lediglich aus einem einzelnen Klick auf diejenige Bedeutung, die im Interesse der Suche steht.

Existierende Methoden würden daraufhin schlicht alle Bilder anzeigen, die zum gewählten Cluster gehören. Dieses Vorgehen lässt jedoch außer Acht, dass nicht nur das Anfragebild, sondern auch die Ergebnisse in der Regel mehrdeutig sind. Bilder nahe der Grenze des ausgewählten Clusters könnten daher übersehen werden, obwohl sie ebenfalls relevant sind. Ferner beschränkt ein solches Vorgehen die Ergebnisse zwangsläufig auf eine Teilmenge der 200 Ergebnisse des initialen Retrievals. Durch die aus dem Nutzer-Feedback gewonnenen Informationen können nun jedoch potenziell auch andere Bilder aus der Datenbank als wesentlich ähnlicher zur Anfrage betrachtet werden als ursprünglich.

Unser Verfahren vermeidet daher binäre Entscheidungen und bewertet stattdessen sämtliche Bilder in der Datenbank neu. Dabei wird sowohl die Ähnlichkeit zum Anfragebild als auch die Übereinstimmung der Richtung im Merkmalsraum mit der Richtung des ausgewählten Clusters berücksichtigt. Anschaulich ausgedrückt werden Bilder, die im Merkmalsraum in derselben Richtung von der Anfrage wie der ausgewählte Cluster liegen, näher an das Anfragebild herangezogen, Bilder in der entgegengesetzten Richtung jedoch weggeschoben.

In Experimenten mit inhärent mehrdeutigen Bildern von der Foto-Plattform Flickr war unser Verfahren in der Lage, die *mean average precision* der Ergebnisse gegenüber dem initialen Ranking um 24% zu steigern. Existierende Ansätze hingegen konnten aufgrund oben genannter Beschränkungen nur die Qualität der ca. ersten 10 Ergebnisse in ähnlichem Maße steigern, jedoch nicht die der gesamten Ergebnisliste.

Ein qualitatives Beispiel ist in Abb. 4 dargestellt. Für die abgebildete Anfrage werden automatisch drei mögliche Interpretationen identifiziert. Für jede dieser Bedeutungen stellen wir hier die am höchsten bewerteten Ergebnisse dar, falls die jeweilige Bedeutung als die relevante markiert würde. Die Fokussierung auf verschiedene Aspekte des Anfragebildes ist deutlich erkennbar: Der erste Cluster erfasst die Bedeutung ‘Hände’, der zweite die Interpretation ‘Kleinkinder’ und der dritte offenbar den Bildtyp ‘Porträt’.

3.3 Adaptive Pooling

Abschließend stellt die Dissertation [Ba20] eine dritte Methode vor, welche eine andere Modalität des Feedbacks erschließt und dem Nutzer nicht nur erlaubt, bestimmte Bilder als relevant zu kennzeichnen, sondern Bereiche innerhalb dieser Bilder auszuwählen, welche für das Suchvorhaben von eminenter Bedeutung sind. Daraufhin wird die Merkmalsrepräsentation sämtlicher Bilder im Datensatz angepasst, um ähnlichen Merkmalen wie denen des ausgewählten Bildabschnitts ein höheres Gewicht zu verleihen.

Zu diesem Zweck lernen wir ein sogenanntes *Exemplar-LDA*-Modell auf den aus dem markierten Bereich extrahierten Merkmalen. Dieser Typ von linearen Modellen kann mithilfe von vorab berechneten Statistiken auf nur einem einzigen Positivbeispiel trainiert werden. Als lineares Modell lässt es sich zudem effizient auf sämtliche lokalen Merkmale der Bilder in der Datenbank anwenden, um deren Übereinstimmung mit der ausgewählten Region zu bewerten. Die lokalen Merkmale werden schließlich gemäß dieser Übereinstimmung gewichtet und in globale Bildrepräsentationen aggregiert.

Wir evaluierten den Nutzen dieser Methode auf spezialisierten Datensätzen mit Bildern von verschiedenen Spezies von Schmetterlingen und Vögeln. In diesen Fällen kann Expertenwissen hilfreich sein, um die charakteristischen Merkmale einer bestimmten Spezies im Anfragebild zu markieren und dadurch die Suche gezielt zu verfeinern. Im konkreten Fall verbesserte unser Ansatz die *mean average precision* gegenüber den nicht verfeinerten Ergebnissen um 224% und um 63% im Vergleich zu existierenden Verfahren.

4 Zusammenfassung

Das Forschungsgebiet des *Content-based Image Retrievals* hat sich über Jahre auf die inhaltsbasierte Suche nach Bildern eines bestimmten individuellen Objekts konzentriert. Für allgemeinere Anwendungsszenarien lassen aktuelle Systeme sowohl ein hinreichendes semantisches Verständnis von Bildern und der Welt sowie den nötigen Grad an Interaktivität zur Eingrenzung der Suche vermissen. Die hier beschriebene Dissertation [Ba20] stellt Lösungen für beide Problemfelder vor.

Hinsichtlich der interaktiven Bildersuche haben wir drei Verfahren entwickelt, die Nutzerfeedback verschiedener Art und Komplexität abdecken, von der schnellen automatischen Erkennung potenzieller Bildinterpretationen, über die aktive Akquise möglichst informativen Relevanzfeedbacks bis hin zur Auswahl wichtiger Bereiche in einzelnen Bildern. In allen Fällen übertreffen unsere Ansätze den vorherigen Stand der Kunst deutlich.

Zur Berücksichtigung semantischer Relationen zwischen Bildern integrieren wir Weltwissen über die Ähnlichkeiten von Klassen in Form von Taxonomien in tiefe Lernverfahren. Dazu wird der Merkmalsraum explizit so konstruiert, dass nicht primär visuell ähnliche, sondern semantisch ähnliche Klassen nahe beieinander liegen. Die auf diese Weise berechneten Bildrepräsentationen verbessern nicht nur die semantische Konsistenz der Ergebnisse der Bildersuche, sondern bieten auch Vorteile für andere Anwendungsfelder wie beispielsweise Klassifikation und Lernen aus wenigen Beispielen. Ein naheliegender nächster

Schritt besteht in der Betrachtung zusätzlicher semantischer Aspekte (z. B. Stimmungen und Aktivitäten) sowie der Kombination von semantischer und visueller Ähnlichkeit.

Literaturverzeichnis

[Ba20] Barz, Björn: Semantic and Interactive Content-based Image Retrieval. Dissertation, Univ. Jena. Cuvillier Verlag, Dezember 2020.

[BD21] Barz, Björn; Denzler, Joachim: Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era. In: ICPR 2020 Workshop on Content-Based Image Retrieval. 2021.

[Be20] Bertinetto, Luca; Mueller, Romain; Tertikas, Konstantinos; Samangooei, Sina; Lord, Nicholas A.: Making Better Mistakes: Leveraging Class Hierarchies with Deep Networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 12503–12512, June 2020.

[BL15] Babenko, Artem; Lempitsky, Victor: Aggregating Local Deep Features for Image Retrieval. In: IEEE International Conference on Computer Vision (ICCV). S. 1269–1277, Dec 2015.

[Br20] Brown, Andrew; Xie, Weidi; Kalogeiton, Vicky; Zisserman, Andrew: Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval. In: European Conference on Computer Vision (ECCV). Springer International Publishing, Cham, 2020.

[Do15] Dougherty, Conor: Google Photos Mistakenly Labels Black People 'Gorillas'. New York Times, July 2015. `https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/`. Accessed 2019-05-21.

[Sm00] Smeulders, Arnold WM; Worring, Marcel; Santini, Simone; Gupta, Amarnath; Jain, Ramesh: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 22:1349–1380, 12 2000.

Björn Barz wurde 1991 in Kassel geboren und studierte von 2011 bis 2016 Informatik an der Friedrich-Schiller-Universität Jena mit den Schwerpunkten Digitale Bildverarbeitung und Computerlinguistik. Seine Bachelorarbeit zum Thema "Interaktives Lernen von Objektdetektoren" wurde mit dem Wissenschaftspreis der Stadt Jena für anwendungsorientierte Abschlussarbeiten ausgezeichnet. Seine Masterarbeit verfasste er zum Thema der unüberwachten Detektion anomaler Abschnitte in multivariaten Zeitreihen. Die Ergebnisse dieser Arbeit wurden zudem in der renommierten Fachzeitschrift IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) veröffentlicht. Von 2016 bis 2020 forschte er im Rahmen einer Promotion am Lehrstuhl für Digitale Bildverarbeitung von Prof. Dr. Joachim Denzler, ebenfalls in Jena, auf dem Themengebiet der inhaltsbasierten Bildersuche. Dabei beschäftigte er sich insbesondere mit interaktiven Verfahren sowie semantischen Bildrepräsentation. Die Arbeit zu letztgenannter Thematik wurde 2019 auf der IEEE Winter Conference on Applications of Computer Vision mit dem Best Paper Award ausgezeichnet. Momentan ist Björn Barz Postdoktorand am Lehrstuhl für Digitale Bildverarbeitung.

Automatisierte Hybride Zeitreihenprognose: Design, Benchmarking und Anwendungsfälle[1]

André Bauer[2]

Abstract: Moderne Cloud-Umgebungen unterliegen Lastschwankungen und entsprechend schnellen und unerwarteten Änderungen. Um ausreichend Rechenressourcen rechtzeitig bereitzustellen, müssen sogenannte Auto-Skalierer den zukünftigen Ressourcenbedarf vorhersagen. Allerdings haben bestehende Arbeiten zur Zeitreihenprognose und zur automatischen Skalierung der Cloud zwei große Probleme. Erstens gibt es keinen vollautomatischen und generischen Prognoseansatz, der die vorhandenen Prognosemethoden so kombiniert, dass ihre Stärken genutzt und ihre Schwächen vermieden werden, um genaue Vorhersagen mit einer verlässlichen Laufzeit zu liefern. Zweitens wird bestehenden Auto-Skalierern misstraut, ein zuverlässiges und kosteneffizientes autonomes Ressourcenmanagement für moderne Cloud-Umgebungen zu bieten, da die Sorge besteht, dass ungenaue oder verzögerte Anpassungen zu finanziellen Verlusten führen können. Um diese Probleme zu lösen, stellt die Dissertation drei Beiträge vor: (i) Einen Prognose-Benchmark, der das Problem der begrenzten Vergleichbarkeit zwischen bestehenden Prognosemethoden adressiert; (ii) Eine automatisierte hybride Zeitreihen-Prognosemethode; (iii) Einen neuartiger hybrider Auto-Skalierer für koordinierte Skalierung von Anwendungen.

1 Einleitung

Heutzutage leben wir in einer digitalisierten Welt. Sowohl unser berufliches als auch unser privates Leben ist von verschiedenen IT-Diensten durchzogen, welche typischerweise in verteilten Computersystemen (z.B. Cloud-Umgebungen) betrieben werden. Die Betreiber solcher Systeme sind aufgrund des hohen Digitalisierungsgrades mit schnellen und wechselnden Anforderungen konfrontiert. Insbesondere Cloud-Umgebungen unterliegen starken Lastschwankungen und entsprechenden schnellen und unerwarteten Änderungen des Bedarfs an Rechenressourcen. Um dieser Herausforderung zu begegnen, können so genannte Auto-Skalierer, wie z.B. der schwellenwertbasierte Mechanismus von Amazon Web Services EC2, eingesetzt werden, um eine elastische Skalierung der Rechenressourcen zu ermöglichen. Doch trotz dieser Gelegenheit werden geschäftskritische Anwendungen nach wie vor mit deutlich überdimensionierten Rechenkapazitäten betrieben, um einen stabilen und zuverlässigen Dienstbetrieb zu gewährleisten. Diese Strategie wird aufgrund des mangelnden Vertrauens in Auto-Skalierer und der Sorge verfolgt, dass ungenaue oder verzögerte Anpassungen zu finanziellen Verlusten führen könnten.

Um die Ressourcenkapazität rechtzeitig anpassen zu können, müssen die zukünftigen Ressourcenanforderungen "vorhergesehen" werden. Denn die Reaktion auf Veränderungen,

[1] Englischer Titel der Dissertation: "Automated Hybrid Time Series Forecasting: Design, Benchmarking, and Use Cases"

[2] Universität Würzburg, Fakultät für Mathematik und Informatik, andre.bauer@uni-wuerzburg.de

sobald diese beobachtet werden, führt zu einer inhärenten Verzögerung. Mit anderen Worten, es sind genaue Prognosemethoden erforderlich, um Systeme proaktiv anzupassen. Ein wirksamer Ansatz in diesem Zusammenhang ist die Zeitreihenprognose, welche auch in vielen anderen Bereichen angewandt wird. Die Kernidee besteht darin, vergangene Werte zu untersuchen und vorherzusagen, wie sich diese Werte im Laufe der Zeit entwickeln werden. Nach dem "No-Free-Lunch Theorem" [WM97] gibt es keinen Algorithmus, der für alle Szenarien am besten funktioniert. Daher ist die Auswahl einer geeigneten Prognosemethode für einen gegebenen Anwendungsfall eine wesentliche Herausforderung. Denn jede Methode hat - abhängig vom spezifischen Anwendungsfall - ihre Vor- und Nachteile. Deshalb basiert üblicherweise die Wahl der Prognosemethode auf Trial-and-Error oder auf Expertenwissen, welches nicht vollständig automatisiert werden kann. Beide Ansätze sind teuer und fehleranfällig.

Obwohl Auto-Skalierung und Zeitreihenprognose etablierte Forschungsgebiete sind, können die bestehenden Ansätze die genannten Herausforderungen nicht vollständig bewältigen: (i) Bei unserer Untersuchung zur Zeitreihenvorhersage stellten wir fest, dass die meisten der überprüften Artikel nur eine geringe Anzahl von (meist verwandten) Methoden berücksichtigen und ihre Performanz auf einem kleinen Datensatz von Zeitreihen mit nur wenigen Fehlermaßen bewerten, während sie keine Informationen über die Ausführungszeit der untersuchten Methoden liefern. Daher können solche Artikel nicht als Hilfe für die Wahl einer geeigneten Methode für einen bestimmten Anwendungsfall herangezogen werden; (ii) Bestehende hybride open-source Prognosemethoden [Ce17, THA18, BHB16, TL18, Sm20], die sich mindestens zwei Methoden zunutze machen, um das "No-Free-Lunch Theorem" anzugehen, sind rechenintensiv, schlecht automatisiert, für einen bestimmten Datensatz ausgelegt oder haben eine unvorhersehbare Laufzeit. Methoden, die eine hohe Varianz in der Ausführungszeit aufweisen, können nicht für zeitkritische Szenarien angewendet werden (z.B. Auto-Skalierung), während Methoden, die auf einen bestimmten Datensatz zugeschnitten sind, Einschränkungen für mögliche Anwendungsfälle mit sich bringen (z.B. nur jährliche Zeitreihen vorhersagen); (iii) Auto-Skalierer skalieren typischerweise eine Anwendung entweder proaktiv oder reaktiv. Obwohl es einige hybride Auto-Skalierer [AETE12, Ji13, Ur08, Wu16, Iq11] gibt, fehlt es ihnen an ausgeklügelten Lösungen zur Kombination von reaktiver und proaktiver Skalierung. Beispielsweise werden Ressourcen nur proaktiv freigesetzt, während die Ressourcenzuweisung vollständig reaktiv (inhärent verzögert) erfolgt; (iv) Die Mehrheit der vorhandenen Mechanismen berücksichtigt bei der Skalierung einer Anwendung in einer öffentlichen Cloud-Umgebung nicht das Preismodell des Anbieters, was häufig zu überhöhten Kosten führt. Auch wenn einige kosteneffiziente Auto-Skalierer vorgeschlagen wurden, berücksichtigen sie nur den aktuellen Ressourcenbedarf und vernachlässigen ihre Entwicklung im Laufe der Zeit. Beispielsweise werden Ressourcen oft vorzeitig abgeschaltet, obwohl sie vielleicht bald wieder benötigt werden.

Um den genannten Herausforderungen und den Defiziten der bisherigen Arbeiten zu begegnen, wurden in der Dissertation [Ba20] drei Ziele formuliert:

Ziel I: *Einen Prognose-Benchmark bereitstellen, um gleiche Voraussetzungen für die Bewertung und den Vergleich der Leistung von Prognosemethoden in*

einem breiten Rahmen zu schaffen, der eine Vielzahl von Bewertungsszenarien abdeckt.

Ziel II: *Bereitstellung einer vollautomatischen und generischen hybriden Prognosemethode, die automatisch relevante Informationen aus einer gegebenen Zeitreihe extrahiert und diese nutzt, um bestehende Methoden so zu kombinieren, dass eine hohe Prognosegenauigkeit bei gleichzeitig geringer Laufzeit-Varianz erreicht wird.*

Ziel III: *Entwicklung eines hybriden Auto-Skalierers, der die koordinierte Skalierung von Anwendungen ermöglicht, indem er proaktive Skalierung (basierend auf der entwickelten Prognosemethode) mit reaktiver Skalierung als Fallback-Mechanismus kombiniert, um maximale Zuverlässigkeit der Ressourcenanpassungen zu gewährleisten.*

Die auf diesen Zielen basierenden Beiträge der Dissertation sind in Abschnitt 2 zusammengefasst. Anschließend wird im Abschnitt 3 die Grundidee des Forschungsbeitrag II (eine automatisierte hybride Zeitreihen-Prognosemethode namens *Telescope*) präsentiert. In Abschnitt 4 werden dann die wichtigsten Ergebnisse der Auswertung von Telescope vorgestellt. Abschließend erfolgt ein Resümee der Dissertation.

2 Übersicht der Dissertationsbeiträge

Im Rahmen der Dissertation [Ba20] wurden folgende drei Beiträge vorgestellt: (i) Der erste Beitrag - ein *Prognosebenchmark*[3] - behandelt das Problem der begrenzten Vergleichbarkeit zwischen bestehenden Prognosemethoden; (ii) Der zweite Beitrag stellt eine automatisierte hybride Zeitreihen-Prognosemethode namens *Telescope*[4] vor, die sich der Herausforderung des "No-Free-Lunch Theorem" [WM97] stellt; (iii) Der dritte Beitrag stellt *Chamulteon*, einen neuartigen hybriden Auto-Skalierer für die koordinierte Skalierung von Anwendungen bereit, der Telescope zur Vorhersage der Lastintensität als Grundlage für eine proaktive Ressourcenbereitstellung nutzt. Im Folgenden werden die drei Beiträge der Arbeit zusammengefasst.

Prognosebenchmark Um gleiche Ausgangsbedingungen für die Bewertung von Prognosemethoden anhand eines breiten Spektrums zu schaffen, schlagen wir einen neuartigen Benchmark vor, der Prognosemethoden auf der Grundlage ihrer Performanz in einer Vielzahl von Szenarien automatisch bewertet und ein Ranking erstellt. Der Benchmark umfasst vier verschiedene Anwendungsfälle, die jeweils 100 heterogene Zeitreihen aus verschiedenen Bereichen abdecken. Der Datensatz wurde aus öffentlich zugänglichen Zeitreihen zusammengestellt und so konzipiert, dass er eine viel höhere Diversität aufweist als bestehende Prognosewettbewerbe. Neben dem neuen Datensatz führen wir zwei neue Maße ein, die verschiedene Aspekte einer Prognose beschreiben. Wir haben den entwickelten Benchmark zur Bewertung von Telescope angewandt.

[3] Prognosebenchmark: `https://github.com/DescartesResearch/ForecastBenchmark`
[4] Telescope: `https://github.com/DescartesResearch/telescope`

Telescope Um eine generische Prognosemethode bereitzustellen, stellen wir einen neuartigen, auf maschinellem Lernen basierenden Prognoseansatz vor, der automatisch relevante Informationen aus einer gegebenen Zeitreihe extrahiert. Genauer gesagt, Telescope extrahiert automatisch intrinsische Zeitreihenmerkmale und zerlegt die Zeitreihe dann in Komponenten, wobei für jede dieser Komponenten ein Prognosemodell erstellt wird. Jede Komponente wird mit einer anderen Methode prognostiziert und dann wird die endgültige Prognose aus den vorhergesagten Komponenten unter Verwendung eines regressionsbasierten Algorithmus des maschinellen Lernens zusammengestellt. In mehr als 1300 Experimentstunden, in denen 15 konkurrierende Methoden (einschließlich Ansätze von Facebook [TL18] und Uber [Sm20]) auf 400 Zeitreihen verglichen wurden, übertraf Telescope alle Methoden und zeigte die beste Prognosegenauigkeit in Verbindung mit einer niedrigen und zuverlässigen Ausführungszeit. Im Vergleich zu den konkurrierenden Methoden, die im Durchschnitt einen Prognosefehler (genauer gesagt, den symmetric mean absolute forecast error) von 29% aufwiesen, wies Telescope einen Fehler von 20% auf und war dabei 2556 mal schneller. Insbesondere die Methoden von Uber und Facebook wiesen einen Fehler von 48% bzw. 36% auf und waren 7334 bzw. 19-mal langsamer als Telescope.

Chamulteon Um eine zuverlässige Auto-Skalierung zu ermöglichen, stellen wir einen hybriden Auto-Skalierer vor, der proaktive und reaktive Techniken kombiniert, um verteilte Cloud-Anwendungen, die mehrere Dienste umfassen, koordiniert und kostengünstig zu skalieren. Genauer gesagt, werden proaktive Anpassungen auf der Grundlage von Prognosen von Telescope geplant, während reaktive Anpassungen auf der Grundlage tatsächlicher Beobachtungen der überwachten Lastintensität ausgelöst werden. Um auftretende Konflikte zwischen reaktiven und proaktiven Anpassungen zu lösen, wird ein komplexer Konfliktlösungsalgorithmus implementiert. Außerdem überprüft Chamulteon Anpassungen im Hinblick auf das Preismodell des Cloud-Anbieters, um die anfallenden Kosten in öffentlichen Cloud-Umgebungen zu minimieren. In mehr als 400 Experimentstunden, in denen fünf konkurrierende Auto-Skalierungsmechanismen unter fünf verschiedene Arbeitslasten, vier verschiedene Anwendungen und drei verschiedene Cloud-Umgebungen evaluiert wurden, zeigte Chamulteon die beste Auto-Skalierungsleistung und Zuverlässigkeit bei gleichzeitiger Reduzierung der berechneten Kosten. Die konkurrierenden Methoden lieferten während (durchschnittlich) 31% der Versuchszeit zu wenige Ressourcen. Im Gegensatz dazu reduzierte Chamulteon diese Zeit auf 8% und die SLO-Verletzungen (Service Level Objectives) von 18% auf 6%, während es bis zu 15% weniger Ressourcen verwendete und die berechneten Kosten um bis zu 45% senkte.

3 Grundidee von Telescope

Die Annahme der Datenstationarität[5] ist eine inhärente Einschränkung für die Zeitreihenprognose. Jede Zeitreiheneigenschaft, welche die Stationarität verletzt, wie z. B. ein nicht konstanter Mittelwert, eine nicht konstante Varianz oder ein multiplikativer Effekt, stellt eine Herausforderung für die richtige Modellbildung dar [MSA18]. Daher transformiert Telescope automatisch die gegebene Zeitreihe, leitet intrinsische Merkmale aus

[5] Bei einer stationären Zeitreihe ändern sich die statistischen Eigenschaften (wie z. B. Mittelwert, Varianz und Autokovarianz) im Laufe der Zeit nicht.

der Zeitreihe ab, wählt einen geeigneten Satz von Merkmalen aus und behandelt jedes Merkmal separat. Mit anderen Worten, wir integrieren verschiedene Methoden, um nichtstationäre Zeitreihen zu behandeln.

Algorithm 1: Telescope Prognose

```
   Input: Time series ts, horizon h
   Result: Forecast of ts
1  [ts, freqs] = Preprocessing(ts);
2  if freqs[1] > 1 then // ts ist saisonal
3  |  features = FeatureExtraction(ts, freqs);
4  |  model = ModelBuilding(ts, features);
5  |  forecast = Forecasting(model, h);
6  else // ts ist nicht saisonal und kann daher nicht zerlegt werden
7  |  forecast = ARIMA(ts, h); // Fallback: ARIMA Prognose
8  end
9  forecast = Postprocessing(forecast);
10 return forecast
```

Der Arbeitsablauf von Telescope kann in fünf Phasen geteilt werden und ist in Algorithmus 1 vereinfacht dargestellt und erhält als Eingabe eine univariate Zeitreihe *ts* und den Horizont *h*. Der Horizont gibt an, wieviele Werte auf einmal vorhergesagt werden müssen. In der ersten Phase (Zeile 1) wird die Zeitreihe vorverarbeitet und die Frequenzen[6] der zugrunde liegenden Muster in der Zeitreihe werden extrahiert. Falls die Zeitreihe saisonal ist, umfassen die zweite und dritte Phase von Telescope die Extraktion relevanter intrinsischer Zeitreihenmerkmale (Zeile 3) und die Erstellung eines Modells, das die Zeitreihe auf der Grundlage dieser Merkmale beschreibt (Zeile 4). Anschließend wird das Modell verwendet, um das Verhalten der zukünftigen Zeitreihe vorherzusagen (Zeile 5). In dem Fall, dass die Zeitreihe nicht saisonal ist (Zeile 7), wird die Zeitreihe mit ARIMA (Auto-Regressive Integrated Moving Average) [BJ70] modelliert und vorhergesagt. Schließlich wird die Prognose entsprechend der Preprocessing-Phase nachbearbeitet und zurückgegeben. Im Folgenden wird jede Phase im Detail erläutert.

Preprocessing Da Vorhersagemethoden, insbesondere Methoden des maschinellen Lernens, mit sich ändernder Varianz und multiplikativen Effekten innerhalb einer Zeitreihe Schwierigkeiten haben [SK12], wird die Zeitreihe transformiert. Genauer gesagt wendet Telescope die Box-Cox-Transformation [BC64] an, da sie sowohl die Varianz als auch die multiplikativen Effekte der Zeitreihe reduziert, was zu einem verbesserten Vorhersagemodell [MSA18] führt. Parallel zur Transformation extrahiert Telescope die dominantesten Frequenzen [7] aus der Zeitreihe durch Anwendung eines Periodogramms [Sc99].

Feature Extraction In dieser Phase ermittelt Telescope intrinsische Zeitreihenmerkmale zur Bewältigung typischer Probleme oder Schwierigkeiten, die bei der Modellierung einer Zeitreihe auftreten können: (i) Die Zeitreihe hat mehrere sich überlagernde wiederkehrende Muster und (ii) die Zeitreihe verletzt die stationäre Eigenschaft. Um das erste Problem anzugehen, bestimmt Telescope für jede dominante Frequenz die zugehörigen Fourier-

[6] In der Zeitreihenanalyse bezeichnet die Frequenz die Länge eines wiederkehrenden Musters.
[7] Mit dominant meinen wir die häufigste Periode wie z.B. Tage in einem Jahr.

Terme der Zeitreihe zur Modellierung der verschiedenen Muster. Genauer gesagt werden für jede dominante Frequenz sowohl ein Sinus als auch ein Kosinus mit der Periodenlänge der entsprechenden Frequenz extrahiert. Um die Nicht-Stationarität zu behandeln, besteht die Kernidee von Telescope darin, die Zeitreihe zu zerlegen und dann jeden Teil separat zu behandeln. Zu diesem Zweck wird die Zeitreihe mit STL (Seasonal and Trend decomposition using Loess) [Cl90] in ihre Komponenten Trend (langfristige Entwicklung der Zeitreihe), Saisonalität (wiederkehrendes Muster innerhalb einer regemäßigen Periode) und Unregelmäßigkeit (verbleibende Teil der Zeitreihe, der nicht durch Trend oder Saisonalität beschrieben wird) zerlegt.

Model Building Um ein geeignetes Vorhersagemodell zu erstellen, das die Zeitreihe beschreibt, implementiert Telescope XGBoost (eXtreme Gradient Boosting) [CG16], um die Beziehung zwischen der Zeitreihe und den intrinsischen Merkmalen zu finden. Da ein starker Trend sowohl die Varianz erhöht als auch die Stationarität verletzt, entfernt Telescope den Trend aus der Zeitreihe. Die resultierende enttrendete Zeitreihe ist nun trend-stationär. Obwohl auch die Saisonalität die Stationarität verletzen kann, eignen sich Methoden des maschinellen Lernens zur Mustererkennung. Folglich lernt XGBoost während seines Trainingsverfahrens, wie die enttrendete Zeitreihe durch die intrinsischen Merkmale Fourier-Terme und Saisonalität beschrieben werden kann.

Forecasting Um die Zeitreihe vorhersagen zu können, müssen die verschiedenen Komponenten getrennt voneinander vorhergesagt werden. Die Saisonalität und die Fourier-Terme sind per Definition wiederkehrende Muster, so dass sie einfach fortgesetzt werden können. Die resultierende vorhergesagte Saisonalität und die vorhergesagten Fourier-Terme werden in Verbindung mit dem Vorhersagemodell als Merkmale für die Vorhersage der zukünftigen enttrendeten Zeitreihen verwendet. Genauer gesagt, regressiert das maschinelle Lernverfahren für jeden Zeitpunkt der Vorhersage einen neuen Wert auf Basis der entsprechenden Werte der Merkmale. Parallel zur Vorhersage der wiederkehrenden Muster wird auch der Trend vorhergesagt. Da der Trend keine wiederkehrenden Muster enthält, ist eine fortschrittliche Prognosemethode erforderlich. Zu diesem Zweck verwenden wir ARIMA [BJ70], da es in der Lage ist, den Trend auch aus wenigen Punkten zu schätzen. Nachdem der Trend vorhergesagt wurde, werden im letzten Schritt dieser Phase die vorhergesagte enttrendete Zeitreihen und der vorhergesagte Trend aufsummiert.

Postprocessing Da die Zeitreihe in der Preprocessing-Phase transformiert wurde, wird die vorhergesagte Zeitreihe mit der inversen Box-Cox-Transformation zurücktransformiert.

4 Evaluation von Telescope

Um die Qualität von Telescope zu evaluieren, haben wir unsere hybride Prognosemethode auf 400 verschiedenen Zeitreihen mit 15 unterschiedlichen Prognosemethoden verglichen. Aus Platzgründen vergleichen wir Telescope in diesem Abschnitt mit vier aktuellen hybriden Prognosemethoden (BETS [BHB16], ES-RNN [Sm20], FFORMS [THA18] und Prophet [TL18]) und die nach unseren Experimenten [Ba20] besten Methoden aus dem Bereich der “klassischen” Zeitreihenprognose (sARIMA [BJ70]) als auch des maschinellen

Lernens (XGBoost [CG16]). Um die Vorhersagegenauigkeit zu messen, berücksichtigen wir den sMAPE (symmetical mean absolute percentage error) [Ma93], wobei $\bar{e}_{\text{sM}}$ den durchschnittlichen Fehler und $\sigma_{e_{\text{sM}}}$ die Standardabweichung des Fehlers widerspiegelt. Neben der Genauigkeit messen wir auch die Laufzeit der einzelnen Methoden. Die Laufzeit wird durch die Zeit normalisiert, die eine naïve Prognosemethode benötigt (im Schnitt 0,1 Sekunde pro Vorhersage), wobei $\bar{t}_{sN}$ die durchschnittliche normierte Zeit und $\sigma_{t_{sN}}$ die Standardabweichung der normierten Zeit reflektiert. Als Beispiel für die Vorhersagen zeigt Abbildung 1 die konkurrierenden Prognosemethoden auf der Zeitreihe Air Passsengers, welche oft als Referenz verwendet wird.

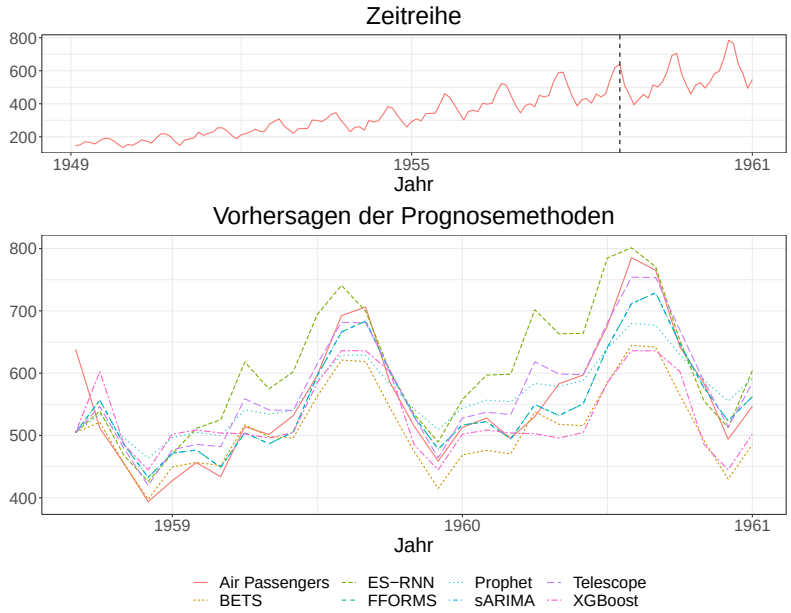


Abb. 1: Vorhersagen für alle konkurrierenden Methoden auf der Zeitreihe Air Passsengers.

Tab. 1: Vorhersagefehler- und Laufzeitvergleich über alle Zeitreihen.

Maß	**BETS**	**ES-RNN**	**FFORMS**	**Prophet**	**sARIMA**	**Telescope**	**XGBoost**
$\bar{e}_{\text{sM}}$ [%]	25,52	47,87	21,15	35,56	20,63	**19,95**	23,85
$\sigma_{e_{\text{sM}}}$ [%]	34,25	66,99	36,87	$2{,}30{\cdot}10^{2}$	35,63	**31,35**	34,67
$\bar{t}_{sN}$	$6{,}14{\cdot}10^{4}$	$1{,}61{\cdot}10^{6}$	$5{,}23{\cdot}10^{5}$	$2{,}04{\cdot}10^{3}$	$8{,}48{\cdot}10^{5}$	$1{,}43{\cdot}10^{2}$	**6,73**
$\sigma_{t_{sN}}$	$1{,}07{\cdot}10^{6}$	$4{,}92{\cdot}10^{6}$	$7{,}93{\cdot}10^{6}$	$6{,}15{\cdot}10^{3}$	$9{,}08{\cdot}10^{6}$	98,67	**15,59**

Tabelle 1 zeigt die Ergebnisse, die über alle Zeitreihen gemittelt wurden, für alle konkurrierenden Methoden. Jede Zeile steht für ein Maß, jede Spalte für eine Methode, und die besten Werte (je niedriger, desto besser) sind fett hervorgehoben. Die genaueste Prognosemethode ist Telescope (19,95%), gefolgt von sARIMA (20,63%). Den höchsten Fehler weist ES-RNN mit 47,87% auf. Die mit Abstand schnellste Methode ist XGBoost (6,73). Telescope ($1{,}43{\cdot}10^{2}$) hat die zweitniedrigste Laufzeit. Die langsamste Methode ist ES-RNN ($1{,}61{\cdot}10^{6}$). Obwohl sARIMA die zweitbeste Vorhersagegenauigkeit hat, ist es im

Durchschnitt fast 6000 Mal langsamer als Telescope. Beispielsweise betrug die maximale Laufzeit von sARIMA für eine Zeitreihe 465.574 Sekunden, was fast 5,5 Tagen entspricht.

Obwohl unser Datensatz mit 400 verschiedenen Zeitreihen ein breites Spektrum an Anwendungsfällen umfasst, lassen sich die Evaluierungsergebnisse möglicherweise nicht auf alle Zeitreihen aus allen Bereichen verallgemeinern. Neben dem Datensatz versuchten wir auch, repräsentative Vorhersagemethoden zu untersuchen, die auf verschiedenen Techniken basieren, wie beispielsweise die Methoden von Facebook (Prophet) und Uber (ES-RNN). Wir verwenden jedoch alle Methoden mit ihren Standardeinstellungen. Folglich können die beobachteten Ergebnisse abweichen, wenn die Prognosemethoden auf die einzelnen Zeitreihen angepasst werden.

5 Resümee

Heutzutage leben wir in einer schnelllebigen Welt, und so sind viele Bereiche Trends und unterschiedlichen Anforderungen unterworfen. So müssen Cloud-Umgebungen beispielsweise mit Lastschwankungen und entsprechend schnellen und unerwarteten Änderungen des Bedarfs an Rechenressourcen zurechtkommen. Da die Reaktion auf Änderungen, sobald sie beobachtet werden, eine inhärente Verzögerung mit sich bringt, muss der zukünftige Ressourcenbedarf vorhergesagt werden, um notwendige Schritte im Voraus zu identifizieren. Eine nützliche und etablierte Technik in diesem Zusammenhang ist die Zeitreihenprognose, die auch in vielen anderen Bereichen eingesetzt wird. Obwohl die Zeitreihenprognose die proaktive automatische Skalierung der benötigten Ressourcen in Cloud-Umgebungen ermöglicht, werden geschäftskritische Anwendungen immer noch mit stark überprovisionierten Ressourcen betrieben, um einen stabilen und zuverlässigen Servicebetrieb zu gewährleisten. Diese Strategie wird vor allem aufgrund von zwei Hauptproblemen der bestehenden Arbeiten verfolgt: Erstens gibt es keinen vollautomatischen und generischen Prognoseansatz, der die vorhandenen Prognosemethoden so kombiniert, dass ihre Stärken genutzt und ihre Schwächen vermieden werden, um genaue Prognosen mit einer zuverlässigen Laufzeit zu liefern. Zweitens wird bestehenden Auto-Skalierer misstraut, ein zuverlässiges und kosteneffizientes autonomes Ressourcenmanagement für moderne Cloud-Umgebungen zu ermöglichen, da die Sorge besteht, dass ungenaue oder verzögerte Anpassungen zu finanziellen Verlusten führen können.

Um den genannten Herausforderungen zu begegnen, wurden in der Dissertation [Ba20] drei Beiträge vorgestellt, welche als wichtige Meilensteine auf dem Gebiet der Zeitreihenprognose und der automatischen Skalierung in Cloud-Umgebungen angesehen werden können. (i) In dieser Arbeit wird zum ersten Mal ein Prognosebenchmark präsentiert, der eine Vielzahl verschiedener Bereiche mit einer hohen Diversität zwischen den analysierten Zeitreihen abdeckt. Auf der Grundlage des zur Verfügung gestellten Datensatzes und des automatischen Auswertungsverfahrens trägt der vorgeschlagene Benchmark dazu bei, die Vergleichbarkeit von Prognosemethoden zu verbessern. Die Benchmarking-Ergebnisse von verschiedenen Prognosemethoden ermöglichen die Auswahl der am besten geeigneten Prognosemethode für einen gegebenen Anwendungsfall. (ii) Telescope bietet den ersten generischen und vollautomatischen Zeitreihen-Prognoseansatz, der sowohl genaue als

auch zuverlässige Prognosen liefert, ohne Annahmen über die analysierte Zeitreihe zu treffen. Dementsprechend macht es teure, zeitaufwändige und fehleranfällige Verfahren überflüssig, wie z.B. Trial-and-Error oder das Hinzuziehen eines Experten. Dies eröffnet neue Möglichkeiten, insbesondere in zeitkritischen Szenarien, in denen Telescope genaue Vorhersagen mit einer kurzen und zuverlässigen Antwortzeit liefern kann. Obwohl Telescope für diese Arbeit im Bereich des Cloud Computing eingesetzt wurde, gibt es, wie die Auswertung zeigt, keinerlei Einschränkungen hinsichtlich der Anwendbarkeit von Telescope in anderen Bereichen. Darüber hinaus wird Telescope, das auf GitHub zur Verfügung gestellt wurde, bereits in einer Reihe von interdisziplinären datenwissenschaftlichen Projekten eingesetzt, z.B. bei der vorausschauenden Wartung im Rahmen von Industry 4.0, bei der Vorhersage von Herzinsuffizienz in der Medizin oder als Bestandteil von Vorhersagemodellen für die Entwicklung von Bienenstöcken. (iii) Im Kontext der elastischen Ressourcenverwaltung ist Chamulteon ein wichtiger Meilenstein für die Stärkung des Vertrauens in Auto-Skalierern. Der komplexe Konfliktlösungsalgorithmus ermöglicht ein zuverlässiges und genaues Skalierungsverhalten, das Verluste durch übermäßige Ressourcenzuweisung oder SLO-Verletzungen (Service Level Objectives) reduziert. Mit anderen Worten, Chamulteon bietet zuverlässige Ressourcenanpassungen, die die berechneten Kosten minimieren und gleichzeitig die Benutzerzufriedenheit maximieren.

Literaturverzeichnis

[AETE12] Ali-Eldin, Ahmed; Tordsson, Johan; Elmroth, Erik: An Adaptive Hybrid Elasticity Controller for Cloud Infrastructures. In: IEEE NOMS 2012. IEEE, S. 204–212, 2012.

[Ba20] Bauer, André: Automated Hybrid Time Series Forecasting: Design, Benchmarking, and Use Cases. Dissertation, University of Würzburg, Germany, 2020.

[BC64] Box, George EP; Cox, David R: An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), S. 211–252, 1964.

[BHB16] Bergmeir, Christoph; Hyndman, Rob J; Benítez, José M: Bagging Exponential Smoothing Methods Using STL Decomposition and Box–Cox Transformation. International journal of forecasting, 32(2):303–312, 2016.

[BJ70] Box, G.E.P.; Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day, 1970.

[Ce17] Cerqueira, Vítor; Torgo, Luís; Pinto, Fábio; Soares, Carlos: Arbitrated Ensemble for Time Series Forecasting. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, S. 478–494, 2017.

[CG16] Chen, Tianqi; Guestrin, Carlos: Xgboost: A Scalable Tree Boosting System. In: ACM SIGKDD 2016. ACM, S. 785–794, 2016.

[Cl90] Cleveland, Robert B; Cleveland, William S; McRae, Jean E; Terpenning, Irma: STL: A Seasonal-Trend Decomposition Procedure based on Loess. Journal of Official Statistics, 6(1):3–73, 1990.

[Iq11] Iqbal, Waheed; Dailey, Matthew N; Carrera, David; Janecek, Paul: Adaptive Resource Provisioning for Read Intensive Multi-Tier Applications in the Cloud. Future Generation Computer Systems, 27(6):871–879, 2011.

[Ji13] Jiang, Jing; Lu, Jie; Zhang, Guangquan; Long, Guodong: Optimal Cloud Resource Auto-Scaling for Web Applications. In: 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing. IEEE, S. 58–65, 2013.

[Ma93] Makridakis, Spyros: Accuracy measures: theoretical and practical concerns. International journal of forecasting, 9(4):527–529, 1993.

[MSA18] Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios: Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward. PloS one, 13(3):e0194889, 2018.

[Sc99] Schuster, Arthur: The Periodgram of Magnetic Declination as Obtained from the Records of the Greenwich Observatory during the Years 1871-1895. Transactions of the Cambridge Philosophical Society, 18:107–135, 1899.

[SK12] Sugiyama, Masashi; Kawanabe, Motoaki: Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. The MIT Press, 2012.

[Sm20] Smyl, Slawek: A Hybrid Method of Exponential Smoothing and Recurrent Neural Networks for Time Series Forecasting. International Journal of Forecasting, 36(1):75–85, 2020.

[THA18] Talagala, Thiyanga S; Hyndman, Rob J; Athanasopoulos, George: Meta-Learning How to Forecast Time Series. Bericht, Monash University, Department of Econometrics and Business Statistics, 2018.

[TL18] Taylor, Sean J; Letham, Benjamin: Forecasting at Scale. The American Statistician, 72(1):37–45, 2018.

[Ur08] Urgaonkar, Bhuvan; Shenoy, Prashant; Chandra, Abhishek; Goyal, Pawan; Wood, Timothy: Agile Dynamic Provisioning of Multi-tier Internet Applications. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 3(1):1–39, 2008.

[WM97] Wolpert, D. H.; Macready, W. G.: No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation, 1(1):67–82, Apr 1997.

[Wu16] Wu, S.; Li, B.; Wang, X.; Jin, H.: HybridScaler: Handling Bursting Workload for Multi-tier Web Applications in Cloud. In: 15th International Symposium on Parallel and Distributed Computing (ISPDC), 2016. S. 141–148, 2016.

André Bauer wurde 1989 in Deutschland geboren. Nach dem Erwerb der mittleren Reife, trat er 2006 eine Ausbildung zum Bauzeichner an. Nach dem Abschluss der Ausbildung entschied er sich sein Abitur nachzuholen und erlangte 2011 seine fachgebundene Hochschulreife. Er studierte Informatik an der Universität Würzburg. Nach seinem Masterabschluss fing er seine Promotion bei Prof. Dr. Samuel Kounev an und schloss diese 2020 mit Auszeichnung ab. Während seiner Promotion beschäftigte er sich mit Elastizität im Cloud Computing, Autoskalierung und Ressourcenmanagement, sowie der Datenanalyse und Modellbildung mittels Kombinationen von maschinellem Lernen und Verfahren aus der Zeitreihenanalyse. Seit der Disputation leitet er die Forschergruppe für “Data Science Engineering Group” am Lehrstuhl für Software Engineering der Universität Würzburg. Darüber hinaus ist er gewählter Release Manager der RG-Cloud Arbeitsgruppe und Newsletter Editor der SPEC Research Group.

Kooperative Absichtserkennung mittels maschineller Lernverfahren – Radfahrerschutz im Kontext des hochautomatisierten Fahrens [1]

Maarten Bieshaar[2]

Abstract: Das Radfahren wird im Verkehr der Zukunft eine zentrale Rolle spielen. Um Unfälle zu vermeiden, ist es entscheidend, Radfahrer frühzeitig zu erkennen und deren Absichten vorherzusagen. Fahrzeuge, die mit Sensoren, Datenverarbeitungssystemen und Kommunikationsfähigkeiten ausgestattet sind, erstellen und pflegen ein lokales Modell ihrer Verkehrsumgebung. Gruppen von kooperierenden und interagierenden Fahrzeugen, sowie Roadside Units, und Radfahrer, die mit Smart Devices (z.B. Smartphone und Smartwatch) und anderen am Körper getragenen Sensoren ausgestattet sind, tauschen Informationen aus. Sie bilden ein multimodales Sensorsystem mit dem Ziel, Radfahrer und deren Absichten zuverlässig zu erfassen. Die kollektive Intelligenz aller Verkehrsteilnehmer erlaubt es den Wahrnehmungshorizont der einzelnen Verkehrsteilnehmer über deren eigene sensorische Fähigkeiten hinaus zu erweitern und somit eine bessere Erkennung der Absichten von Radfahrer zu ermöglichen.[3]

1 Einführung

1.35 Millionen Verkehrstote pro Jahr wurde in der Ausgabe 2018 des von der Weltgesundheitsorganisation veröffentlichten globalen Statusberichts zur Verkehrssicherheit gemeldet [Wo18] . Darüber hinaus erleiden schätzungsweise 20 bis 50 Millionen weitere Menschen nicht-tödliche Verletzungen, welche zumeist in einer dauerhaften Behinderung enden. Ungeschützte Verkehrsteilnehmer (engl. Vulnerable Road Users, kurz VRU), d.h. Radfahrer und Fußgänger sind besonders gefährdet, da sie über kein passives Sicherheitssystem verfügen, das die Auswirkungen von Unfällen potenziell abmildern könnte. In den vergangenen Jahren haben sich zahlreiche Bürgerinitiativen für eine „Verkehrswende“ und damit für die Förderung des nachhaltigen und klimafreundlichen Radverkehrs starkgemacht. Obwohl die ersten Erfolge zu verzeichnen sind (d.h., dem Radverkehr wird in vielen Städten mehr Raum zugestanden) bleibt das Radfahren auf deutschen Straßen weiterhin gefährlich. 2019 starben auf deutschen Straßen im Durchschnitt 1 – 2 Radfahrer pro Tag.

Moderne Fahrerassistenzsysteme, wie bspw. automatische Warnassistenten, können die Sicherheit von ungeschützten Verkehrsteilnehmern, z.B. durch eine frühzeitige Warnung

[1] Englischer Titel der Dissertation: „Cooperative Intention Detection using Machine Learning – Advanced Cyclist Protection in the Context of Automated Driving“

[2] Universität Kassel, Intelligente Eingebettete Systeme, Kassel, Deutschland, mbieshaar@uni-kassel.de

[3] Aus Gründen der Lesbarkeit wird im Text die männliche Form verwendet, nichtsdestoweniger beziehen sich die Angaben auf Angehörige aller Geschlechter.

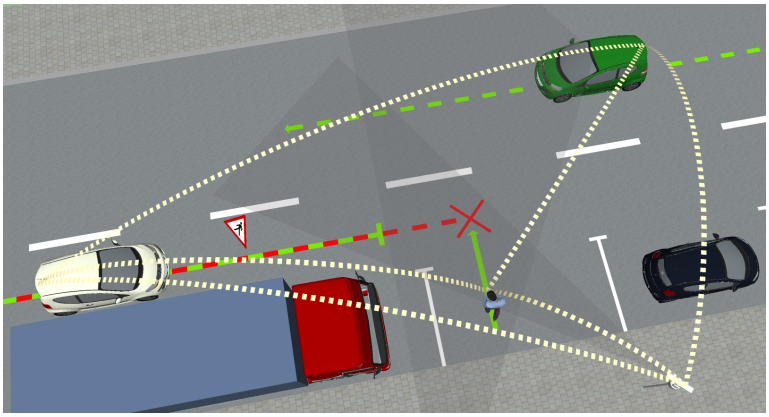


Abb. 1: Schematische Darstellung einer kritische Situation aus [Bi21]. Ein Fußgänger möchte die Straße kreuzen, wobei er durch einen LKW verdeckt wird. Die Situation wird durch die Kooperation (angedeutet durch die gestrichelten gelben Pfeile) und Informationsaustausch zwischen den Fahrzeugen, der RSU und dem Smartphone, das der Fußgänger bei sich trägt, aufgelöst.

deutlich erhöhen. Diese Systeme prognostizieren die Absicht von ungeschützten Verkehrsteilnehmern, d.h. die geplante, aber noch nicht ausgeführte Aktion, basierend auf lokal verfügbaren sensorischen Informationen. Moderne (sowie in Zukunft auch hoch-automatisierte) Fahrzeuge sind mit zahlreichen Sensoren ausgestattet, u.a. Kameras, LiDAR und RADAR. Dennoch kann es infolge einer Verdeckung (z.B. Radfahrer befindet sich im toten Winkel) oder einer Fehlfunktion der Sensorik stets zu gefährlichen und bisweilen sogar tödlichen Situationen kommen. Fehler, insbesondere in der maschinellen Wahrnehmung, haben zumeist eine gravierende Beeinträchtigung des Gesamtsystems zur Folge. Kommunikation und Kooperation mit anderen Verkehrsteilnehmern und mit Sensoren ausgestatteter Infrastruktur (engl. Roadside Unit, kurz RSU) stellt eine Möglichkeit dar Komplettausfälle des eigenen Wahrnehmungssystems zu kompensieren und mit Verdeckungen umzugehen.

Die der Arbeit zu Grunde liegende Vision des zukünftigen kooperative-interagierenden, hoch-automatisierten und vor allem sicheren Verkehrs lässt sich am folgenden Beispiel erläutern [Bi17]. Ein Fußgänger, der hinter einem Hindernis die Straße überquert, ist in Abb. 1 dargestellt. Das herannahende weiße Fahrzeug kann den verdeckten Fußgänger und seine Absicht, die Straße zu überqueren, nicht rechtzeitig erfassen. Das entgegenkommende Fahrzeug und eine RSU (hier illustriert durch eine Überwachungskamera) können jedoch den Fußgängern und dessen Absicht erkennen. Diese erkannte Absicht kann dem weißen Fahrzeug via moderner drahtloser Car2X-Kommunikation mitgeteilt werden, so dass der Fahrer oder das Fahrzeug selbst (im Falle des automatisierten Fahrens) eine Notbremsung einleiten kann. Eine weitere Informationsquelle sind die vom Fußgänger bei sich getragene Smart Devices, z.B. ein Smartphone, eine Smartwatch oder andere Wearables. Es wird demnach die kollektive Intelligenz aller Fahrzeuge, der RSU sowie der vom Fußgänger bei sich getragenen Smart Devices genutzt, um die Absicht des Fußgängers, die Straße zu überqueren, zu erkennen.

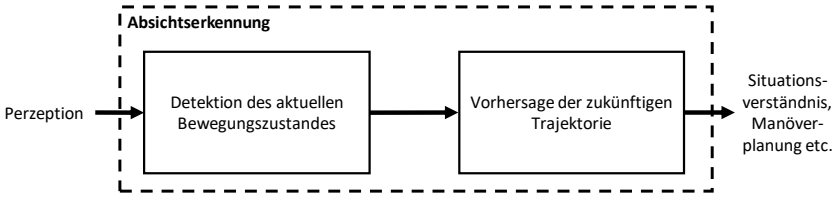


Abb. 2: Schematische Darstellung des zweistufigen Prozesses zur Absichtserkennung: Detektion des aktuellen Bewegungszustandes gefolgt von der Vorhersage der zukünftigen Trajektorie.

Das Ziel und die globale Vision dieser Arbeit (siehe [Bi21]) sind kooperativ-interagierende, hoch-automatisierte Fahrzeuge, die das Verhalten von ungeschützten Verkehrsteilnehmern berücksichtigen können. Die kollektive Intelligenz aller Fahrzeuge, ergänzt durch Informationen von RSU und von den ungeschützten Verkehrsteilnehmern bei sich getragenen Smart Devices, wird ausgenutzt, um die Sicherheit von ungeschützten Verkehrsteilnehmern zu erhöhen, indem deren Absichten prognostiziert werden. Im Folgenden werden alle Fahrzeuge, RSU und mit Smart Devices ausgestatteten ungeschützte Verkehrsteilnehmer als Agenten bezeichnet. Wir gehen nicht davon aus, dass alle Agenten immer mit der identischen Sensorik ausgestattet sind oder dass alle Informationen immer verfügbar sind. Stattdessen wird ein opportunistischer Ansatz angestrebt, d.h. es werden alle verfügbaren Informationen so gut wie möglich genutzt. Die Beiträge dieser Arbeit zu dem skizzierten zukünftigen Verkehrsszenario sind Algorithmen und Techniken zur opportunistischen kooperativen Absichtserkennung von Radfahrern.

2 Radfahrer-Absichtserkennung

Im Rahmen dieser Arbeit bezeichnen die Begriffe „Absicht“ bzw. „Intention“ eine geplante, aber noch nicht gestartete bzw. durchgeführte Aktion. Die Absichtserkennung (oder auch Intentionserkennung) wird als ein zweistufiger Prozess (siehe Abb. 2) (vgl. [Bi17]) betrachten. Im ersten Schritt wird die Bewegung des Radfahrers als eine Abfolge von Bewegungszuständen, z.B. Warten, Starten, Fahren und Abstoppen, gesehen. Das erste Ziel ist die frühzeitige Erkennung von Bewegungsübergängen, d.h. Änderungen des Bewegungszustands. Im zweiten Schritt wird die Bewegung bestimmter Körperpunkte, z.B. des Körperschwerpunkts, der Gelenke oder des Kopfes, im 3D-Raum berücksichtigt. Ziel ist es die zukünftige Trajektorie dieser Punkte für einen maximalen Zeitraum von 2.5 s vorherzusagen. Der Grund für die zweistufige Modellierung ist, dass das frühzeitige Wissen über eine Bewegungszustandsänderung die Vorhersage der zukünftigen Radfahrertrajektorie erheblich verbessert [Go16]. Sowohl die Erkennung des aktuellen Bewegungszustandes als auch die Trajektorienvorhersage sind Teil dessen, was wir als Absichtserkennung bezeichnen. Für die Detektion des Bewegungszustands sowie auch für die Prädiktion der zukünftigen Trajektorie werden maschinelle Lernverfahren verwendet. Für die Absichtserkennung betrachten wir je nach Agenten unterschiedliche Sensormodalitäten bspw. Videosequenzen einer Kamera oder abstraktere Daten wie die vergangenen Positionen, Ge-

schwindigkeiten, Beschleunigungen, Ausrichtungen und Posen des Radfahrers. Die erkannte Absicht bzw. die prädizierte zukünftige Trajektorie kann anschließend zur Situationsanalyse oder zur sicheren Manöverplanung verwendet werden [ES17].

3 Forschungsfragen und Wesentliche Beiträge

Die wesentlichen Beiträge der Arbeit lassen mittels der folgenden drei zentralen Forschungsfragen zusammenfassen:

Probabilistische Radfahrer-Trajektorienvorhersage – *Wie kann die zukünftige Trajektorie eines Radfahrers probabilistisch vorhergesagt werden?*
Prognosen über das zukünftige Verhalten von Radfahrern sind von Natur aus mit Unsicherheiten behaftet, da nicht alle Variablen, die das zukünftige Verhalten der Radfahrer beeinflussen, bekannt sind. Ziel der probabilistischen Trajektorienvorhersage ist es die (Un-)Sicherheit der Vorhersage zu modellieren und zu quantifizieren. Die Quantifizierung der Unsicherheit bei der Prognose ist ein wesentlicher Aspekt für sicheres hochautomatisiertes Fahren [Ei17]. Im Rahmen dieser Arbeit wird eine neuartige probabilistische Vorhersagemethode vorgestellt. Dabei wird die Unsicherheit mittels einer Wahrscheinlichkeitsverteilung modelliert. Die in dieser Arbeit neu entwickelte Methode auf Basis von neuronalen Netzwerken sowie der Generalisierung der Quantilen Regression erlaubt eine situationsabhängige, stets zuverlässige und zu gleich genaue Schätzung der zukünftigen Radfahrertrajektorie sowie der mit der Prognose einhergehenden Unsicherheit. Im Rahmen einer umfangreichen Studie wird in der Arbeit gezeigt, dass mit dieser Methode eine Steigerung der Güte der probabilistischen Prognosen von bis zu 99 % im Vergleich zu klassischen maschinellen Lernverfahren erreicht werden kann.

Radfahrer als zusätzliche Sensoren – *Wie können Smart Devices für die Radfahrer-Absichtserkennung eingesetzt werden?*
Laut einer Umfrage des Branchenverbandes Bitkom gab es 2018 in Deutschland 57 Millionen Smartphone-Nutzer, d.h. acht von zehn Menschen in Deutschland besitzen und nutzen ein Smartphone. Smart Devices, d.h. Smartphones, Smartwatches und anderen Wearables, verfügen über zahlreiche Sensoren darunter GPS, Accelerometer, Gyroskop und Magnetometer. Ziel ist es die Sensoren der Smart Devices für die Radfahrer-Absichtserkennung zu nutzen. Durch die Verfügbarkeit neuer Kommunikationstechnologien (z.B. 5G) wird es zudem möglich sein, dass Smart Devices mit Fahrzeugen oder RSU vernetzt werden können [SvSM17]. Ein Radfahrer ist somit selbst in Lage Information an Fahrzeuge oder RSU über die eigene Position, die aktuelle Bewegung (z.B. wartend oder fahrend) sowie der zukünftigen prädizierten Trajektorie zu schicken. Im Rahmen dieser Arbeit wird ein neuartiges Konzept zur Radfahrer-Absichtserkennung mittels Smart Devices unter Verwendung maschineller Lernmethoden vorgestellt. Darüber hinaus wird das Potential von intelligenten Fahrradhelmen zur Radfahrer-Absichtserkennung dargelegt. Zudem wird gezeigt, dass die Vernetzung mehrerer am Körper getragener smarter Geräte in Form eines neuartigen Fahrrad Ad Hoc Netzwerk (engl. Bicycle Area Network kurz BicAN), zu einer deutlich verbesserten Absichtserkennung führt.

Kooperative Absichtserkennung – *Wie können Absichten von Radfahrern kooperativ erkannt werden?*
Das Ziel dieser Arbeit ist, dass die kollektive Intelligenz aller Verkehrsteilnehmer sowie RSU und Smart Devices genutzt wird, um die Absichten von Radfahrer im Rahmen eines kooperativen Prozesses frühzeitig und vor allem zuverlässig zu erkennen. Die Kooperation beschreibt hier im Wesentlichen einen Fusionsprozess, der auf unterschiedlichen Ebene stattfinden kann, z.B. auf Ebene der Eingabe in den Absichtserkennungsprozesses (Kooperative Wahrnehmung), oder auf Ebene der detektierten Bewegungszustände, oder der prädizierte zukünftigen Trajektorien. Eine Besonderheit ist hierbei die Heterogenität der Sensoren der beteiligten Agenten. Dies reicht von Bildern von Fahrzeug- und Verkehrskameras bis hin zu den Inertialsensoren der Smart Devices. Darüber hinaus sind auch Effekte, wie z.B. Verzögerungen, verlorene oder außerhalb der Sequenz ankommende Nachrichten, welche durch die Verwendung des unterliegenden drahtlosen Kommunikationsmediums (z.B., 5G oder ITS-G5) entstehen können, beim Entwurf der Methoden zur kooperativen Absichtserkennung zu betrachten.

4 Probabilistische Radfahrer-Trajektorienvorhersage

Die probabilistische Trajektorienprognose hat das Ziel die zukünftige Aufenthaltswahrscheinlichkeit des Radfahrers zu modellieren, d.h. anstatt einer Punktprognose wird eine Wahrscheinlichkeitsverteilung über mögliche zukünftige Aufenthaltsorte prädiziert. Die vorhergesagten Verteilungen sollen dabei nicht statisch sein, sondern sich der beobachteten Situation anpassen. Ziel der probabilistischen Vorhersagemethoden ist es immer möglichst scharfe (d.h. kompakte) Verteilungen vorherzusagen. Die Randbedingung ist dabei, dass die vorhersagten Verteilungen zuverlässig sind, d.h. wenn man aus einer Verteilung den Bereich mit 95 % Aufenthaltswahrscheinlichkeit extrahiert, dann sollte der Radfahrer auch tatsächlich in 95 % der Fälle in diesem Bereich sein. Letzteres wird auch als Zuverlässigkeit der Vorhersage bezeichnet. Insbesondere die Zuverlässigkeit von probabilistischen Prognosen für Radfahrer ist eine große Herausforderung, z.B. bei Übergängen zwischen Bewegungszuständen. Mit sehr großer Wahrscheinlichkeit wird der Radfahrer das was er gerade macht auch in der kommenden Sekunde machen, d.h. wenn der Radfahrer geradeaus fährt, dann ist die Wahrscheinlichkeit, dass er dies auch in der nächsten Sekunden machen wird, sehr hoch. Es besteht trotzdem noch zu einem geringen Anteil die Möglichkeit, dass er plötzlich abbremst oder abstoppt. Die für die Modellierung notwendigen multivariaten Verteilungen sind schief und haben schwere Ränder, weshalb klassische Methoden wie bspw. neuronale Netzwerke mit Gaußscher Likelihood nicht geeignet sind, bzw. keine zuverlässigen Vorhersagen zulassen (siehe [Ze19]). Im Rahmen dieser Arbeit wird eine neuartige Methode zur Modellierung beliebiger, multivariater, unimodaler, sternförmigen Verteilungen vorgestellt. Diese neue Methode ist eine Erweiterung der weit verbreiteten Quantilen Regression (QR) [KB78] auf multivariate Zielgrößen. Der große Vorteil der QR ist die nicht-parametrische Natur der Vorhersagen, d.h. es können beliebige Verteilungen modelliert werden insbesondere auch schiefe und solche mit schweren Rändern. QR-Methoden sind allerdings nur für die Vorhersage univariater (d.h. eindimensionaler) Zielgröße geeignet, da sie auf der Vorhersage von Quantilen beruhen. Da es jedoch bei multivariaten (d.h. mehrdimensionalen) Zielgrößen keine inhärente Ordnung

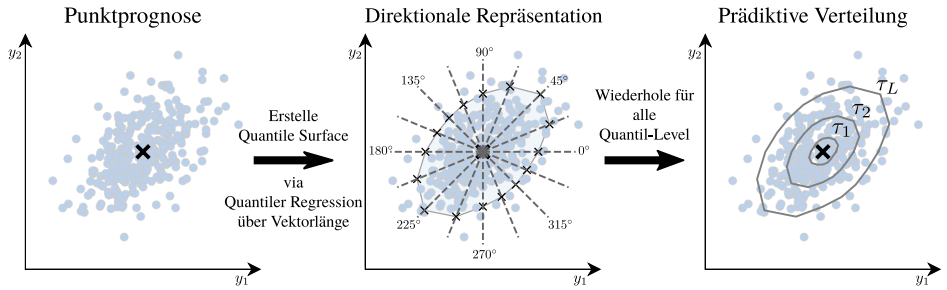


Abb. 3: Illustration der Quantile Surface Vorhersagemethode: Deterministische Punktprognose und Erstellung der Quantile Surfaces mittels direktionaler Repräsentation.

gibt, stellt die Erweiterung von Quantilen und damit der QR Methode auf multivariate Zielgrößen eine große Herausforderung dar.

Die neue Methode, genannt Quantile Surface (QS) Vorhersage, besteht aus zwei Schritten. Im ersten Schritt wird eine deterministische Punktprognose für die zukünftige Trajektorie erstellt. Dieses Netzwerk prognostiziert auf Basis von Eingabemerkmalen (hier die vergangene Radfahrertrajektorie) das Lagemaß (z.B. den geometrischen Median) der vorherzusagenden Verteilung. Im zweiten Schritt werden jetzt ausgehend von dieser Punktprognose Ränder von Gebieten (den sogenannten Quantile Surfaces) geschätzt, die einen bestimmten Träger (engl. Support) der Verteilung beinhalten, z.B. 95 % der zu erwartenden Beobachtungen. Eine Beobachtung entspricht hier einer tatsächlich beobachteten Radfahrerposition im Vorhersagezeitraum. Die Kernidee ist die Verwendung einer sogenannten direktionalen Repräsentation der Beobachtungen ausgehend von der deterministischen Punktprognose. Im zweidimensionalen Fall entspricht diese Art der Repräsentation einer Darstellung mittels Polarkoordinaten. Die Quantile Surface für ein bestimmtes Quantile-Level ergibt sich durch die Anwendung der QR Methode über die Vektorlänge des Vektors von der deterministischen Punktprognose zur Beobachtung, d.h. es wird versucht das Ausmaß des Gebietes in eine bestimmte Richtung zu beschreiben. Die komplette prädizierte Verteilung ergibt sich aus der Anwendung dieser Methodik für unterschiedliche Richtungen und diskrete Quantil-Level τ, z.B. $5\%, 10\% \ldots, 95\%$. Das Verfahren ist exemplarisch in Abb.3 dargestellt.

Für die QR wird kein lineares Modell verwendet, sondern ein neuronales Netzwerk mit nicht-linearen Aktivierungsfunktionen. Dieses Netzwerk bekommt als Eingabe die vergangene Trajektorie des Radfahrers sowie die Richtung, für welche das Ausmaß der QS bestimmt werden soll. Die Ausgabe des Netzwerks sind diskrete Quantil-Level für das Ausmaß der QS in die jeweilige Richtung. Die Parameter des Netzwerkes werden mittels der Quantile Fehlerfunktion (auch bekannt als Pinnball Fehlerfunktion) optimiert. Mit Hilfe des QS Verfahrens können zuverlässige und zugleich scharfe Prognosen für die zukünftige Radfahrertrajektorie erstellt werden. Im Rahmen eine Studie mit 107 Radfahrern erreicht die QS-Vorhersagemethode eine 99% Verbesserung bzgl. Güte der probabilistischen Vorhersage (gemessenen mittels des Continuous Ranked Probability Scores) verglichen zu klassischen Methoden.

5 Radfahrer als zusätzliche Sensoren

Heutzutage trägt fast jeder Mensch ein Smartphone oder anderes Wearable Device bei sich. Es gibt zahlreiche Studien, welche die Verwendung von Smart Devices für die Aktivitätserkennung (d.h. die Erkennung der aktuellen Bewegung) von Fußgängern einsetzten (siehe [SvSM17] für eine detaillierte Auflistung). Die Verwendung von Smart Devices für die Absichtserkennung (d.h. auch die Trajektorienvorhersage) von Radfahrern wurde jedoch bisher nur wenig untersucht. Im Rahmen dieser Arbeit wird diese Forschungslücken geschlossen und zahlreiche Aspekte der Smart Device gestützten Radfahrer-Absichtserkennung untersucht, darunter die Smart Device gestützte Bewegungszustandsdetektion und Trajektorienvorhersage sowie insbesondere auch die Untersuchung des Einflusses der Trageposition des Gerätes auf die Erkennungsleistung.

Smart Device gestützte Lokalisierung und Geräte-Trageposition Im Rahmen einer Fallstudie wird gezeigt, dass die Positionierungsgenauigkeit aktueller Smart Devices (d.h. handelsüblicher Android Smartphones) noch nicht ausreichend ist, um eine rein GPS-gestützte Absichtserkennung durchzuführen. Insbesondere bei Geräten, welche an Positionen getragen werden, bei denen das GPS-Signal stark durch den Körper abgeschirmt wird (z.B. der Hosentasche), ist die Positionsgenauigkeit noch nicht ausreichend. Andere Geräte, welche an mehr exponierten Positionen getragen werden, z.B. am Handgelenk oder auf dem Helm, zeigen deutlich bessere Ergebnisse. Zudem wird auch gezeigt, dass die Geräte-Trageposition mittels maschineller Lernverfahren auf Basis der Inertialsensoren zuverlässig klassifiziert werden kann [Bi18]. Diese Information kann zur nachträglichen Bewertung der Lokalisierungsgenauigkeit verwendet werden.

Bewegungszustandsdetektion und Trajektorienvorhersage Ein zentraler Beitrag für die Einbindung von Radfahrern als zusätzlichen Sensoren ist ein neuartiger Prozess zur schnellen und zuverlässigen Bewegungszustandserkennung unter Verwendung der Inertialsensorik (d.h. Accelerometer und Gyroskop). Herzstück dieses Prozesses sind Methoden der menschlichen Aktivitätserkennung (engl. Human Activity Recongition, kurz HAR) in Kombination mit Methoden des maschinellen Lernens [BSS18]. Im Rahmen der Untersuchung mit 107 Radfahrern wird gezeigt, dass die Ergebnisse dieser neuartigen Detektionsmethode vergleichbar sind mit denen eines rein Fahrzeugkamera gestützten Ansatzes. Die Gerät-Trageposition hat auch hier einen starken Einfluss auf die Detektion, so eignet sich bspw. das Smartphone, welches in der Hosentasche getragen wird, besonders gut für die Radfahrer-Anfahrtserkennung. Der Ansatz zur Bewegungszustandserkennung wird durch einige Modifikationen, darunter der Verwendung eines neuronalen Netzwerks, auch für die Trajektorienprädiktion weiterentwickelt. Nach bestem Wissen und Gewissen ist dies das erste Mal, dass der Einsatz von Smart Devices für die Radfahrer-Trajektorienprädiktion untersucht wird. Die Ergebnisse hier sind nicht vergleichbar mit denen eines rein Fahrzeugkamera gestützten Ansatzes, kommen diesem aber in einigen Aspekten (z.B. die Prädikationen von abbiegenden Radfahrern) bereits sehr nahe. Als ein limitierender Faktor stellt sich hierbei insbesondere die Lokalisierungsgenauigkeit der Smart Devices heraus.

Verwendung mehrerer kooperierender Smart Devices und Smarthelmet In Zukunft werden die Menschen anstatt eines einzelnen Smartphones eine Vielzahl von smarten End-

geräten mit sich führen. Dies reicht von Smartwatches über mit Sensorik ausgestattete Kleidung bis hin zu intelligenten Fahrrädern sowie intelligente mit Sensorik ausgestatteten Helmen (engl. Smarthelmet). Im Rahmen dieser Arbeit wird ein neuartiger Ansatz skizziert in dem diese Geräte durch ein Fahrrad Ad Hoc Netzwerk (engl. Bicycle Area Network kurz BicAN) miteinander verbunden sind, Informationen austauschen und kooperieren. Im Rahmen einer umfangreichen Fallstudie mit 51 Radfahrern und drei Smart Devices, d.h. einem Smartphone in der Hosentasche, einer Smartwatch am Handgelenk und einem Smarthelmet, kann gezeigt werden, dass die Kombination der Informationen aller Geräte zu einer deutlich verbesserten Bewegungszustandsdetektion führt. Darüber hinaus wird das Potenzial eines Smarthelmets im Kontext einer verbesserten Eigenlokalisierung und Orientierungsschätzung für die Radfahrer-Absichtserkennung aufgezeigt.

6 Kooperative Methoden zur Absichtserkennung

Im Rahmen dieser Arbeit stellen werden drei Ansätze zur kooperativen Radfahrer-Absichtserkennung vorgestellt: 1.) Fusion mittels eines kooperativ-erstellten probabilistischen Modells der vergangenen Trajektorie des Radfahrers, 2.) Kooperation zur verbesserten Detektion des aktuellen Bewegungszustands und 3.) Kooperation auf Ebene von vorhergesagten Trajektorien. Um die Ansätze evaluieren zu können, wurde im Rahmen dieser Arbeit in drei großangelegten Messkampagnen Daten von 107 Radfahrern an einer öffentlichen Kreuzung in Aschaffenburg aufgezeichnet. Von jedem dieser Radfahrer liegen Daten sowohl von Verkehrskameras, als auch Daten aus einem Versuchsträger sowie auch von den vom Radfahrer bei sich getragenen Smart Devices vor. Die Evaluation der unterschiedlichen Verfahren findet offline mit den erhobenen Daten statt. Ein weiterer wesentlicher Beitrag dieser Arbeit ist die Bewertung der unterschiedlichen Verfahren hinsichtlich ihrer Umsetzbarkeit mit den im aktuellen ETSI Standard definierten Nachrichtentypen.

Probabilistische Trajektorienfusion mittels Orthogonaler Polynome Die zugrundeliegende Idee dieses kooperativen Ansatzes ist, dass alle Agenten die aktuelle Position der beobachteten Radfahrer teilen. Die vergangene Trajektorie eines Radfahrers bzw. die Messungen der Positionen aller Agenten über dessen Position wird durch ein Polynom mit orthogonalen Basisfunktionen approximiert. Die Fusion wird somit durch die Approximation der Positionsmessungen mittels eines Polynoms realisiert. Die Koeffizient des orthogonalen Basispolynome des approximierenden Polynoms werden als Merkmale für die Absichtserkennung verwendet (siehe [Go16]). Die Stärke des Ansatzes im Vergleich zu klassischen Verfahren wie Kalman Filtern ist, dass dieser Ansatz ohne Modifikationen mit zeitlich verzögerten Messungen sowie außerhalb der Sequenz ankommende Messungen umgehen kann. Auf Grund der methodischen Nähe zu kooperativen Tracking-Verfahren ist dieses Verfahren mit den Nachrichtentypen, welche im ETSI Standard definiert sind, bereits heute umsetzbar.

Kooperative Bewegungszustandsdetektion Das Ziel der kooperativen Bewegungszustandsdetektion ist es durch die Fusion der Informationen und Messungen von mehreren Agenten die Bewegungszustandsdetektion zu verbessern. Hierfür werden unterschiedliche Methoden untersucht, u.a. die Konkatenation des Eingaberaums mehrere Agenten, ein

Bayesscher Ansatz zur Fusion von Entscheidungen und Methoden des maschinellen Lernens (z.B. das Anlernen von Ensembles). Kooperation auf Ebene der Detektion der Bewegungszustände stellt sich als äußerst vielversprechend heraus, so kann durch die Fusion die Robustheit der Bewegungszustandsdetektion um bis zu 43% im Vergleich zu einem nicht-kooperativen Ansatz gesteigert werden. Zudem werden auch Übergänge zwischen Bewegungszuständen deutlich schneller erkannt. Da für die Kooperation auf Basis der Bewegungszustandsdetektion keine exakte Positionierungsinformation notwendig ist, stellt diese Art der Kooperation eine hervorragende Möglichkeit dar, um Smart Devices zu integrieren. Für die praktische Umsetzung der kooperativen Bewegungszustandsdetektion sind lediglich kleinere Erweiterung des ETSI Standards notwendig.

Kooperative Trajektorienvorhersage Bei diesem Ansatz handelt es um die Fusion auf Ebene der Trajektorienvorhersagen, d.h. die Agenten tauschen Vorhersagen aus. Zur Fusion wird ein neuartiger Ansatz auf Basis des Coopetitive Soft Gating Ensembles (kurz CSGE) [De18] entwickelt und untersucht. Die Vorhersage von unterschiedlichen Agenten werden auf Basis von drei Komponenten (einer globalen, einer situationsabhängigen und einer temporalen Komponente) bewertet. Auf Basis dieser Bewertung wird eine Gewichtung für die Fusion der Vorhersage der Agenten ermittelt. Das CSGE ist um die Fähigkeit erweitert worden zeitliche verzögerte Vorhersage (z.B. auf Grund eines Kommunikation Verzögerungen) mit zu berücksichtigen. Die kooperative Trajektorienvorhersage mittels des CSGE erweist hinsichtlich des Vorhersagefehlers als der beste Ansatz. Für dieses Verfahren kann sogar gezeigt werden, dass im Vergleich zu einem nicht-kooperativen Ansatz ein statistisch signifikanter besserer durchschnittlicher Rang (bzgl. des Vorhersagefehlers) erzielt wird. Für die praktische Umsetzung des Verfahrens im realen Verkehr sind deutliche umfangreichere Erweiterung des ETSI Standards nötig.

7 Zusammenfassung und Fazit

In dieser Arbeit wird ein opportunistischer Ansatz zur kooperativen Absichtserkennung von Radfahrern vorgestellt [Bi21]. Dabei werden unterschiedliche neue Aspekte betrachtet, darunter eine neuartige probabilistische Vorhersagemethode, unterschiedlichste Verfahren für die Verwendung von Radfahrern als zusätzliche Sensoren und zu guter letzte zahlreiche Verfahren für die kooperative Absichtserkennung auf diversen Ebenen. Die Ergebnisse dieser Untersuchungen sind vielversprechend und haben das Potential in Zukunft das Radfahrern sicherer zu machen. Die nächsten Schritte für zukünftige Forschung sind größere Messkampagnen und Versuchsreihen in denen die Realisierung der Verfahren im online Betrieb mit mehreren Fahrzeugen, RSU und Radfahrern mit Smart Devices untersucht wird.

Literaturverzeichnis

[Bi17] Bieshaar, M.; Reitberger, G.; Zernetsch, S.; Sick, B.; Fuchs, E.; Doll, K.: Detecting Intentions of Vulnerable Road Users Based on Collective Intelligence. In: AAET - Automatisiertes und vernetztes Fahren. Braunschweig, Germany, S. 67–87, 2017.

[Bi18] Bieshaar, M.: Organic Computing: Doctoral Dissertation Colloquium. Kapitel Where is my Device? - Detecting the Smart Device's Wearing Location in the Context of Active Safety for Vulnerable Road Users, S. 27–37. Kassel University Press, 2018.

[Bi21] Bieshaar, M.: Cooperative Intention Detection using Machine Learning–Advanced Cyclist Protection in the Context of Automated Driving. Intelligent Embedded Systems. Kassel University Press, 2021. (Dissertation, Universität Kassel, Fachbereich Elektrotechnik/Informatik).

[BSS18] Bieshaar, M.and Depping, M.; Schneegans, J.; Sick, B.: Starting Movement Detection of Cyclists using Smart Devices. In: International Conference on Data Science and Advanced Analytics (DSAA). Turin, Italy, S. 313–322, 2018.

[De18] Deist, S.; Bieshaar, M.; Schreiber, J.; Gensler, A.; Sick, B.: Coopetitive Soft Gating Ensemble. In: Workshop on Self-Improving System Integration (SISSY). Trento, Italy, 2018.

[Ei17] Eilbrecht, J.; Bieshaar, M.; Zernetsch, S.; Doll, K.; Sick, B.; Stursberg, O.: Model-predictive planning for autonomous vehicles anticipating intentions of vulnerable road users by artificial neural networks. In: Symposium Series on Computational Intelligence (SSCI). Honolulu, HI, S. 2869–2876, 2017.

[ES17] Eilbrecht, J.; Stursberg, O.: Cooperative driving using a hierarchy of mixed-integer programming and tracking control. In: Intelligent Vehicles Symposium (IV). Los Angeles, CA, S. 673–678, 2017.

[Go16] Goldhammer, M.: Selbstlernende Algorithmen zur videobasierten Absichtserkennung von Fußgängern. Intelligent Embedded Systems. Kassel University Press, 2016. (Dissertation, Universität Kassel, Fachbereich Elektrotechnik/Informatik).

[KB78] Koenker, R.; Bassett, G.: Regression Quantiles. Econometrica, 46:33–50, 1978.

[SvSM17] Scholliers, J.; van Sambeek, M.; Moerman, K.: Integration of vulnerable road users in cooperative ITS systems. European Transport Research Review (ETRR), 9(2):15, 2017.

[Wo18] World Health Organization: Global Status Report on Road Safety 2018. World Health Organization, Geneva, Switzerland, 2018.

[Ze19] Zernetsch, S.; Reichert, H.; Kress, V.; Doll, K.; Sick, B.: Trajectory Forecasts with Uncertainties of Vulnerable Road Users by Means of Neural Networks. In: Intelligent Vehicles Symposium (IV). Paris, France, S. 810–815, 2019.

Maarten Bieshaar wurde am 8. Mai 1990 in Fritzlar, Hessen geboren. Er schloss 2013 den B.Sc. und 2015 den M.Sc. in Informatik an der Universität Paderborn ab. Ab 2016 arbeitet er als wissenschaftlicher Mitarbeiter am Fachgebiet Intelligente Eingebettete Systeme (IES) unter der Leitung von Bernhard Sick an der Universität Kassel. Dort begann er ebenfalls im Jahr 2016 mit seiner Promotion in Informatik, welche er in 2020 mit Auszeichnung abschloss. Seit 2020 leitet er am Fachgebiet IES die Gruppe „AI for Motion“, welche sich mit KI-Themen im Automobilumfeld beschäftigt.

Maschinelles Lernen für Ressourcenplanung in Verteilten Systemen[1]

Michael Borkowski[2]

Abstract: Verteilte Rechensysteme sind aus der heutigen digitalen Welt nicht mehr wegzudenken: Suchmaschinen wie Google, Cloud-Speichersysteme wie Dropbox, Streaming-Dienste wie Netflix oder wissenschaftliche Großrechner führen komplexe Aufgaben auf verteilter IT-Infrastruktur aus. Dabei müssen entsprechende Systeme laufend Ressourcenoptimierung betreiben. Beispielsweise können durch Aktivierung von Ressourcen kurz vor Lastspitzen und anschließender Passivierung enorme Kostenersparnisse erzielt werden. Statt konventioneller Wenn-Dann-Beziehungen oder starrer Regelkreise beschreibe ich in meiner Dissertation adaptive und Vorhersage-basierte Techniken, wie sie in einer dynamischen Umgebung wie dem heutigen Internet unabdingbar sind. Hierfür verwende ich Modelle für maschinelles Lernen, insbesondere künstliche neuronale Netze und Kalman-Filter. Meine Ergebnisse zeigen, dass der Einsatz solcher Methoden Kosten und Ressourcenverbrauch senkt sowie die Verfügbarkeit und Verlässlichkeit der Systeme erhöht.

1 Einführung

Das Versagen eines verteilten Systems kann 66.240 US-Dollar kosten – pro Minute. So erlitt Amazon Web Services (AWS) im April 2011 einen Teilausfall, der binnen 12 Stunden einen Gesamtschaden von 48 Millionen US-Dollar verursachte [Ah17]. Zahlen wie diese verdeutlichen den heutigen wirtschaftlichen Stellenwert von verteilten Systemen. Sie machen deutlich, dass Ansätze zur Verbesserung der Systemstabilität und -performanz in der heutigen digitalen Welt unentbehrlich sind. Um ihre Systeme kostenoptimiert zu betreiben, können Betreiber auf Methoden des maschinellen Lernens zurückgreifen. Beispiele solcher Methoden entwickle und beschreibe ich in meiner Dissertation [Bo20].

Verteilte Systeme spielen eine entscheidende Rolle in vielen Aspekten der heutigen digitalen Infrastruktur und ermöglichen Cloud-Speicherlösungen [Gr15], Smart Cities [PLM17] oder das Internet of Things (IoT) [Bo16]. Die Forschung innerhalb der verteilten Systeme umfasst Bereiche wie Cloud Computing [Ar10], Data Stream Processing (DSP) [Ca18b], Business Process Management (BPM) [Sc15] oder dezentrale Consensus-Technologien wie Blockchains [Zh18]. Zu den gängigen Zielen gehören Elastizität, Flexibilität und Skalierbarkeit, während gleichzeitig Kosten [Sc15] oder Qualitätseinbußen [Ca18a] eingeschränkt werden sollen.

Moderne verteilte Systeme weisen aufgrund ihrer verteilten Architektur einen hohen Grad an Komplexität auf. Die für den Betrieb und die Wartung solcher Systeme erforderlichen Komponenten bilden eine unübersichtliche Landschaft, welche oft sehr heterogen ist, da

[1] Englischer Titel der Dissertation: „Predictive Approaches for Resource Provisioning in Distributed Systems“

[2] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Flugführung, michael.borkowski@dlr.de

einzelne Komponenten oft von verschiedenen Interessengruppen bereitgestellt und auf uneinheitlicher Hardware mit unterschiedlichen Technologien betrieben werden.

In vielen Situationen muss ein verteiltes System Ressourcenplanung durchführen: Ressourcen wie etwa Rechenleistung müssen in ausreichender – aber nicht überschüssiger – Menge zur Verfügung gestellt werden (Skalierung), Aufgaben müssen auf die verfügbaren Ressourcen aufgeteilt werden (Platzierung bzw. Allocation) und die zeitliche Abfolge der Ausführung einzelner Aufgaben muss bestimmt werden (Zeitplanung bzw. Scheduling).

In modernen verteilten System müssen vielfältige, sich oft kurzfristig ändernde Einsatzbereiche berücksichtigt werden. Es ist daher für einen kostenoptimierten Betrieb unbedingt notwendig, dass die Ressourcenplanung eines verteilten Systems nicht durch starre Regeln (etwa Last-Schwellwerte, die für das Hinzufügen bzw. Entfernen von Rechenressourcen verwendet werden) oder konventionelle Wenn-Dann-Regelwerke definiert werden, wie dies aktuell meist der Fall ist. Vielmehr müssen proaktiv, Vorhersage-basiert und automatisiert Entscheidungen zur Ressourcenplanung getroffen werden. Beispielsweise kann ein System bereits vor einem vorhergesagten Anstieg der Last Rechenressourcen hinzufügen, um Einbußen in der Verfügbarkeit zu vermeiden.

In meiner Dissertation [Bo20] präsentiere ich Vorhersage-basierte Ansätze zur Ressourcenplanung in verteilten Systemen, welche zur Senkung von Kosten und gleichzeitig zur Erhöhung der Systemstabilität und Verfügbarkeit führen. Die Ansätze verwenden Techniken aus dem Bereich des maschinellen Lernens (ML), insbesondere künstliche neuronale Netze (ANNs). Ich zeige mittels Vergleichen zum Stand der Technik, wie diese Ansätze zu einem optimierten Betrieb verschiedener Arten von verteilten Systemen beitragen.

2 Ressourcenplanung in Verteilten Systemen

Der Prozess der Ressourcenplanung, etwa die Zuweisung von Ressourcen zu bestimmten Aufgaben, ist eines der wichtigsten Gebiete in verteilten Systemen [Ha17]. Oft wird das Placement-Problem, also das Platzieren von Anwendungen auf Virtuellen Maschinen (VMs) oder VMs auf Physischen Maschinen (PMs), betrachtet [Ca16]. Die Ressource, auf der eine Aufgabe platziert wird, ist für eine bestimmte Zeit und in einem bestimmten Ausmaß belegt, weswegen die genaue Zuordnung zwischen Ressourcen und Aufgaben einen erheblichen Einfluss auf die gesamte Ressourcenauslastung hat. Dies wirkt sich auf die Betriebskosten und die Leistung des gesamten Systems aus [CLN12].

Zusätzlich zum Finden einer geeigneten Platzierung müssen moderne verteilte Systeme auch Elastizität sicherstellen [Du11, LBMAL14], was durch eine dynamische und automatische Anpassung an Änderungen der Arbeitslast sowie Aktivierung und Passivierung von Ressourcen erreicht wird. Diese Fähigkeit wird als Skalierbarkeit bezeichnet [HKR13]. Die Entscheidung, wann und wie skaliert werden soll ist ein nicht-triviales Problem, und oft gibt es mehrere widersprüchliche Optimierungszielmetriken [Co13]. Beispielsweise kann ein Cloud-Computing-Anbieter daran interessiert sein, die Betriebskosten seiner Infrastruktur zu reduzieren, muss aber gleichzeitig eine bestimmte Dienstgüte (engl. Quality

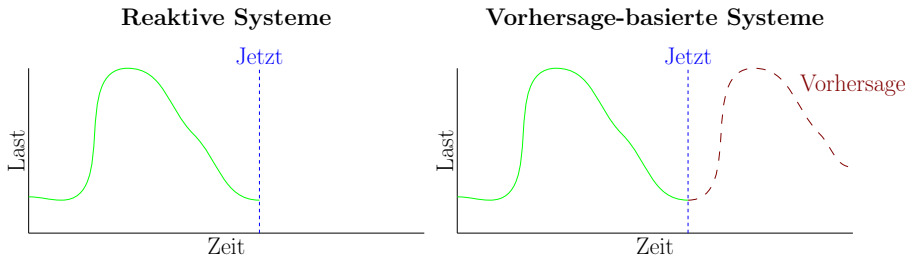


Abb. 1: Reaktive und Vorhersage-basierte Verfahren und ihr Umgang mit Werten einer Systemmetrik

of Service – QoS) einhalten, um bestimmte Leistungsvereinbarungen (engl. Service Level Agreements – SLAs) zu erfüllen und somit Strafzahlungen zu vermeiden [Ca18a].

Gängige Lösungen zur Ressourcenplanung können in reaktive sowie proaktive Ansätze unterteilt werden [LBMAL14], wie in Abbildung 1 illustriert ist. Reaktive Ansätze zeichnen sich dadurch aus, dass eine Metrik beobachtet wird, und aufgrund des aktuellen Werts Entscheidungen getroffen werden. Ein klassisches Beispiel hierfür ist ein Schwellwert-gesteuertes Skalierungsverhalten, bei dem zusätzliche Rechenleistung aktiviert wird, wenn die Systemlast einen vordefinierten oberen Schwellwert überschreitet, und wieder passiviert wird, wenn die Systemlast den unteren Schwellwert unterschreitet [LBMAL14]. Im Gegensatz dazu zeichnen sich proaktive, insbesondere Vorhersage-basierte Ansätze dadurch aus, dass nicht nur der aktuelle Wert einer Metrik, sondern auch eine Vorhersage über ihren zukünftigen Wert getroffen wird. Anhand dieser Vorhersage kann proaktiv auf mögliche zukünftige Veränderungen reagiert werden, etwa in Form von zusätzlichen Rechenressourcen bereits vor einer Lastspitze [LBMAL14].

In meiner Dissertation [Bo20] untersuche ich Vorhersage-basierte Ansätze zur Ressourcenplanung in verteilten Systemen. Insbesondere schaffe ich Grundlagen zum Treffen von Vorhersagen in diesem Bereich, wie etwa die Auswahl an Methoden und Werkzeugen, die in verschiedenen Situationen zu brauchbaren Vorhersagen führen. Ich stelle dar, wie diese Methoden auf verschiedene Probleme und Anwendungsfälle in modernen verteilten Systemen angewendet werden können, und bewerte ihre Leistung – also die Auswirkung auf die Betriebskosten und die Leistungsfähigkeit der Systeme – quantitativ.

In meiner Arbeit betrachte ich dabei verschiedene Arten von verteilten Systemen, insbesondere Cloud Computing [BSH16], BPM-Systeme [Bo19a], DSP [BHS19] und Blockchain-Technologien [Bo19b]. Ich untersuche verschiedene Aspekte von Vorhersage-basierten Ansätzen, darunter die Vorhersage und Filterung von Metrikwerten sowie die Abschätzung der Eintrittswahrscheinlichkeit von Fehlerzuständen in Geschäftsprozessen.

3 Maschinelles Lernen, Neuronale Netze und Kalman-Filter

In den vorgeschlagenen Ansätzen zur Vorhersage-basierten Ressourcenplanung verwende ich Methoden des ML, also des maschinellen Erlernens von Mustern und Regelmäßigkeiten aus Trainingsdaten, ohne dass dabei durch menschliches Zutun diese Muster explizit als

Regeln definiert werden müssen. Dies setzt sich deutlich von regelbasierten Methodiken ab, da bei Anwendung von ML im Allgemeinen das notwendige Fachwissen über die zugrundeliegenden Charakteristika entfällt [Hu14]. Stattdessen erlernt das System diese Charakteristika anhand von beispielhaften Daten. Es existiert eine Vielfalt an Möglichkeiten, ein solches System zu trainieren, und es kann zwischen überwachtem, teilüberwachtem und unüberwachtem Lernen [SA13] sowie zwischen Online- und Offline-Lernen unterschieden werden [BDKM97].

Insbesondere ANN sind ein weitläufig verwendetes Werkzeug bei der Vorhersage von Werten anhand von Trainingsdaten [Ha98]. ANN bestehen zumeist aus mehreren Schichten sogenannter künstlicher Neuronen, die miteinander verbunden sind. Beispielhaft ist dies in Abbildung 2 dargestellt. Ein ANN nimmt einen Vektor an Werten als Eingabe an der Eingabeschicht an, und verwendet oft mehrere verborgene Schichten und schlussendlich die Ausgabeschicht, um je nach Anwendungsfall einen oder mehrere Ausgabewerte zu erzeugen. Die Ausgabe eines ANN stellt dabei die Vorhersage auf Basis des zur Verfügung gestellten Eingabevektors dar.

Ein weiteres wichtiges Werkzeug im Kontext von ML sind Kalman-Filter (KF) [KB61]. Im Wesentlichen dienen KF dazu, Daten aus mehreren, verschiedenartigen Sensorquellen mit verschiedenen und zeitlich variablen Unschärfen zu einer gemeinsamen Messgröße zusammenzufassen, und stellen dabei auch die Unschärfe dieser Messgröße zur Verfügung. Zur Anwendung von KF wird ein System mittels eines Zustandsvektors definiert. Jeder Einzelwert des Zustandsvektors stellt dabei eine Systemgröße dar, die von Interesse ist, wie etwa CPU- und RAM-Auslastung bei einem verteilten System [Gu12, CFF14]. Außerdem wird ein Eingabevektor definiert, welcher die Umgebung des beobachteten Systems bzw. dessen Eingabe (darauf wirkende Einflüsse) definiert. Jeder Einzelwert des Eingabevektors stellt dabei einen äußeren Einfluss auf das System dar, wie etwa die Menge der zu verarbeitenden Daten. Darüber hinaus wird eine Zustandsübergangsmatrix definiert, anhand der aus dem aktuellen Zustands- und Eingabevektor ein neuer Zustandsvektor erzeugt wird. Diese Matrix definiert die Systemdynamik. Während reguläre KF rein lineare Übergangsmatrizen verwenden, verallgemeinert der erweiterte KF (EKF) den Ansatz für nichtlineare Modelle [Ja07]. Anstelle von Matrizen verwenden EKF Funktionen als

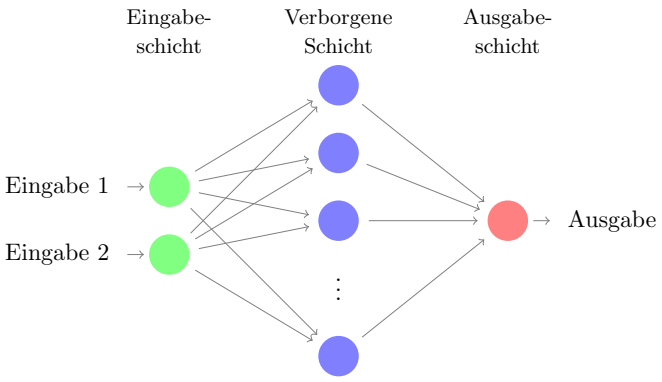


Abb. 2: Skizze eines mehrschichtigen ANN, vgl. [Bo20]

Übergangsmodelle, um so nichtlineare Systemdynamiken abzubilden. Eine Besonderheit von KF ist es, dass der tatsächliche Systemzustand nicht direkt bekannt sein muss, sondern nur indirekt über Beobachtungen abgeleitet werden kann. Dabei können sowohl die Beobachtungen als auch der Systemzustand Rauschen unterliegen, welches vom KF modelliert und als Ergebnis ausgegeben wird.

KF arbeiten somit in Zyklen, bei denen aus einer bekannten Eingabe und einer bekannten Beobachtung der (grundsätzlich unbekannte) innere Systemzustand ableitbar gemacht wird. Abbildung 3 zeigt exemplarisch einen Ausschnitt aus dem Zyklusdurchlauf im Systemmodell von KF, bei denen die Eingabe, der Zustand sowie die Übergangsfunktion f und Messfunktion h zu sehen sind. Eine detailliertere Beschreibung samt Formeln des KF-Modells findet sich in meiner Dissertation [Bo20].

4 Ansätze und Ergebnisse

Wie in Aschnitt 1 beschrieben, existiert eine große Vielfalt an Arten von verteilten Systemen in verschiedenen Bereichen, beispielsweise Cloud Computing [Ar10], BPM [Sc15] oder DSP [Ca18b]. Jeder dieser Bereiche bringt ein eigenes Anforderungsprofil an Vorhersage-basierten Ansätze mit sich, außerdem finden sich in jedem der Bereiche je nach Situation verschiedenartige Anwendungsfälle.

Verteilte Systeme müssen daher differenziert betrachtet werden. In meiner Dissertation beleuchte ich zunächst, wie diversen Anforderungen an verteilte Systeme mit verschiedenen Methoden entgegengetreten werden kann. So gibt es etwa Anwendungsfälle, in denen Vorhersagen zeitverzögert getroffen werden kann, und Offline-Lernen das Mittel der Wahl darstellt [BDKM97]. In anderen Situationen ist die Vorhersage hingegen bereits während der Ausführung des Systems notwendig – in diesen Fällen müssen Online-Learning-Techniken angewendet werden [La96]. Eine andere bedeutende Hürde beim maschinellen Lernen

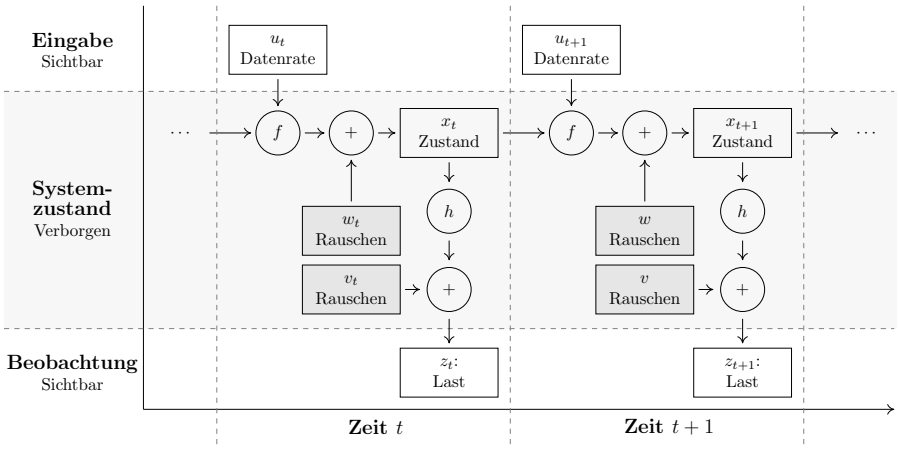


Abb. 3: Übersicht des Systemmodells in KF, vgl. [Bo20]

besteht in der Verfügbarkeit von Trainingsdaten. Wie oben beschrieben wird zwischen überwachtem, teilüberwachtem und unüberwachtem Lernen unterschieden [SA13], und die anzuwendende Methode hängt unter anderem von der Art der zugrundeliegenden Daten ab. Wenn diese Daten als sogenannte beschriftete Daten (labelled data) vorliegen, also als Eingabedaten gemeinsam mit den Soll-Ausgabedaten, werden überwachte Trainingstechniken (supervised learning) eingesetzt. Wenn hingegen keine oder nur begrenzte Soll-Ausgabedaten vorliegen, werden entweder auf teilüberwachte (semi-supervised) oder unüberwachte (unsupervised) Techniken angewendet.

In meiner Dissertation beschreibe ich daher einige Anwendungsfälle für maschinelles Lernen in verteilten Systemen, welche im Folgenden kurz umrissen werden, und stelle Lösungsmöglichkeiten vor. Ein Überblick über die einzelnen Beiträge findet sich in Tabelle 1. Ich variiere dabei sowohl die Domäne (Cloud Computing, BPMS, DSP) als auch die Zielaufgabe (Vorhersage des Ressourcenverbrauchs, Vorhersage von Fehlerzuständen, Vorhersage der Systemlast).

Tab. 1: Übersicht über die Beiträge der Dissertation zu Anwendungsfällen des maschinellen Lernens im Bereich der verteilten Systeme

Domäne	Methode	Ergebnis	Referenz
Cloud Computing	Regression mittels ANNs	Verbesserung d. Abweichung: 23 %	[BSH16]
EBS/BPMS	Klassifikation mittels ANNs	Präzision d. Vorhersage: 87 %	[Bo19a]
DSP	Skalierung mittels EKF	Verminderung d. Skalierungen: 88 %	[BHS19]

4.1 Vorhersage von Ressourcenverbrauch im Cloud Computing

Cloud Computing bezeichnet die Bereitstellung von IT-Ressourcen auf Bedarfsbasis in Form von Cloud-Diensten [Ar10], also als Dienste, welche bei Bedarf abgerufen und nach Verbrauch verrechnet werden. Der Anbieter eines Cloud-Dienstes muss eine Infrastruktur betreiben, die für Konsumenten zur Verfügung steht. Entsprechend liegt ein Hauptaugenmerk von Cloud-Anbietern darauf, diese Infrastruktur kostengünstig zu betreiben. Im Vordergrund stehen hierbei die Platzierung von Aufgaben (also Anforderungen von Rechenressourcen) auf Infrastrukturressourcen (etwa VMs) sowie die Skalierung, also die Aktivierung und Passivierung von Infrastruktur auf Basis des Bedarfs. Für diese beiden Vorgänge ist es essenziell, den Ressourcenbedarf der eintreffenden Aufgaben abzuschätzen, um diese Aufgaben optimal auf die bestehende bzw. neu zu aktivierende Infrastruktur zu verteilen.

Zu diesem Zweck präsentiere ich in Kapitel 3 meiner Dissertation sowie in der dazugehörigen Publikation [BSH16], wie ANNs zur Regression eingesetzt werden können. Vorhergesagt wird hierbei der Verbrauch von Systemressourcen für jede eintreffende Aufgabe auf Basis vergangener Aufgabenausführungen. Zur Evaluierung verwende ich einen Datensatz, welcher von Travis CI, einem Cloud-basierten Anbieter von Continuous Integration, gesammelt wurde. Vorhergesagt wird die benötigte Zeit sowie CPU-Auslastung.

Mittels mehrschichtigen ANNs erreiche ich im Median eine Verkleinerung der Vorhersageabweichung von 20 % (Zeit) bzw. 23 % (CPU-Auslastung).

4.2 Abschätzung der Fehlereintrittswahrscheinlichkeiten in Geschäftsprozessen

Geschäftsprozesse stellen in unserer vernetzten Welt ein wichtiges Werkzeug für große Unternehmen dar, bei dem Abläufe definiert und oft auch automatisiert werden. Die Abläufe können bei komplexen Prozessen vielfache Verzweigungen und Parallelitäten aufweisen und verschiedene Geschäftspartner umfassen [Fd12].

In Kapitel 4 meiner Dissertation sowie der dazugehörigen Publikation [Bo19a] stelle ich vor, wie aus Ereignissen, die während eines Geschäftsprozesses auftreten, die Eintrittswahrscheinlichkeit von Fehlern für verschiedene Prozessschritte abgeschätzt, Fehler also vorhergesagt werden können. Eine (hinreichend wahrscheinliche) Vorhersage von Fehlerzuständen kann bei Geschäftsprozessen dazu dienen, dass Geschäftspartner schon vorzeitig reagieren können, indem entweder vorbeugende Maßnahmen getroffen oder neue Prozessinstanzen gestartet werden. Als Beispiel dient die Lieferung von verderblichen Gütern, bei der aufgrund von Ereignissen (etwa zu hoher Temperatur im Kühlcontainer) ein Fehler (eine mangelhafte Lieferung) vorhergesagt werden kann.

Ich präsentiere einen Ansatz, der mittels rekurrenten und gefalteten ANNs (recurrent and convolutional ANNs) die während eines Geschäftsprozesses auftretenden Ereignisse konsolidiert und für jeden zukünftigen Prozessschritt eine Fehlerwahrscheinlichkeit generiert. Hierbei präsentiere ich auch Optimierungstechniken zur Eingrenzung des Suchraums, um etwa mit Prozessschleifen umzugehen und lange Laufzeiten der Vorhersage zu verhindern. Die Aggregation der Ergebnisse zeigt eine Präzision von 87 % in der Fehlervorhersage.

4.3 Reduktion von Skaliervorgängen im Data Stream Processing

DSP-Systeme zeichnen sich oft durch variierende Last aus und müssen demnach auch Skalierbarkeit und Elastizität aufweisen [AdSVB18]. Ein dabei auftretendes Problem liegt in den indirekten Kosten, die durch Skaliervorgänge entstehen: Insbesondere das Aktivieren von Ressourcen kann Kosten verursachen, etwa das Hochfahren von zusätzlichen VMs [GGW10, MH11]. Somit sind Skaliervorgänge, auch wenn sie für Elastizität notwendig sind, auf das notwendige Minimum zu begrenzen. Im Kapitel 5 meiner Arbeit sowie in der dazugehörigen Publikation [BHS19] betrachte ich ein DSP-System mit seiner Variabilität als dynamisches System, von dem die Systemlast gemessen wird, kurzzeitige Schwankungen hierbei allerdings als unerwünschtes Rauschen herausgerechnet werden sollen. Dies trägt zu einer besseren Vorhersage der tatsächlichen Systemlast bei.

Ich verwende hierfür eine EKF-basierte Methodik und beschreibe, wie diese in einem DSP-System angewendet werden kann. Für die Evaluierung verwende ich unter anderem einen Echtdatensatz aus einem biomedizinischen Labor, bei dem Mikroskopaufnahmen für Tissue Engineering verarbeitet werden. Die variable Last wurde aus den im Labor anfallenden Datenmengen übernommen, um die Experimente möglichst nahe an der Realität

zu halten. Ich erreiche in meinen Experimenten unter anderem eine Verminderung der Skaliervorgänge um bis zu 88 %, führe aber auch eine Kostenanalyse durch, in der ich detailliert beschreibe, in welchen Fällen der EKF-basierte Ansatz vorteilhaft ist.

5 Fazit

In meiner Dissertation zeige ich, dass der Einsatz von Vorhersage-basierten Ansätzen für Ressourcenplanung in verteilten Systemen die Performanz und Verlässlichkeit der Systeme deutlich steigert. Die Verwendung von maschinellem Lernen führt dabei zu einer maßgeblichen Verbesserung der Wirtschaftlichkeit. Ich präsentiere konkrete Ansätze, die auf dynamische, selbstlernende und autonome Art und Weise verteilten Systemen ermöglichen, proaktiv auf Lastwechsel und eine sich ändernde Umgebung zu reagieren. Die gemessenen Werte zeigen eine Verbesserung von bis zu 88 % im Vergleich zum Stand der Technik.

Literaturverzeichnis

[AdSVB18] Assunção, Marcos Dias; da Silva Veith, Alexandre; Buyya, Rajkumar: Distributed Data Stream Processing and Edge Computing: A Survey on Resource Elasticity and Future Directions. Journal of Network and Computer Applications, 103:1–17, 2018.

[Ah17] Ahmad, Waqar; Hasan, Osman; Pervez, Usman; Qadir, Junaid: Reliability modeling and analysis of communication networks. Journal of Network and Computer Applications, 78:191–215, 2017.

[Ar10] Armbrust, Michael; Fox, Armando; Griffith, Rean; Joseph, Anthony D.; Katz, Randy; Konwinski, Andy; Lee, Gunho; Patterson, David; Rabkin, Ariel; Stoica, Ion; Zaharia, Matei: A View of Cloud Computing. Communications of the ACM, 53(4):50–58, 2010.

[BDKM97] Ben-David, Shai; Kushilevitz, Eyal; Mansour, Yishay: Online Learning versus Offline Learning. Machine Learning, 29(1):45–63, 1997.

[BHS19] Borkowski, Michael; Hochreiner, Christoph; Schulte, Stefan: Minimizing Cost by Reducing Scaling Operations in Distributed Stream Processing. PVLDB, 12(7):724–737, 2019.

[Bo16] Botta, Alessio; de Donato, Walter; Persico, Valerio; Pescapé, Antonio: Integration of Cloud Computing and Internet of Things: A Survey. Future Generation Computer Systems, 56:684–700, 2016.

[Bo19a] Borkowski, Michael; Fdhila, Walid; Nardelli, Matteo; Rinderle-Ma, Stefanie; Schulte, Stefan: Event-Based Failure Prediction in Distributed Business Processes. Information Systems, 81:220–235, 2019.

[Bo19b] Borkowski, Michael; Sigwart, Marten; Frauenthaler, Philipp; Hukkinen, Taneli; Schulte, Stefan: DeXTT: Deterministic Cross-Blockchain Token Transfers. IEEE Access, 7(1):111030–111042, 2019.

[Bo20] Borkowski, Michael: Predictive Approaches for Resource Provisioning in Distributed Systems. Dissertation, TU Wien, 2020.

[BSH16] Borkowski, Michael; Schulte, Stefan; Hochreiner, Christoph: Predicting Cloud Resource Utilization. In: 9th IEEE/ACM International Conference on Utility and Cloud Computing (UCC). IEEE/ACM, S. 37–42, 2016.

[Ca16] Cardellini, Valeria; Grassi, Vincenzo; Lo Presti, Francesco; Nardelli, Matteo: Optimal Operator Placement for Distributed Stream Processing Applications. In: 10th ACM International Conference on Distributed and Event-based Systems (DEBS). ACM, S. 69–80, 2016.

[Ca18a] Cardellini, Valeria; Grbac, Tihana Galinac; Nardelli, Matteo; Tanković, Nikola; Truong, Hong-Linh: QoS-Based Elasticity for Service Chains in Distributed Edge Cloud Environments. In: Autonomous Control for a Reliable Internet of Services, S. 182–211. Springer, 2018.

[Ca18b] Cardellini, Valeria; Lo Presti, Francesco; Nardelli, Matteo; Russo Russo, Gabriele: Optimal Operator Deployment and Replication for Elastic Distributed Data Stream Processing. Concurrency and Computation: Practice and Experience, 30(9):article e4334, 2018.

[CFF14] Corradi, Antonio; Fanelli, Mario; Foschini, Luca: VM Consolidation: A Real Case Based on OpenStack Cloud. Future Generation Computer Systems, 32:118–127, 2014.

[CLN12] Chaisiri, Sivadon; Lee, Bu-Sung; Niyato, Dusit: Optimization of Resource Provisioning Cost in Cloud Computing. IEEE Transactions on Services Computing, 5(2):164–177, 2012.

[Co13] Copil, Georgiana; Moldovan, Daniel; Truong, Hong-Linh; Dustdar, Schahram: Multi-level Elasticity Control of Cloud Services. In: International Conference on Service-Oriented Computing (ICSOC). LNCS 8274. Springer, S. 429–436, 2013.

[Du11] Dustdar, Schahram; Guo, Yike; Satzger, Benjamin; Truong, Hong-Linh: Principles of Elastic Processes. IEEE Internet Computing, 15(5):66–71, 2011.

[Fd12] Fdhila, Walid; Rinderle-Ma, Stefanie; Baouab, Aymen; Perrin, Olivier; Godart, Claude: On Evolving Partitioned Web Service Orchestrations. In: 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA). S. 1–6, 2012.

[GGW10] Gong, Zhenhuan; Gu, Xiaohui; Wilkes, John: PRESS: Predictive Elastic Resource Scaling for Cloud Systems. In: International Conference on Network and Service Management (CNSM). IEEE, S. 9–16, 2010.

[Gr15] Gracia-Tinedo, Raúl; Tian, Yongchao; Sampé, Josep; Harkous, Hamza; Lenton, John; García-López, Pedro; Sánchez-Artigas, Marc; Vukolic, Marko: Dissecting UbuntuOne: Autopsy of a Global-Scale Personal Cloud Back-End. In: Internet Measurement Conference (IMC). ACM, S. 155–168, 2015.

[Gu12] Gulisano, Vincenzo; Jimenez-Peris, Ricardo; Patino-Martinez, Marta; Soriente, Claudio; Valduriez, Patrick: StreamCloud: An Elastic and Scalable Data Streaming System. IEEE Transactions on Parallel and Distributed Systems, 23(12):2351–2365, 2012.

[Ha98] Haykin, Simon: Neural Networks: A Comprehensive Foundation. Prentice Hall, 2nd edition. Auflage, 1998.

[Ha17] Hao, Fang; Kodialam, Murali; Lakshman, T. V.; Mukherjee, Sarit: Online Allocation of Virtual Machines in a Distributed Cloud. IEEE/ACM Transactions on Networking, 25(1):238–249, 2017.

[HKR13] Herbst, Nikolas Roman; Kounev, Samuel; Reussner, Ralf H: Elasticity in Cloud Computing: What It Is, and What It Is Not. In: 10th International Conference on Autonomic Computing (ICAC). Jgg. 13. USENIX, S. 23–27, 2013.

[Hu14] Huang, Gao; Song, Shiji; Gupta, Jatinder N. D.; Wu, Cheng: Semi-Supervised and Unsupervised Extreme Learning Machines. IEEE Transactions on Cybernetics, 44(12):2405–2417, 2014.

[Ja07] Jazwinski, Andrew H.: Stochastic Processes and Filtering Theory. Dover, 2007.

[KB61] Kalman, Rudolph E.; Bucy, Richard S.: New Results in Linear Filtering and Prediction Theory. Journal of Basic Engineering, 83(1):95–108, 1961.

[La96] Langley, Pat: Elements of Machine Learning. Morgan Kaufmann, 1996.

[LBMAL14] Lorido-Botran, Tania; Miguel-Alonso, José; Lozano, Jose Antonio: A Review of Auto-Scaling Techniques for Elastic Applications in Cloud Environments. Journal of Grid Computing, 12(4):559–592, 2014.

[MH11] Mao, Ming; Humphrey, Marty: Auto-Scaling to Minimize Cost and Meet Application Deadlines in Cloud Workflows. In: International Conference for High Performance Computing, Networking, Storage and Analysis (SC). IEEE, 2011. article 49.

[PLM17] Petrolo, Riccardo; Loscri, Valeria; Mitton, Nathalie: Towards a Smart City Based on Cloud of Things, a Survey on the Smart City Vision and Paradigms. Transactions on Emerging Telecommunications Technologies, 28(1):article e2931, 2017.

[SA13] Sathya, Ramadass; Abraham, Annamma: Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. International Journal of Advanced Research in Artificial Intelligence, 2(2):34–38, 2013.

[Sc15] Schulte, Stefan; Janiesch, Christian; Venugopal, Srikumar; Weber, Ingo; Hoenisch, Philipp: Elastic Business Process Management: State of the Art and Open Challenges for BPM in the Cloud. Future Generation Computer Systems, 46:36–50, 2015.

[Zh18] Zheng, Zibin; Xie, Shaoan; Dai, Hong-Ning; Chen, Xiangping; Wang, Huaimin: Blockchain Challenges and Opportunities: A Survey. International Journal of Web and Grid Services, 14(4):352–375, 2018.

Michael Borkowski wurde am 26. März 1991 in Wien geboren. Er schloss 2015 sein Diplomstudium im Bereich Software Engineering an der TU Wien ab. Von 2015 bis 2019 war er als Projektassistent und Doktorand an der Distributed Systems Group der TU Wien tätig, wo er in den Bereichen des maschinellen Lernens, der verteilten Rechensysteme sowie Blockchains forschte und 2020 mit Auszeichnung promovierte. Seit April 2019 ist er am Deutschen Zentrum für Luft- und Raumfahrt (DLR) in Braunschweig als wissenschaftlicher Mitarbeiter im Bereich Unbemannte Luftfahrzeugsysteme (UAS) tätig und wirkt unter anderem an zahlreichen H2020-EU-Projekten mit. Seine wissenschaftliche Arbeit umfasst 22 begutachtete Publikationen, darunter fünf Artikel in wissenschaftlichen Fachzeitschriften, unter anderem in Elsevier Information Systems sowie Proceedings of the VLDB Endowment und eine mit *Best Paper Award* ausgezeichnete Publikation bei der IEEE International Conference on Blockchain.

Formale Analyse von variantenreichen und kontextsensitiven Systemen[1]

Philipp Chrszon[2]

Abstract:

Moderne Informations- und Kommunikationssysteme sind zunehmend von Variantenreichtum und Dynamik geprägt. In der Softwaretechnologie wurden Konzepte wie Features und Rollen eingeführt, um die Variabilität innerhalb einer Systemfamilie bzw. kontextabhängige Adaptionen zu erfassen. Eine zentrale Herausforderung bei der Entwicklung featureorientierter und rollenbasierter Systeme stellen Interaktionen dar, d.h. emergentes Verhalten, welches sich aus der Kombination von Features bzw. Rollen ergibt. Das Ziel der Dissertation ist die Entwicklung von formalen Methoden, um eine frühzeitige Entdeckung von Interaktionen zu ermöglichen, welche die funktionalen und nicht-funktionalen Eigenschaften eines Systems beeinflussen. Dazu werden Formalismen, Modellierungssprachen und zugehörige Analysewerkzeuge entwickelt, welche die Konzepte von Features und Rollen explizit unterstützen.

1 Einführung

Informations- und Kommunikationstechnologie gewinnt noch immer stetig an Bedeutung, sowohl im täglichen Leben als auch in der Industrie. Sie erweitert die Möglichkeiten der Kommunikation, steigert die Produktivität, erlaubt ein immer höheres Maß der Automatisierung und berührt damit nahezu alle Bereiche des Alltags. Natürlich erhöht sich dadurch unsere Abhängigkeit von Hardware- und Softwaresystemen, insbesondere von deren Stabilität, Korrektheit und Qualität. Hardware- und Softwarefehler können einen hohen finanziellen Schaden verursachen und im schlimmsten Fall gar Menschenleben gefährden. Aber auch nicht-funktionale Eigenschaften wie z.B. der Energieverbrauch und die Zuverlässigkeit haben einen entscheidenden Einfluss auf die Wirtschaftlichkeit eines Systems. Diesen Qualitätsanforderungen bei der Entwicklung, dem Betrieb und der Wartung moderner Systeme gerecht zu werden, wird jedoch zunehmend schwieriger. Die Komplexität von Systemen wächst nicht nur durch immer weitreichendere Anforderungen und das wachsende Potential moderner Hardware, sondern auch in Folge von Vernetzung, Mobilität und Parallelität. Zudem müssen Systeme vermehrt an verschiedene Einsatzgebiete, Anforderungen, länderspezifische Regelungen sowie Hardwareplattformen angepasst werden, sodass nicht nur ein einzelnes System, sondern eine Vielzahl an Varianten eines Systems

[1] Englischer Titel der Dissertation: "Formal Analysis of Variability-Intensive and Context-Sensitive Systems"

[2] Technische Universität Dresden, Institut für Theoretische Informatik, Nöthnitzer Str. 46, 01187 Dresden, philipp.chrszon@tu-dresden.de

entwickelt werden müssen. Ein weiterer Trend ist der steigende Bedarf nach Systemen, die sich an einen veränderten Kontext anpassen können. So kann beispielsweise das Verhalten eines Systems von seiner Position, der Tageszeit, den verfügbaren Netzwerkverbindungen und dem Ladezustand abhängig sein. Der Kontext kann sich aber auch aus der dynamischen Kollaboration mit anderen Systemen und Geräten ergeben, wie etwa in Systems of Systems oder in Smart-Home Umgebungen. Der Variantenreichtum und die Dynamik moderner Systeme stellt eine zusätzliche Herausforderung beim Sicherstellen von deren Korrektheit und Qualität dar. Davon motiviert ist das Ziel der Dissertation die Entwicklung formaler Methoden, welche auf die (quantitative) Analyse variantenreicher und kontextsensitiver Systeme zugeschnitten sind.

In der Softwaretechnik wurde eine Reihe von Ansätzen für den Entwurf sowie die Implementierung von variantenreichen und von kontextsensitiven adaptiven Systemen entwickelt. *Features* sind ein etabliertes Konzept, um die Gemeinsamkeiten und Unterschiede zwischen Varianten einer Systemfamilie, d.h. die Variabilität, zu erfassen. Ein Feature beschreibt dabei eine optionale oder zusätzliche Funktionalität [Za03]. Bei der featureorientierten Entwicklung von Softwareproduktlinien [AK09] steht die systematische Wiederverwendung von Komponenten, genannt Feature-Module, im Vordergrund, um eine hohe Effizienz und Wirtschaftlichkeit beim Erstellen neuer Produkte zu erreichen. Dazu werden, üblicherweise mittels eines automatisierten Generatorwerkzeugs, ausgewählte Feature-Module mit einem Basissystem kombiniert. Die Integration eines Feature-Moduls in das System kann dabei nicht nur neue Funktionalität hinzufügen, sondern auch vorhandene Funktionalität ändern oder überschreiben (Superimposition). Diese Adaptionen sind i.d.R. statisch und sind nach der Auslieferung des Systems fix. Adaptivität zur Laufzeit und kann mittels *dynamischer Features* umgesetzt werden [GH03].

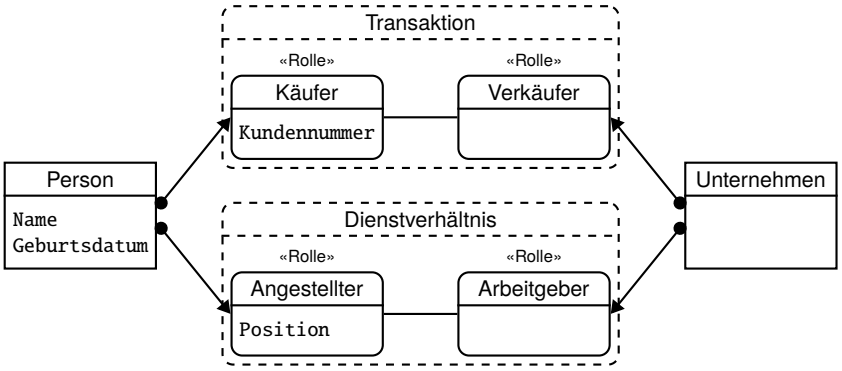


Abb. 1: Beispiel für Rollenmodellierung

Ein weiteres Konzept zur Umsetzung von Laufzeitadaptivität sind *Rollen*. Das Rollenkonzept wurde erstmals von Bachman und Daya [BD77] in der Datenmodellierung eingeführt, um kontextabhängige Informationen und die Evolution von Entitäten abzubilden. Abb. 1 zeigt ein Beispiel. Die Hauptidee besteht darin, die *extrinsischen*, also vom Kontext abhängige, Eigenschaften von den intrinsischen Eigenschaften einer Entität zu trennen und

in Rollen auszulagern. Dies ermöglicht es auf elegante Weise mit transienten Eigenschaften umzugehen. Beispielsweise hat jede Person während der gesamten Lebenszeit einen Namen und ein Geburtsdatum. Die Position der Person innerhalb eines Unternehmens ist hingegen nur während des Arbeitsverhältnisses gültig. Zudem kann eine Nebentätigkeit sehr einfach abgebildet werden, indem eine weitere Rolle hinzugefügt wird. Das Rollenkonzept wurde später in der konzeptionellen Modellierung aufgegriffen, wobei Rollen neben Objekten und Beziehungen als drittes grundlegendes Modellelement eingeführt wurden [Kü14; St00]. Rollen kapseln hier kontextabhängiges Verhalten und kontextabhängige Eigenschaften. Üblicherweise kann das Spielen einer Rolle nicht nur Verhalten hinzufügen, sondern auch das Verhalten des Spielers adaptieren. Zudem können Rollen die Beziehungen zwischen Objekten explizit machen, indem jeder Rolle eine entsprechende Gegenrolle zugeordnet wird, z.B. Käufer und Verkäufer. Ähnlich dazu kann durch die Zuordnung einer Rolle zu einem Kontext (z.B. Transaktion) auch dieser explizit modelliert werden. Das Rollenkonzept geht damit deutlich über das Featurekonzept hinaus. Außerdem erlauben Rollen eine sehr feingranulare Verhaltensadaption auf der Ebene einzelner Komponenten oder Objekte. Sie sind daher besonders für die Modellierung und Implementierung kontextsensitiver und adaptiver Systeme geeignet.

Eine zentrale Herausforderung bei der Entwicklung von featureorientierten und rollenbasierten Systemen sind *Interaktionen*, d.h. emergentes Verhalten welches sich aus der Kombination von zwei oder mehr Features bzw. Rollen ergibt. Während diese Interaktionen oft erwünscht sind, um die Funktionalität eines Systems modular zu implementieren, können auch unerwartete Interaktionen auftreten, die die Korrektheit des Systems oder dessen nicht-funktionale Eigenschaften beeinträchtigen.

Das Ziel der Dissertation [Ch21] ist die Entwicklung formaler Methoden zur Erkennung von Feature- und Rolleninteraktionen. Dazu werden Formalismen zur Modellierung, Modellierungssprachen sowie Analysewerkzeuge entwickelt, die eine explizite Unterstützung von Features und Rollen bieten, um Variabilität und kontextabhängige Adaptionen zu erfassen. Der Fokus liegt hierbei auf Formalismen zur Verhaltensmodellierung, welche eine Analyse mittels Model Checking erlauben. Außerdem werden stochastische Effekte sowie Kostenmaße berücksichtigt, sodass neben Aussagen zur funktionalen Korrektheit des Systems auch Aussagen über dessen quantitative Eigenschaften getroffen werden können, wie z.B. der erwartete Energieverbrauch oder Durchsatz. Die vorgestellten Implementierungen können damit den Entwicklungsprozess komplexer Systeme unterstützen, indem sie unerwünschtes emergentes Verhalten frühzeitig aufzeigen. Die Beiträge der Arbeit sind im einzelnen

1. die Erweiterung des Tools ProFeat [Ch18] zur Modellierung featureorientierter Systeme und die Untersuchung verschiedener Analyseansätze,
2. die Formalisierung von Rollenbasierten Automaten (RBA), um rollenspezifisches Verhalten zu modellieren,
3. der Entwurf und die Implementierung einer rollenbasierten Modellierungssprache,

welche in die Eingabesprache des probabilistischen Model Checkers PRISM [KNP02] übersetzt werden kann,

4. Fallstudien, um die Anwendbarkeit und Skalierbarkeit der entwickelten Werkzeuge zu demonstrieren, und
5. ein Ansatz für die Koordination von Rollen basierend auf der exogenen Koordinationssprache Reo [Ar04].

Die Beiträge zur Modellierung und Analyse von rollenbasierten Systemen werden in den folgenden Abschnitten näher erläutert.

2 Formale Modellierung von kontextabhängigem Verhalten

Im Rahmen der Arbeit wurde ein Modellierungsformalismus für kontextsensitive Systeme entwickelt, der die Beschreibung von kontextspezifischen Adaptionen mittels Rollen erlaubt. Die im folgenden verwendeten Begriffe sind an das *Compartment Role Object Model* (CROM) [Kü15], einem von UML-Klassendiagrammen inspirierten Metamodell, welches zahlreiche existierende Ansätze zur Rollenmodellierung vereint, angelehnt. Wir unterscheiden verschiedene Arten von Komponenten. Die Instanzen von *natürlichen Typen* stellen die Spieler von Rollen dar (z.B. Person, Unternehmen), *Rollen* (z.B. Student, Server) und die Kontexte, in denen Rollen gespielt werden und welche *Compartments* genannt werden (z.B. Universität, Transaktion). Im Gegensatz zum CROM und verwandten Metamodellen beschäftigt sich die Arbeit nicht vordergründig mit der Struktur von Systemen, sondern mit deren operationellem Verhalten.

Rollenbasierte Automaten (RBA) [Ch20] erlauben eine einheitliche Repräsentation des Verhaltens von Rollen, deren Spielern, sowie auch Compartments. Aus diesen grundlegenden Komponenten lassen sich durch Komposition komplexe Systemmodelle zusammensetzen. Der RBA-Formalismus umfasst Kompositionsoperatoren für die Parallelkomposition und für die Bindung von Rollen an Komponenten. Die Bindung *ermöglicht* einer Komponente das *aktive* Spielen einer Rolle [MKK12], d.h. die Ausübung des rollenspezifischen Verhaltens. Ein RBA, welcher einer Rolle repräsentiert, kann dabei an einen beliebigen anderen RBA gebunden werden, einschließlich RBA, die eine andere Rolle oder sogar ein Compartment repräsentieren. Die Bindung von Rollen an Compartments, die wiederum selbst aus rollenbehafteten Komponenten bestehen, ermöglicht eine hierarchische Modellierung von Systemen. Der Parallelkompositionsoperator formalisiert die Interaktion von RBA als Synchronisation über gemeinsame Aktionen. Mithilfe der genannten Operatoren kann durch schrittweise Komposition der RBA für die Systemkomponenten ein einzelner RBA erzeugt werden, der das gesamte Systemverhalten beschreibt. Dieser RBA umfasst zunächst einmal alle möglichen Kombinationen für das Spielen von Rollen. Jedoch sind üblicherweise nicht alle die Möglichkeiten erlaubt oder erwünscht. Eine zusätzliche Komponente, der Rollenspielkoordinator, beschreibt die Regeln für das Annehmen und Ablegen von Rollen.

Mithilfe des Koordinators können somit die Adaptionen des Systems auf sich ändernden Kontext gesteuert werden.

Die Struktur von RBA ähnelt der von Markov-Entscheidungsprozessen (MEP) und kombiniert nicht-deterministische und probabilistische Zustandsübergänge. RBA sind charakterisiert durch *Rollenannotationen* an den Zustandsübergängen. Formal ist die Menge der Rollenannotationen über einer Menge von Rollen R definiert als $\mathbb{A}(R) = \{r, \overline{r}, +r : r \in R\}$. Intuitiv steht eine Annotation r für ein aktives Spielen der Rolle r und $\overline{r}$ drückt aus, dass r explizit nicht gespielt wird. Die dritte Variante, $+r$, erlaubt das Ersetzen einer bereits vorhandenen Transition des Spielers durch eine Transition der Rolle bei der Bindung. Eine Rolle kann somit das bestehende Verhalten ihres Spielers überschreiben und damit adaptieren. Rollenannotationen sind integraler Bestandteil der Transitionsrelation von RBA, wie in folgender Definition gezeigt.

Definition 1. Ein RBA ist ein Tupel $\mathcal{A} = (S, Act, R, \longrightarrow, S^{init})$, wobei S eine endliche Menge von Zuständen, Act eine Menge von Aktionen, $R = \langle B, U \rangle$ ein Rollen-Interface, $\longrightarrow \subseteq S \times Act \times \wp(\mathbb{A}(R)) \times Distr(S)$ die Transitionsrelation und $S^{init} \subseteq S$ die Menge der Initialzustände ist. Hierbei bezeichnet $Distr(S)$ die Menge der stochastischen Verteilungen über S.

Die Annotationen erfüllen zwei Funktionen. Erstens erlauben sie bei der Analyse Aussagen über das aktive Spielen von Rollen und dessen zeitliche Abfolge zu treffen. Zweitens können bei der Komposition eines RBA mit einem Koordinator Transitionen mit bestimmten Annotationen erzwungen oder unterdrückt werden. Das Ergebnis einer solchen Komposition ist ein Standard-MEP, bei dem das Rollenspiel in die Aktionen der Zustandsübergänge kodiert ist. Die MEP-Semantik von RBA erlaubt die Anwendung von existierenden Analysealgorithmen und -werkzeugen, insbesondere probabilistische Model Checker wie z.B. PRISM.

3 Implementierung einer rollenbasierten Modellierungssprache

Um komplexe Modelle von kontextsensitiven adaptiven Systemen kompakt beschreiben zu können, wurde im Rahmen der Dissertation eine Modellierungssprache (RML) mit RBA-Semantik entwickelt, die eine explizite Unterstützung des Rollenkonzepts bietet. Ein Modell beschreibt sowohl die Struktur des Systems als auch das Verhalten der einzelnen Systemkomponenten. Diese Aufteilung ist inspiriert von Modellierungssprachen für featureorientierte Systeme [Ch18; Cl12], die ein Featuremodell mit Verhaltensdefinitionen für die Feature-Module kombinieren. Die Strukturbeschreibung in RML folgt existierenden Sprachen für die konzeptionelle Modellierung [Kü15]. Da die Sprache primär für die Analyse mit probabilistischem Model Checking konzipiert ist, werden die RBA, die das Verhalten der Komponenten definieren, mittels einer dort üblichen Guarded Command Language beschrieben. Konkret erweitert RML die Eingabesprache von PRISM um rollenspezifische Sprachelemente sowie um die Unterstützung für Metaprogrammierung. Ein RML-Modell

kann ein einzelnes System, aber auch eine Familie von Systemen beschreiben, die sich aus verschiedenen Kombinationen der Systemkomponenten ergibt. Dies ist vor allem für die Analyse von Vorteil, da somit das Finden von problematischen Kombinationen vereinfacht wird, die unerwünschtes emergentes Verhalten aufweisen.

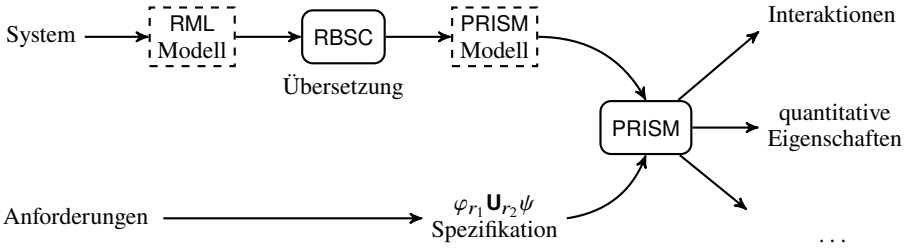


Abb. 2: Schema für die Systemanalyse

Basierend auf der MEP-Semantik von RBA kann ein RML-Modell wie in Abb. 2 gezeigt in ein PRISM-Modell übersetzt werden, sodass die Modellprüfung und quantitative Analyse von PRISM durchgeführt wird. Der Vorteil dieses Ansatzes liegt darin, dass die von PRISM unterstützten vielfältigen Analysemöglichkeiten auch für RML-Modelle anwendbar sind. Bei der Übersetzung bleiben die Rollenannotationen des Modells erhalten. Daher enthalten die von PRISM generierten Gegenbeispiele im Falle einer verletzten Spezifikation nicht nur eine Abfolge von Aktionen und Zuständen, sondern zusätzlich noch die Sequenz der gespielten Rollen. Dies vereinfacht die Identifikation der beteiligten Komponenten und deren (unerwünschte) Interaktionen erheblich.

4 Evaluation

Dieser Abschnitt gibt einen kurzen Überblick über die Evaluation der entwickelten Modellierungs- und Analyseansätze. Auf ein illustratives Beispiel zum Erkennen von unerwünschten Interaktionen zwischen kooperierenden Systemen folgen Ausführungen zur Anwendbarkeit sowie Skalierbarkeit der entwickelten Analysemethoden und -werkzeuge.

4.1 Erkennen von Interaktionen

Wir betrachten das Beispiel einer selbst-adaptiven Produktionszelle, welche aus mehreren Robotern besteht [GOR06]. Die Aufgabe der Zelle ist die schrittweise Verarbeitung von Werkstücken, wobei jeder Roboter mit einem entsprechenden Werkzeug für die Ausführung eines Schrittes ausgestattet ist, also eine bestimmte Rolle im Verarbeitungsprozess übernimmt. So könnte z.B. der erste ein Loch in das Werkstück bohren, der zweite eine Schraube einfügen und der dritte die Schraube festdrehen. Im Falle des Defekts eines Werkzeuges konfiguriert sich die Zelle automatisch neu, indem bestimmte Roboter ihr Werkzeug und

damit ihre Rolle wechseln. Der Wechsel des Werkzeuges nimmt eine gewisse Zeit in Anspruch, daher ist es sinnvoll, dass für jeden Verarbeitungsschritt mindestens ein Roboter zuständig ist. Im Folgenden gehen wir davon aus, dass einer der Roboter einen doppelt so hohen Durchsatz wie die anderen Roboter hat. Um diese zusätzliche Kapazität sinnvoll zu nutzen, könnten zwei Zellen wie in Abb. 3 gezeigt so kombiniert werden, dass sie den schnelleren Roboter (5) gemeinsam nutzen.

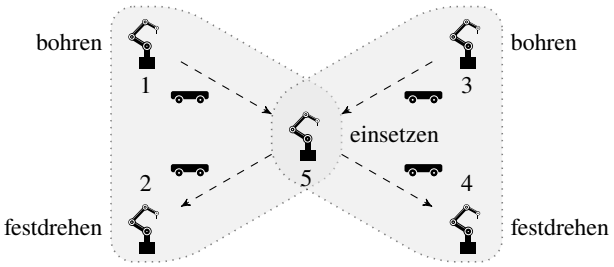


Abb. 3: Zwei automatisierte Produktionszellen, welche sich einen Roboter teilen

Der geteilte Roboter muss zwangsläufig in beiden Zellen dieselbe Aufgabe übernehmen, woraus sich folgende Interaktion ergibt. Angenommen der Bohrer von Roboter 1 bricht ab, woraufhin dieser seinen zugewiesenen Arbeitsschritt nicht mehr ausführen kann. Bei der folgenden Rekonfiguration der linken Zelle werden dann die Rollen der Roboter 1 und 5 getauscht und die Verarbeitung kann fortgesetzt werden. Jedoch übernimmt nun in der rechten Zelle kein Roboter mehr das Einsetzen von Schrauben, was eine Rekonfiguration dieser Zelle auslöst. Hierbei wird die Einsetzen-Rolle wieder Roboter 5 zugewiesenen. Doch dann ist die linke Zelle wiederum nicht mehr für eine durchgängige Verarbeitung konfiguriert, was eine erneute Rekonfiguration zur Folge hat, und immer so weiter. Das beschriebene Szenario lässt sich leicht aus der von PRISM ausgegebenen Zustandsabfolge ablesen. Durch die explizite Annotation des Rollenspiels sind Interaktionen zwischen Komponenten bzw. Rollen oftmals auch ohne Betrachtung der einzelnen Systemzustände erkennbar. Hervorzuheben ist, dass diese Interaktion, welche sich in wechselseitigen Rekonfigurationen äußert, nicht die funktionale Korrektheit des Gesamtsystems beeinflusst. Beide Zellen können bis zur Rekonfiguration der jeweils anderen Zelle ihre Verarbeitung ausführen, jedoch vermindert sich insgesamt der Durchsatz des Systems. Es handelt sich hierbei also um eine *quantitative* Interaktion, welche aufgrund der gewählten Analysemethode dennoch erkennbar ist.

4.2 Anwendbarkeit und Skalierbarkeit

Um die Skalierbarkeit und Anwendbarkeit des rollenbasierten Modellierungs- und Analyseansatzes zu evaluieren, wurde ein Peer-to-Peer Dateitransferprotokoll untersucht. Das Modell ist beliebig skalierbar, wobei die Anzahl der Netzwerkknoten, die Netzwerktopologie sowie die Anzahl der verschiedenen Dateien im Netzwerk konfigurierbar ist. Für die Quantifizierung des Overheads, der durch den rollenbasierten Ansatz verursacht wird, wurde

ein funktional identisches Modell ohne Nutzung rollenspezifischer Sprachkonstrukte in der herkömmlichen PRISM Eingabesprache modelliert und mit PRISM analysiert. Hierbei ist anzumerken, dass das Standardmodell speziell auf das Benchmarkszenario zugeschnitten ist und somit eine weit weniger flexible Komposition der Systemkomponenten erlaubt. Das Ergebnis der Vergleichs der beiden Analysemethoden ist in Abb. 4 dargestellt, wobei die Analysedauer für den rollenbasierten Ansatz relativ zur Analysedauer des Standardmodells eingetragen ist. Mit steigender Größe der Modelle wird der Overhead des rollenbasierten Ansatz vernachlässigbar. Für die größten Modellinstanzen hat sich die Struktur der rollenbasierten Modelle als vorteilhaft erwiesen, sodass hier die Analyse sogar schneller als die Analyse mit Standardtools durchgeführt wurde.

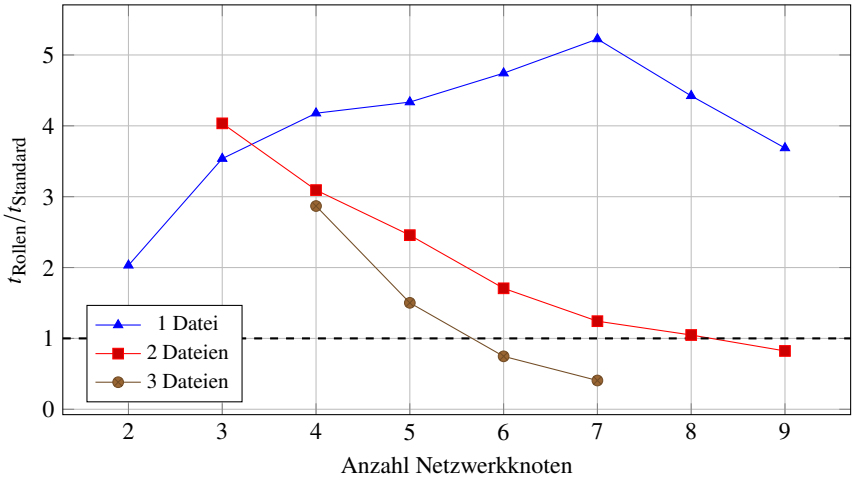


Abb. 4: Analysedauer von Rollen-basierten Modellen relativ zu Standardmodellen

Die Anwendbarkeit von ProFeat für die Analyse stochastischer featureorientierter Systeme wurde anhand von mehreren Fallstudien demonstriert. Die erste Reihe von Fallstudien fokussierte auf den Vergleich von familienbasierter Analyse, bei der die gesamte Systemfamilie in einem Durchlauf analysiert wird, und produktbasierter Analyse, bei der jede Variante des Systems einzeln analysiert wird. Hierbei zeigte sich, dass eine familienbasierte Analyse stochastischer Systeme oftmals nur bei Familien mit sehr vielen Varianten sowie signifikanten strukturellen Gemeinsamkeiten zwischen den Varianten vorteilhaft ist. Die weiteren Fallstudien haben die Anwendbarkeit für “klassische” Softwareproduktlinien und darüber hinaus auch für generelle Systemfamilien gezeigt. Insbesondere erlaubt die Unterstützung von dynamischen Features in ProFeat auch die Modellierung und Analyse von adaptiven Systemen.

5 Zusammenfassung

In der Dissertation wurden die Konzepte von Features und Rollen, welche für die Modellierung und Implementierung von variantenreichen bzw. kontextsensitiven adaptiven Systemen eingeführt wurden, auf formale Methoden mit dem Ziel einer quantitativen Analyse übertragen. Für die Modellierung und Analyse featureorientierter Systeme wurde das Tool ProFeat weiterentwickelt und dessen Anwendbarkeit anhand von mehreren Fallstudien demonstriert. Um kontextabhängiges Verhalten innerhalb von formalen Modellen zu erfassen, wurden rollenbasierte Automaten eingeführt. Basierend darauf wurde eine rollenbasierte Modellierungssprache mit RBA-Semantik entwickelt, welche sich für die Analyse in die Eingabesprache des Model Checkers PRISM übersetzen lässt. Die Eignung des Ansatzes für das Erkennen von (quantitativen) Interaktionen sowie dessen Skalierbarkeit wurde mittels illustrativer Fallstudien gezeigt.

Literatur

[AK09] Apel, S.; Kästner, C.: An Overview of Feature-Oriented Software Development. Journal of Object Technology 8/5, S. 49–84, 2009.

[Ar04] Arbab, F.: Reo: a channel-based coordination model for component composition. Mathematical Structures in Computer Science 14/, S. 329–366, Juni 2004, ISSN: 1469-8072.

[BD77] Bachman, C. W.; Daya, M.: The Role Concept in Data Models. In: Proceedings of the Third International Conference on Very Large Data Bases - Volume 3. VLDB '77, VLDB Endowment, Tokyo, Japan, S. 464–476, 1977, URL: http://dl.acm.org/citation.cfm?id=1286580.1286629.

[Ch18] Chrszon, P.; Dubslaff, C.; Klüppelholz, S.; Baier, C.: ProFeat: feature-oriented engineering for family-based probabilistic model checking. Formal Aspects of Computing 30/1, S. 45–75, 2018.

[Ch20] Chrszon, P.; Baier, C.; Dubslaff, C.; Klüppelholz, S.: From features to roles. In: SPLC '20: 24th ACM International Systems and Software Product Line Conference, Montreal, Quebec, Canada, October 19-23, 2020. ACM, 19:1–19:11, 2020.

[Ch21] Chrszon, P.: Formal Analysis of Variability-Intensive and Context-Sensitive Systems, Diss., TU Dresden, 2021, URL: https://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa2-736915.

[Cl12] Classen, A.; Cordy, M.; Heymans, P.; Legay, A.; Schobbens, P.-Y.: Model checking software product lines with SNIP. International Journal on Software Tools for Technology Transfer 14/5, S. 589–612, 2012, ISSN: 1433-2787.

[GH03] Gomaa, H.; Hussein, M.: Dynamic Software Reconfiguration in Software Product Families. In (van der Linden, F., Hrsg.): Software Product-Family Engineering, 5th International Workshop, PFE 2003, Siena, Italy, November 4-6, 2003, Revised Papers. Bd. 3014. Lecture Notes in Computer Science, Springer, S. 435–444, 2003.

[GOR06] Güdemann, M.; Ortmeier, F.; Reif, W.: Formal Modeling and Verification of Systems with Self-x Properties. In: Autonomic and Trusted Computing, Third International Conference, ATC 2006, Wuhan, China, September 3-6, 2006, Proceedings. S. 38–47, 2006.

[KNP02] Kwiatkowska, M.; Norman, G.; Parker, D.: PRISM: Probabilistic symbolic model checker. In: Computer Performance Evaluation: Modelling Techniques and Tools. Springer, S. 200–204, 2002.

[Kü14] Kühn, T.; Leuthäuser, M.; Götz, S.; Seidl, C.; Aßmann, U.: A Metamodel Family for Role-Based Modeling and Programming Languages. In: Software Language Engineering. Springer, S. 141–160, 2014.

[Kü15] Kühn, T.; Böhme, S.; Götz, S.; Aßmann, U.: A combined formal model for relational context-dependent roles. In: Proceedings of the 2015 ACM SIGPLAN International Conference on Software Language Engineering, SLE 2015, Pittsburgh, PA, USA, October 25-27, 2015. S. 113–124, 2015.

[MKK12] Mizoguchi, R.; Kozaki, K.; Kitamura, Y.: Ontological analyses of roles. In: Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on. IEEE, S. 489–496, 2012.

[St00] Steimann, F.: On the representation of roles in object-oriented and conceptual modelling. Data & Knowledge Engineering 35/1, S. 83–106, 2000.

[Za03] Zave, P.: An experiment in feature engineering. In: Programming methodology. Springer, S. 353–377, 2003.

Philipp Chrszon wurde am 17.07.1989 in Zwickau geboren. Er studierte Informatik an der Technischen Universität Dresden und erlangte 2014 einen Masterabschluss. Seit Beginn seiner Promotionszeit 2014 arbeitet er als wissenschaftlicher Mitarbeiter an der Professur für Algebraische und logische Grundlagen der Informatik an der TU Dresden. Von Oktober 2014 bis Oktober 2017 war er Doktorand im Graduiertenkolleg “Rollenbasierte Software-Infrastrukturen für durchgängig-kontextsensitive Systeme (RoSI)” und anschließend assoziiertes Mitglied. Von Oktober 2017 bis August 2019 war er in den Sonderforschungsbereich 912 HAEC eingebunden. Im September 2020 verteidigte er erfolgreich seine Dissertation.

Vorhersagebasierte Suche für autonomes Spielen[1]

Alexander Dockhorn[2]

Abstract: Lernverfahren befähigen einen Agenten autonom eine für ihn unbekannte Umgebung zu explorieren und ihm gestellte Aufgaben zu erfüllen. Hierbei erlauben Modell-bildende Verfahren dem Agenten, ein geistiges Abbild seiner Umgebung zu konstruieren und in diesem das Resultat seiner Handlungen vorherzusehen. Die Konstruktion und Verwendung eines solchen Modells stellt die Schwerpunkte meiner Dissertation dar. Hierfür wird zunächst eine theoretische Grundlage für die Dekomposition von Forward Modellen geschaffen. Darauf basierend werden das Local Forward Model sowie das Object-based Forward Model als effiziente Modellheuristiken abgeleitet. Besonderes Augenmerk wird zudem auf die Modellierung unsicherer Informationen gelegt. Abschließend werden Anwendungen im Bereich der künstlichen Intelligenz in Spielen und der Robotik für die entwickelten Verfahren demonstriert.

1 Einführung

Die Entwicklung von künstlicher allgemeiner Intelligenz ist eines der wichtigsten langfristigen Ziele in der Forschung zur künstlichen Intelligenz (KI). Dazu gehört die Entwicklung von intelligenten Algorithmen, die eine Vielzahl von Problemen bewältigen können. Während KI-Methoden in den letzten Jahren viele erfolgreiche Anwendungen hervorgebracht haben, sind die Fähigkeiten dieser Agenten oft auf sehr spezifische Problemszenarien beschränkt. In Anbetracht der Tatsache, dass das menschliche Gehirn in der Lage ist, komplex zu denken und sich an neue Kontexte, Aufgaben und Umgebungen anzupassen, hinkt die KI-Forschung bei der Entwicklung von vergleichbar flexibel anwendbaren Computeragenten dem natürlichen Vorbild hinterher.

Die Forschung in der künstlichen allgemeinen Intelligenz kann in die modellorientierte und die anwendungsorientierte Sichtweise unterteilt werden. Bei der modellorientierten Sichtweise werden die Fähigkeiten des menschlichen Gehirns analysiert und teilweise kopiert. Im Gegensatz dazu werden bei der anwendungsorientierten Sichtweise Lösungen durch die Analyse von immer komplexeren Problemen erstellt. Während jede dieser Lösungen oft auf eine bestimmte Anwendung beschränkt ist, sind die Anwendungen im Laufe der Jahre immer fortschrittlicher geworden. Der Vorteil dieses Ansatzes liegt in der Vergleichbarkeit der entwickelten Lösungen auf Basis der jeweiligen Aufgabenstellung.

Spiele, insbesondere digitale Spiele, stellen nützliche Werkzeuge bei der Entwicklung und Bewertung von KI-Agenten dar. Nicht nur verfügen sie über klare und quantifizierbare Ziele, sondern erfordern auch komplexe Entscheidungsprozesse. Digitale Spiele haben zudem den

[1] Englischer Titel der Dissertation: "Prediction-based Search for Autonomous Game-Playing"

[2] Lehrstuhl für Computational Intelligence, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg, E-Mail: alexander.dockhorn@ovgu.de; Website: https://adockhorn.github.io/

Vorteil, dass sie für Computer vollständig zugänglich sind und wir ihre Ein- und Ausgabe ohne großen technischen Aufwand konfigurieren können. Da Spiele auch von menschlichen Spielern gespielt werden, ermöglichen sie interessante Mensch-Computer-Interaktionen und liefern uns große Datensätze des menschlichen Spielverhaltens.

Nachdem viele Jahre die Entwicklung eines KI-Agenten für ein einzelnes Spiel im Fokus der Forschung stand (vgl. Entwicklungen von Schach und Go Agenten), gewinnt die Entwicklung eines einzelnen Agenten, der in der Lage ist, mehrere Spiele zu spielen, zunehmend an Popularität und wird als vielversprechender nächster Schritt zur Verbesserung der Fähigkeiten von KI-Agenten gesehen [Le13]. Im Rahmen der Evaluierung der allgemeinen Spielfähigkeiten eines Agenten erhält er jeweils eine Beschreibung des aktuellen Spiels in Form seines Zustands- und Aktionsraums sowie des Regelwerks. Während der Entwicklungsphase des Agenten sind diese Informationen unbekannt, sodass der Agent selbst lernen muss, wie er diese nutzen kann, um die Spiele effektiv zu spielen. Diese Aufgabe erfordert, dass der Agent Algorithmen zur Wissensrepräsentation, des Lernens, des Schlussfolgerns und der Entscheidungsfindung kombiniert [GLP05], wodurch die notwendige Expertise für das Spielen des Spiels vom Entwickler des Agenten auf den Agenten selbst verlagert wird.

1.1 Übersicht zu Existierenden Lernverfahren

Zunächst wird ein kurzer Überblick über bereits existierende Lösungen der künstlichen Intelligenz in Spielen gegeben, um die im Rahmen meiner Dissertation entwickelten Verfahren in den Forschungskontext einzubetten.

Grundlage der Betrachtungen ist das Agent-Environment Interface [SB18], welches die fünf Komponenten *Agent*, *Umgebung*, *Zustand der Umgebung*, *Aktionen des Agenten* und dessen *Belohnung* umfasst. Der Agent interagiert durch die Ausführung von Aktionen $a \in \mathscr{A}$ mit seiner Umgebung. In Folge dessen verändert sich ihr Zustand $s \in \mathscr{S}$ und der Agent erhält eine numerische Belohnung r, welche über den Erfolg bzw. Misserfolg im Erfüllen der ihm gestellten Aufgabe informiert. Ein solcher Prozess wird als Markov-Entscheidungsproblem modelliert, welches den Zustandsübergang und die Belohnung als Wahrscheinlichkeitsverteilung über die Sequenz der vorherigen Zustände und Aktionen definiert:

$$P(S_{t+1}, R_{t+1} \mid S_0, A_0, \ldots, S_t, A_t)$$

Das Umgebungsmodell kann in ein Zustandsübergangs- und ein Belohnungsmodell aufgeteilt werden:

$$\begin{aligned} \text{Zustandsübergangsmodell:} \quad & p(s_t \mid s_{t-1}, a_{t-1}, \ldots, s_{t-m}, a_{t-m}) \\ \text{Belohnungsmodell:} \quad & q(r_t \mid s_{t-1}, a_{t-1}, \ldots, s_{t-m}, a_{t-m}) \end{aligned}$$

Aus dieser Aufteilung ergeben sich zwei grundlegende Betrachtungsweisen auf das Lernproblem. Während Reinforcement Learning Algorithmen sich auf die Belohnungsverteilung fokussieren, nutzen Model Learning Verfahren die Informationen vorangegangener Interaktionen um das Zustandsübergangsmodell zu approximieren. Letzteres kann dann in

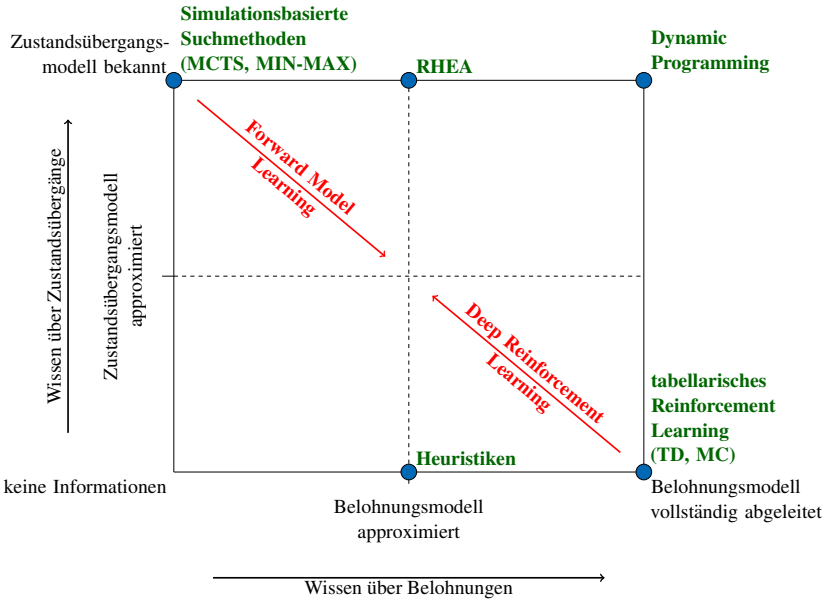


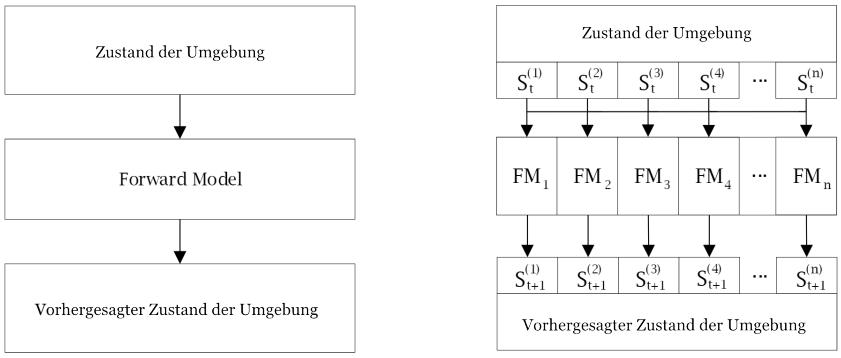
Abb. 1: Vergleich von allgemeinen Lernverfahren basierend auf dem Wissen des Agenten über das Belohnungsmodell und das Zustandsübergangsmodell [Do20].

Verbindung mit Such- und Planungsalgorithmen verwendet werden, um den Wert einer Aktion anhand der zu erwartenden Ergebnisse zu bestimmen.

Der Einsatz von Planungs- bzw. Suchalgorithmen setzt die Kenntnis des Zustandsübergangsmodells voraus. Ist ein solches bekannt, haben Verfahren wie z.B. Monte Carlo Tree Search (MCTS) [Br12] und der Rolling Horizon Evoluationary Algorithm (RHEA) [GLP17] bereits gute Leistungen in einer Vielzahl von Spielen erbringen können [Pe19]. Hingegen ist der Einsatz von tabellarischen Reinforcement Learning Algorithmen [SB18] auf einfache Zustands- und Aktionsräume begrenzt. Unter Verwendung von tiefen neuronalen Netzen konnte durch die Approximation des Belohnungsmodells eine breitere Anwendbarkeit erreicht werden [Mn15]. Die in dieser Arbeit vorgestellten Verfahren des Forward Model Learnings erweitern die Anwendbarkeit von Planungs- und Suchalgorithmen auf Szenarien in denen ein Zustandsübergangsmodell nicht bekannt ist. Abbildung 1 gibt einen Überblick über die in diesem Abschnitt präsentierten Methoden und stellt die strategische Position der Arbeit im Kontext anderer wissenschaftlicher Arbeiten dar.

2 Dekomposition von Forward Model Learning

Das Ziel von Forward Model Learning ist das Erlernen von Zustandsübergangsmodellen aufgrund bisheriger Interaktionen des Agenten mit seiner Umgebung. Forward Models treffen anhand des aktuellen Zustands des Systems und der Aktion des Agentens eine Vorhersage über den nächsten zu beobachtenden Zustand.



(a) End-to-End Forward Model (b) Decomposed Forward Model

Abb. 2: Visueller Vergleich von Forward Model Architekturen [Do20]

In meiner Dissertation untersuche ich Algorithmen zum Lernen von Forward Modellen, welche mithilfe eines Suchalgorithmus verwendet werden können. Im Folgenden wird eine Unabhängigkeit der Vorhersage von vorherigen Zeitschritten angenommen. Hierdurch lässt sich das Zustandsübergangsmodell wie folgt reduzieren:

$$p(s_t \mid s_{t-1}, a_{t-1}, \ldots, s_{t-m}, a_{t-m}) = p(s_t \mid s_{t-1}, a_{t-1})$$

Sind Beobachtungen von Ein- und Ausgaben eines uns unbekannten Zustandsübergangsmodells vorhanden, so können wir anhand eines überwachten Lernverfahrens ein approximatives Modell generieren. Bildet dieses den aktuellen Zustand und die Aktion des Agenten direkt auf den Folgezustand ab, sprechen wir von einem End-to-End Forward Model. Dieses Verfahren wird hauptsächlich in Kombination mit Deep Neural Networks verwendet, welche im Falle hinreichender Trainingsdaten oft in der Lage sind, sehr komplexe Zustandsübergangsmodelle zu erlernen [HS18]. Die Generierung einer ausreichenden Menge an Trainingsdaten ist jedoch zeit- und kostenaufwändig.

Aus diesem Grund werden Dekompositionstechniken genutzt um das Zustandsübergangsmodell in mehrere Komponentenmodelle zu zerlegen. Hierfür werden zunächst voneinander unabhängige Mengen von Variablen anhand einer Abhängigkeitsanalyse bestimmt [DTK18]. Bei einer vollständigen Unabhängigkeit beteiligter Zustandsvariablen kann ein Modell pro Variable erlernt werden, bzw. ein Modell für die Änderung ihres Werts bis zum nächsten Zeitschritt (differentiell):

$$\text{Komponentenmodell:} \quad p(s_t^i \mid s_{t-1}, a_{t-1}, \ldots, s_1, a_1)$$
$$\text{Differentielles Komponentenmodell:} \quad p(s_{t-1}^i - s_t^i \mid s_{t-1}, a_{t-1}, \ldots, s_1, a_1)$$

Die Prädiktion des Folgezustands ergibt sich aus der Aufteilung des Zustands in voneinander unabhängige Komponenten, der Prädiktion dieser Komponenten und der anschließenden Aggregierung der Resultate. Abbildung 2 zeigt einen Vergleich des Prädiktionsablaufs zwischen End-to-End und Decomposed Forward Modellen.

3 Modellbildungsheuristiken

Um die Modellbildung zu beschleunigen kann Vorwissen über das Zustandsübergangsmodell in Form von Modellbildungsheuristiken in den Prozess einfließen. Im Folgenden werden das Local Forward Model und das Object-based Forward Model beschrieben, welche beispielhaft für die Abstraktion eines Zustandsübergangsmodells durch die Annahme von Unabhängigkeiten stehen.

3.1 Local Forward Model

Local Forward Models ähneln in ihrer Funktionsweise Convolutional Neural Networks [GBC16]. Hierbei wird der Folgezustand anhand unabhängiger Beobachtungen einzelner Informationsquellen vorhergesagt, wobei der Folgezustand lediglich vom eigenen Zustand und dem Zustand „benachbarter“ Quellen abhängt. Nachfolgend gehen wir davon aus, dass weit entfernte Informationsquellen voneinander unabhängig sind, und demnach der Folgezustand einer der Quellen jeweils nur von benachbarten Quellen abhängt. Ein solches System ist beispielsweise bei der Betrachtung von räumlichen Interaktionen aus der Vogelperspektive gegeben. Zwei Objekte oder Personen, die weit voneinander entfernt sind, können sich hierbei nicht direkt beeinflussen, wohingegen nebeneinanderliegende Objekte sich gegenseitig durch Interaktionen beeinflussen können.

Die lokale Transitionsfunktion beschreibt die Veränderung einer Zustandsvariable abhängig von dem Wert der Zustandsvariablen in ihrer Nachbarschaft:

$$fm_i : \left(N(S_t^{(i)}), A_t\right) \longmapsto S_{t+1}^{(i)}$$

wobei $N(S_t^{(i)}) \subseteq \mathscr{S}$ die Menge benachbarter Zustandsvariablen der Zustandsvariable i zum Zeitpunkt t beschreibt. Gegeben einer Distanzfunktion d und einem Schwellwert ε kann die lokale Nachbarschaft einer Zustandsvariable durch die folgende Funktion bestimmt werden:

$$N(S_t^{(i)}) = \left\{ S_t^{(j)} \mid d(S_t^{(i)}, S_t^{(j)}) \leq \varepsilon, \quad j \in 1, \ldots, n \right\}$$

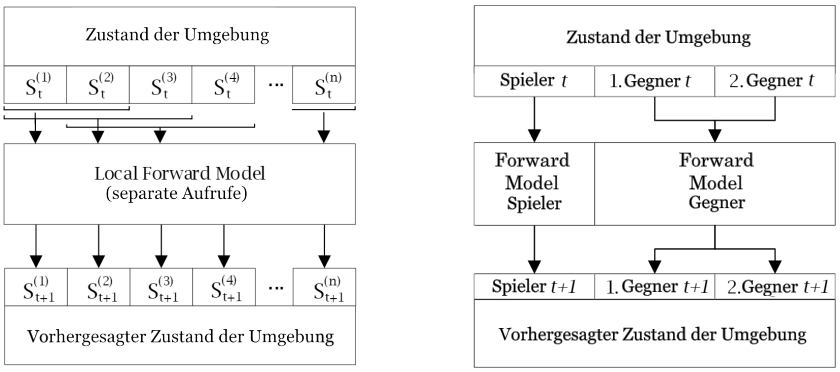
Im Folgenden wird eine Zustandsbeschreibung auf Basis einer Matrix T verwendet, deren Zellen $T(i,j)$ den nominalen Zustand einer Zustandsvariable an der Position (i,j) angibt:

$$T = \begin{bmatrix} T(1,1) & \ldots & T(1,m) \\ \vdots & \ddots & \vdots \\ T(n,1) & \ldots & T(n,m) \end{bmatrix}$$

Die lokale Transitionsfunktion ergibt sich aus:

$$fm_{x,y} : \left(N(x,y)_t, A_t\right) \longmapsto T(x,y)_{t+1}$$

In einer Evaluierung von Lucas et al. [Lu19] zeigen sich Local Forward Models durch wenige Beispiele in der Lage, das Regelwerk von Conway's Game of Life präzise abzubilden. Durch die häufig in Spielen verwendete Zustandsbeschreibung als Matrix bietet sich



(a) Local Forward Model

(b) Object-based Forward Model

Abb. 3: Visueller Vergleich von Modellbildungsheuristiken [Do20]

deren Modellierung durch Local Forward Modelle an. Tests mit dem Spiel Sokoban [Do19] zeigten, dass präzise Modelle anhand weniger Wiederholungen gelernt und auf unbekannte Zustände übertragen werden können.

3.2 Object-based Forward Model

Die Grundlage des Object-based Forward Models bilden nicht räumliche, sondern semantische Abhängigkeiten von Zustandsvariablen. Setzt sich der Zustand aus voneinander abgrenzbaren Entitäten ab, welchen jeweils eine Menge an Zustandsvariablen zugeordnet werden, so können diese separat von anderen Entitäten modelliert werden. Hierbei nutzt das Object-based Forward Model Vorwissen über solch separierbare Zustandsvariablengruppen, um eine entsprechende Aufteilung in voneinander unabhängige Modellkomponenten zu realisieren.

Im Folgenden gehen wir davon aus, dass der Agent über eine objekt-basierte Partitionierung der beobachtbaren Zustandsvariablen verfügt. Eine Partitionierung kann die Mengen von Sensorwerten $1,\ldots,m$ beschreiben, die zu einem gemeinsamen Objekt in der Umgebung gehören, wie z. B. dem Avatar des Agenten oder eines Nicht-Spieler-Charakters.

$$\begin{aligned}\mathscr{S} &= (S^{(1)}, S^{(2)}, \ldots, S^{(n)}) \\ &= (\underbrace{S^{(1,1)}, \ldots, S^{(1,i)}}_{\text{Objekt 1}}, \ldots, \underbrace{S^{(m,1)}, \ldots, S^{(m,k)}}_{\text{Objekt m}})\end{aligned}$$

Für jedes Objekt approximieren wir ein Objektmodell fm_i, welches dessen aktuellen Zustand und die Aktion des Agenten auf den nächsten Zustand abbildet:

$$fm_i : \left((S_t^{(i,1)}, \ldots, S_t^{(i,k)}),\, A_t\right) \longmapsto (S_{t+1}^{(i,1)},\, \ldots,\, S_{t+1}^{(i,k)})$$

Sind die Attribute eines Objekts zudem voneinander unabhängig, kann ein Decomposed Forward Model verwendet werden, um das Objektmodell des Objekts zu modellieren.

$$fm_{i,j} : \left((S_t^{(i,1)}, \ldots, S_t^{(i,k)}),\ A_t \right) \longmapsto S_{t+1}^{(i,j)}$$

Dies führt zu einer mehrstufigen Abstraktion des Systems. Um den nächsten Zustand vorherzusagen, wird der aktuelle Zustand zunächst partitioniert und in Folge dessen jedes Objektmodell aufgerufen. Die Ausgaben dieser werden nachfolgend aggregiert, um den vorhergesagten Zustand zu erhalten. Sollten unterschiedliche Objekte gleiche Verhaltensweisen aufzeigen, so können deren gelernten Modelle zusammengeführt werden (vgl. Abbildung 3).

$$\begin{aligned} fm(S_t, A_t) &= (fm_1(S_t, A_t),\ \ldots,\ fm_n(S_t, A_t)) \\ &= ((fm_{1,1}((S_t^{(1,1)}, \ldots, S_t^{(1,k)}), A_t),\ \ldots,\ fm_{m,k}((S_t^{(m,1)}, \ldots, S_t^{(m,k')}), A_t))) \\ &= (S_{t+1}^{(1)},\ S_{t+1}^{(2)},\ \ldots,\ S_{t+1}^{(n)}) = S_{t+1} \end{aligned}$$

4 Predictive State Determinization

In der bisherigen Modellbildung wurde von einem vollständig beobachtbaren Zustand des Systems ausgegangen. Dies ist jedoch in vielen Spielen nicht gegeben. Das populären Genre der Kartenspiele, in welchen zumeist die Karten der Gegenspieler nicht eingesehen werden können, ist nur eines von vielen Beispielsystemen in welchen der Agent aus unvollständigen Informationen den Zustand des Systems ermitteln muss.

Diese Problematik wurde anhand des digitalen Kartenspiels Hearthstone untersucht, welches durch seine mehr als 2000 verschiedenen Karten mit teils einzigartigen Effekten ein äußerst komplexes Spiel darstellt. Hierbei wählt jeder Spieler 30 Karten um ein individuelles Deck zu konstruieren. Aufgrund von häufig auftretenden Kartenkombinationen ist es möglich, die Karten des gegnerischen Decks anhand der ersten Spielrunden zu erahnen. Zum effektiven Einsatz eines Suchverfahrens bildet der Agent ein Modell zur Vorhersage der restlichen Karten des Gegenspielers und somit zur Vervollständigung der Zustandsbeobachtung. Hierfür wurden zwei Modelle entwickelt, welche in den folgenden Abschnitten näher erläutert werden.

Im ersten vorgestellten Verfahren wird angenommen, dass jede vom Gegner gespielte Karte unabhängige Informationen über die nächste zu spielende Karte liefert. Hierbei wird davon ausgegangen, dass Karten, die in der Vergangenheit bereits häufig zusammen auftraten, dies auch in Zukunft tun werden. Die Idee ist durch eine Arbeit von Elie Bursztein [Bu16] motiviert und wurde im Rahmen der Dissertation um vier weitere Varianten zur Zählung gemeinsam auftretender Karten ergänzt. Hierbei wurde eine Datenbank von Decks menschlicher Spieler analysiert um die jeweiligen Häufigkeitswerte zu bestimmen. Aufgrund der Vielzahl an Karten wurde die Zählung auf Bigramme begrenzt. Dies birgt den Nachteil, dass größere Zusammenhänge der Deck-Gestaltung nicht abgebildet werden können.

Um die Art des gegnerischen Decks zu erfassen, wurde die Datenbank bestehender Decks im Rahmen eines neu entwickelten Fuzzy Multiset Clustering Verfahrens auf Deck-Archetypen

reduziert. Hierbei wird jedes Deck als Multiset beschrieben, wobei die Kombination ähnlicher Decks als Fuzzy Multiset [Ya86] beschrieben werden kann. Für ein solches wurde ein Clustering Verfahren entworfen [DSK08], welches in prototypischen Deck-Zusammenstellungen resultiert. Durch die Definition der Distanz eines Prototypens zur Menge aller vom Gegner bereits gespielten Karten ist es möglich, die Art des gegnerischen Decks abzuschätzen und die übrigen Karten des Gegners vorherzusagen [DK20c].

Beide Verfahren ermöglichen es, Vorhersagen über die gegnerischen Karten zu treffen. Da diese Abschätzung jedoch fehlerhaft sein kann, ist es nötig, diesen potentiellen Fehler auch innerhalb des Suchverfahrens abzubilden. Hierfür wurde der Ensemble-MCTS Algorithmus entwickelt [Do18], welcher einen separaten Suchlauf für jede Prädiktion durchführt und deren Ergebnisse aggregiert um eine robuste Entscheidung zu treffen.

5 Ergebnisse und Schlussfolgerungen

Die in meiner Dissertation vorgestellten Verfahren wurden anhand einer Vielzahl von Anwendungen evaluiert. Abschließend werden diese und darauf aufbauende Studien kurz vorgestellt und deren Bedeutung für das Forschungsgebiet hervorgehoben.

Die in den Abschnitten 2 und 3 vorgestellten Model Learning Verfahren wurden anhand von 30 Spielen des General Video Game AI (GVGAI) Frameworks evaluiert [Pe19]. Hierbei lernte der Agent die Spiele innerhalb von 100 Runden zu spielen ohne jegliches Vorwissen. Insofern die Unabhängigkeitsannahmen der Modellheuristiken erfüllt waren, erzielten die gelernten Modelle eine nahezu perfekte Vorhersage der Folgezustände und ermöglichten den effizienten Einsatz getesteter Suchverfahren. Die resultierenden Agenten übertrafen deutlich die zuvor besten Agenten des mit dem Frameworks assoziierten GVGAI-Wettbewerbs.

Methoden der Predictive State Determinization wurden anhand des zuvor beschriebenen Spiels Hearthstone evaluiert. Da für diese Anwendung keine vergleichbaren Studien existierten, wurde der Hearthstone AI Wettbewerb von mir ins Leben gerufen. Als Teil der IEEE Conference on Games konnten in 3 Wettbewerbsjahren insgesamt 100 Einreichungen von Wissenschaftlern aus der ganzen Welt gesammelt werden. In einem Vergleich mit den besten Einreichungen der Jahre 2018 und 2019 erzielte das von mir entwickelte Verfahren die besten Leistungen.

In Folge der Dissertation konnten die vorgestellten Methoden weiterentwickelt und auf weitere Anwendungsfelder übertragen werden. Durch den Einsatz von Active Learning [DK20a] wird der Trainingsprozess weiter beschleunigt. Die Einbindung der Modellsicherheit in den Planungsprozess führte zu robusteren Ergebnissen während des Trainings und der Evaluierung [DK20a]. Während die präsentierten Model Learning Verfahren noch auf diskrete Zustands- und Aktionsräume limitiert waren, konnte eine Erweiterung auf kontinuierliche Räume abgeleitet werden [DK20b]. Diese führte zum erfolgreichen Einsatz der vorgestellten Verfahren im Kontext einer Auswahl von Motion Control Szenarien.

Die präsentierten Verfahren stellen einen wichtigen Beitrag für Model Learning und Vorhersage-basierter Suche dar. Im Gegensatz zu populären Deep Learning Verfahren ist

das Ergebnis des Modells und der Suche interpretierbar. Durch die strukturelle Analyse von Abhängigkeiten der Ein- und Ausgabedaten können selbst komplexe Zustandsübergangsmodelle effizient repräsentiert und erlernt werden. Durch die Modellierung der Sicherheit des resultierenden Modells zeigt sich die Vorhersage-basierte Suche robust gegenüber unvorhergesehenen Zuständen des Systems und trägt damit maßgeblich zum sicheren Einsatz in der Praxis bei. Die Generalisierbarkeit des Verfahrens auf weitere Anwendungsgebiete konnte bereits anhand von Szenarien der Robotik demonstriert werden [DK20b] und wird in zukünftigen Arbeiten weiter verfolgt.

Literaturverzeichnis

[Br12] Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S.: A Survey of Monte Carlo Tree Search Methods. IEEE Transactions on Computational Intelligence and AI in Games, 4(1):1–43, 2012.

[Bu16] Bursztein, Elie: I am a legend: Hacking hearthstone using statistical learning methods. In: 2016 IEEE Conference on Computational Intelligence and Games (CIG). IEEE, S. 1–8, sep 2016.

[DK20a] Dockhorn, Alexander; Kruse, Rudolf: Balancing Exploration And Exploitation in Forward Model Learning. In: Advances in Intelligent Systems Research and Innovation, S. 1–20. Elsevier, 2020.

[DK20b] Dockhorn, Alexander; Kruse, Rudolf: Forward Model Learning for Motion Control Tasks. In: 2020 IEEE 10th International Conference on Intelligent Systems (IS). IEEE, S. 1–5, 8 2020.

[DK20c] Dockhorn, Alexander; Kruse, Rudolf: Predicting cards using a fuzzy multiset clustering of decks. International Journal of Computational Intelligence Systems, 13(1):1207–1217, 2020.

[Do18] Dockhorn, Alexander; Frick, Max; Akkaya, Ünal; Kruse, Rudolf: Predicting Opponent Moves for Improving Hearthstone AI. In (Medina, Jesus; Ojeda-Aciego, Manuel; Verdegay, Jose Luis; Pelta, David A.; Cabrera, Inma P.; Bouchon-Meunier, Bernadette; Yager, Ronald R., Hrsg.): 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2018. Springer International Publishing, S. 621–632, 2018.

[Do19] Dockhorn, Alexander; Lucas, Simon M; Volz, Vanessa; Bravi, Ivan; Gaina, Raluca D; Perez-Liebana, Diego: Learning Local Forward Models on Unforgiving Games. In: 2019 IEEE Conference on Games (CoG). IEEE, London, S. 1–4, 8 2019.

[Do20] Dockhorn, Alexander: Prediction-based Search for Autonomous Game-Playing. Dissertation, Otto von Guericke University Magdeburg, 2020.

[DSK08] Dockhorn, Alexander; Schwensfeier, Tony; Kruse, Rudolf: Fuzzy Multiset Clustering for Metagame Analysis. In: Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). Atlantis Press, S. 536–543, 2019/08.

[DTK18] Dockhorn, Alexander; Tippelt, Tim; Kruse, Rudolf: Model Decomposition for Forward Model Approximation. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, S. 1751–1757, 11 2018.

[GBC16] Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron: Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

[GLP05] Genesereth, Michael; Love, Nathaniel; Pell, Barney: General Game Playing: Overview of the AAAI Competition. AI Magazin, 26(2):62–72, 2005.

[GLP17] Gaina, R. D.; Lucas, S. M.; Perez-Liebana, D.: Rolling horizon evolution enhancements in general video game playing. In: 2017 IEEE Conference on Computational Intelligence and Games (CIG). S. 88–95, 2017.

[HS18] Ha, David; Schmidhuber, Jürgen: Recurrent World Models Facilitate Policy Evolution. In (Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; Garnett, R., Hrsg.): Advances in Neural Information Processing Systems. Jgg. 31. Curran Associates, Inc., S. 2450–2462, 2018.

[Le13] Levine, John; Congdon, Cb; Ebner, Marc; Kendall, Graham: General video game playing. Dagstuhl Follow-Ups, S. 1–7, 2013.

[Lu19] Lucas, Simon M; Dockhorn, Alexander; Volz, Vanessa; Bamford, Chris; Gaina, Raluca D; Bravi, Ivan; Perez-Liebana, Diego; Mostaghim, Sanaz; Kruse, Rudolf: A Local Approach to Forward Model Learning: Results on the Game of Life Game. In: 2019 IEEE Conference on Games (CoG). IEEE, S. 1–8, 8 2019.

[Mn15] Mnih, Volodymyr; Kavukcuoglu, Koray; Silver, David; Rusu, Andrei a; Veness, Joel; Bellemare, Marc G; Graves, Alex; Riedmiller, Martin; Fidjeland, Andreas K; Ostrovski, Georg; Petersen, Stig; Beattie, Charles; Sadik, Amir; Antonoglou, Ioannis; King, Helen; Kumaran, Dharshan; Wierstra, Daan; Legg, Shane; Hassabis, Demis: Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2 2015.

[Pe19] Perez-Liebana, Diego; Lucas, Simon M.; Gaina, Raluca D.; Togelius, Julian; Khalifa, Ahmed; Liu, Jialin: General Video Game Artificial Intelligence, Jgg. 3. Morgan & Claypool Publishers, 2019. https://gaigresearch.github.io/gvgaibook/.

[SB18] Sutton, Richard S.; Barto, Andrew G.: Reinforcement Learning. The MIT Press, Cambridge, 2. Auflage, 2018.

[Ya86] Yager, Ronald R.: On the Theory of Bags. International Journal of General Systems, 13(1):23–37, nov 1986.

Alexander Dockhorn ist Postdoctoral Research Associate an der Queen Mary University of London. Seinen Doktortitel erhielt er an der Otto-von-Guericke-Universität in Magdeburg. In seiner Forschung untersucht er die Fähigkeiten von prädiktionsbasierten Suchagenten in Spielen mit einem besonderen Interesse zur Modellierung von Unsicherheiten. Als Mitglied des Institute of Electrical and Electronics Engineers (IEEE) beteiligt er sich aktiv an der Mitgestaltung der Forschungslandschaft und fungiert als Vorsitzender des IEEE CIS Competitions Sub-Committees, sowie als Mitglied zahlreicher weiterer Kommittees. Seit 2017 organisiert er den Hearthstone-KI-Wettbewerb, um die Vergleichbarkeit von KI-Agenten in Kartenspielen zu fördern. Eine vollständige Liste seiner Projekte und Veröffentlichungen finden Sie auf seiner Webseite: https://adockhorn.github.io/

Konzepte und Schnittstellen für eine Nahtlose Interaktion zwischen Virtueller und Physischer Realität[1]

Ceenu George[2]

Abstract: Virtual Reality (VR, Virtuelle Realität) ermöglicht es Benutzern, realitätsnahe Modelle von Produkten zu erstellen und eine immersive soziale Interaktion mit entfernten Kollegen zu erleben. VR nutzt unsere visuelle Dominanz, um diese Erfahrungen zu vermitteln, und versucht Nutzer davon zu überzeugen, dass sie sich in einer anderen Realität befinden. Während ihr Bewusstsein jedoch in VR präsent ist, befindet sich ihr Körper in der physische Realität (PR). Da der Nutzer die PR nicht sehen kann, bringt dies erhebliche Unsicherheiten in die Interaktion. In dieser Arbeit gehe ich auf diese Unsicherheit ein, indem ich Konzepte und Schnittstellen entwickle, die es dem Nutzer ermöglichen, in VR zu bleiben und gleichzeitig ein Bewusstsein für die PR zu behalten. Sie behalten dieses Bewusstsein bei, ohne das Head-mounted-display (HMD) abnehmen zu müssen. Ich bezeichne das als *nahtlose* Interaktion mit der PR. Die übergreifende Forschungsvision dieser Arbeit ist daher, die Trennung zwischen der virtuellen und der physischen Realität zu reduzieren.

1 Einführung

Die Forschung an Virtueller Realität (VR) hat seit der Einführung von kostengünstigen HMDs für Verbraucher im Jahr 2016 durch Oculus und HTC einen Anstieg an Interesse erfahren. Obwohl es das Konzept von HMDs schon seit Sutherland's Einführung des "Schwertes von Damokles" im Jahr 1968 gibt [Ma99], wurde es bis vor kurzem vorwiegend in Forschungslabors eingesetzt. Mit dieser Entwicklung vom Labor- zum Verbraucherprodukt entstanden neue Nutzungsmöglichkeiten: Museen haben VR-Erlebnisse implementiert, die es den Benutzern ermöglichen Artefakte immersiver zu erleben oder sie bequem von zu Hause aus zu betrachten. Des weiteren nutzt die Industrie VR als erschwingliche und schnelle Alternative für die Prototypenentwicklung (z. B. Automobilbranche). In solchen Situationen dient VR als Erweiterung zu bestehenden physischen Bildschirmen - PC oder Mobiltelefon - und bietet ein immersiveres Erlebnis mit einem größeren Sichtfeld.

Da VR-fähige HMDs kabellos werden, stehen sie vor ähnlichen Herausforderungen wie etablierte, ubiquitäre Geräte wie Mobiltelefone. Eine dieser Herausforderungen ist die Trennung zwischen der virtuellen und der physischen Realität [AM00]. Diese wird mit immersiven HMDs verstärkt, da die Nutzer sowohl mental als auch visuell von der realen Welt getrennt sind. Der Grad der wahrgenommenen Trennung wird sogar als Erfolgsmaßstab für VR-Anwendungen etabliert – allgemein bezeichnet als *Präsenz* [Sl96].

[1] Englischer Titel der Dissertation: Virtual Reality Interfaces for Seamless Interaction with the Physical Reality [Ge20a]

[2] Medieninformatik, LMU München, ceenu.george@ifi.lmu.de

Obwohl der Wechsel zwischen der physischen und der virtuellen Realität eine Herausforderung für etablierte, ubiquitäre Systeme bleibt (z.B. beim SMS schreiben oder Autofahren), ermöglicht die begrenzte Bildschirmfläche solcher Geräte ein Bewusstsein für beide Realitäten. Beispielsweise bietet der Bildschirm eines Mobiltelefons dem Nutzer die Möglichkeit ständig um sich zu schauen, um zu sehen, ob sich jemand nähert oder um ein unerwartetes Geräusch zu überprüfen. Bei HMDs ist dies jedoch nur möglich, indem man den Nutzer zwingt sein Headset abzunehmen, welches die *Präsenz* beeinträchtigt [Sl96].

Um diese Beeinträchtigung zu vermeiden, habe ich Konzepte und Schnittstellen entwickelt, die eine Interaktion mit der PR ermöglichen, ohne dass das Headset abgenommen werden muss - Ich bezeichne das als *nahtlose* Interaktion mit der PR. Anstatt sich von einer Realität zu trennen, um in die andere einzutauchen, unterstützen meine Konzepte und Schnittstellen den Nutzer in VR präsent zu sein, während er sich der PR bewusst bleibt. Die übergreifende Forschungsvision, die diese Arbeit leitet, ist daher, diese Trennung zwischen der virtuellen und der physischen Realität zu reduzieren.

Um ein tieferes Wissen über das HMD-Nutzerverhalten zu erlangen, habe ich am Anfang meiner Promotion eine einleitende, qualitative Forschung durchgeführt. Die Resultate dieser Forschung bilden die Struktur meiner Dissertation und werden im nächsten Kapitel vorgestellt.

2 Einleitende Forschung

Im vorherigen Abschnitt habe ich die Notwendigkeit hervorgehoben, Schnittstellen und Konzepte zu untersuchen, die eine nahtlose Interaktion zwischen virtueller und physischer Realität ermöglichen. Ich habe dies durch frühere Arbeiten motiviert, bei einem nutzerzentrierten Ansatz ist es jedoch unerlässlich, die Details eines solchen Bedarfs in Form einer Feldstudie mit möglichen, zukünftigen Benutzern zu untersuchen.

RQ0: Was versteht ein Laie unter der Nutzung von virtueller Realität (VR) und welche Erwartungen hat ein Laie daran?

Um RQ0 zu untersuchen, führten wir eine einleitende Untersuchung (N=34) durch [GSH19]. Aus dieser Untersuchung ergaben sich eine Reihe von Themen, die die Grundlage meiner Forschung sind und die Struktur dieser Arbeit bestimmen:

- Fokus 1: Nutzbare Sicherheit und Privatsphäre für VR: Beobachtbarkeit von HMD-Benutzerinteraktionen
- Fokus 2: Kommunikation mit umstehenden Kollegen
- Fokus 3: Verwaltung der Präsenz in der physischen und virtuellen Realität

3 Vertiefende Forschung

3.1 Fokus 1: Nutzbare Sicherheit und Privatsphäre für VR: Beobachtbarkeit von HMD-Benutzerinteraktionen

Die Privatsphäre in VR wurde in unserer einleitenden Untersuchung [GSH19] von unseren Teilnehmern thematisiert. Wir fanden heraus, dass Nutzer sich Sorgen machten, von umstehenden Personen während der Interaktion beobachtet zu werden. Die Beobachtbarkeit durch Umstehende wurde jedoch nicht ausschließlich als negativ empfunden. Die Beziehung zum Beobachter spielte eine Rolle dabei, ob die Beobachtbarkeit als positiv oder negativ eingestuft wurde.

In diesem Abschnitt möchte ich unsere Arbeit zu den negativen Aspekten der Beobachtbarkeit vorstellen, die innerhalb des HCI-Forschungsgebiets der *Nutzbaren Sicherheit und Privatsphäre* angesiedelt ist. Positive Aspekte der Beobachtbarkeit werden in Fokus 2 behandelt.

Beobachtbarkeit wird innerhalb der Nutzbaren Sicherheit- und Privatsphären-Forschung als *Shoulder-surfing* bezeichnet. Es wird hierbei eine Situation umschrieben, in der ein Angreifer den Bildschirm eines Laptop- oder Smartphone-Benutzers beobachtet, ohne dass dieser davon weiß. Dies geschieht in der Regel außerhalb des unmittelbaren Sichtfelds des Benutzers, oft von hinten und über die Schulter. Bislang gibt es nur wenige Empfehlungen für HMD-Interaktionen in diesem Zusammenhang. Daher ist das erste Ziel dieser Arbeit zu verstehen, welche Interaktionen/Gesten eines HMD-Benutzers von einem Unbeteiligten beobachtet werden können.

RQ1 A: Welche Tätigkeiten eines HMD-Benutzers können von einem Umstehenden beobachtet werden?

Eiband et al. fanden heraus, dass bei Smartphones in 9% der Shoulder-surfing-Fälle Anmeldedaten und Passwörter des Smartphonenutzers von Umstehenden beobachtet werden konnten [Ei17]. Gezielte Angriffe auf Anmeldedaten und Passwörter werden vor allem im Zusammenhang mit ubiquitären Geräten untersucht, da Nutzer diese an öffentlichen Orten mit unbekannten Personen verwenden können (z.B. in öffentlichen Verkehrsmitteln). Dabei wurde gezeigt, dass das Risiko von Shoulder-surfing als gering wahrgenommen wird, unabhängig davon, wie hoch es tatsächlich ist [Ha14]. Daher muss das Shoulder-surfingrisiko durch das System gemildert werden, anstatt sich darauf zu verlassen, dass der Nutzer Maßnahmen ergreift.

Im Kontext von HMDs ist es unklar, ob frühere Arbeiten an ubiquitären Geräten auf VR übertragen werden können und ob neue Authentifizierungskonzepte aus einer Nutzbaren Sicherheitsperspektive entworfen werden müssen.

RQ1 B: Wie kann man Authentifizierungsmechanismen für VR-Nutzer erstellen und testen, die sowohl nutzbar als auch sicher sind?

Beitrag. Um RQ1a zu erforschen haben wir untersucht, welche Tätigkeiten von Umstehenden beobachtet werden können [Ge19a]. Die Tätigkeiten basierten auf gängigen Tätigkeiten, die derzeit auf ubiquitären Geräten erledigt werden, wie sie aus einer von uns durchgeführten Tagebuchstudie [Ge19a] abgeleitet wurden. Basierend auf diesen Ergebnissen haben wir fünf Tätigkeiten in unsere Studie aufgenommen: Anschauen eines Videos, Tippen eines Textes, Authentifizieren, Lesen eines Textes und 3D-Manipulation. Wir fanden heraus, dass Umstehende in der Lage waren, sowohl Tätigkeitswechsel (83% Erfolgsquote) als auch die individuellen Taten (77% Erfolgsquote) innerhalb weniger Sekunden nach dem Tätigkeitswechsel erfolgreich zu identifizieren. Wir geben auch Design-Empfehlungen (z.B. Pointing vs. TapPINg zur Erhöhung der Sicherheit) für die Implementierung von weniger beobachtbaren VR-Interaktionen und schlagen Konzepte vor, die die Privatsphäre von HMD-Nutzern in kollaborativen, physischen Arbeitsräumen mit nicht-HMD Nutzern gewährleisten.

Im Bezug auf RQ1b untersuchen wir, ob bestehende Authentifizierungsmethoden von ubiquitären Geräten, wie PIN und Muster von Mobiltelefonen, auf VR übertragen werden können [Ge17]. Wir fanden heraus, dass sie ähnlich nutzbar und gleichzeitig sicherer sind. Zusätzliche Faktoren, die speziell für HMDs gelten, wurden ebenfalls untersucht: Ein 2D-Handybildschirm bietet nur beobachtbare Interaktionen mit Gesten, während das HMD zusätzliche Eingabemodalitäten, wie z.B. einen Laserpointer, bietet. In ähnlicher Weise ist VR nicht durch die Größe es physischen Bildschirms begrenzt und bietet somit die Möglichkeit, unterschiedliche Größen/Abstände von Eingabemethoden für die PIN-/Mustereingabe und Feedback zu erforschen. Um den 3D-Raum für die Authentifizierung zu erforschen, haben wir iterativ einen neuartigen Authentifizierungsmechanismus für VR entworfen - genannt *RoomLock* [Ge19b]. Innerhalb dieses Systems müssen die Teilnehmer eine Reihe virtueller Objekte, die in einem dreidimensionalen virtuellen Raum platziert sind, als Passwort auswählen. Wir fanden heraus, dass *RoomLock* vergleichbar benutzbar und gleichzeitig sicherer – weniger shoulder surfing Möglicheiten – ist, als die übertragen Lösungen aus [Ge17]. Da die Kopf- und Blickinteraktion auf den meisten HMDs verfügbar ist, haben wir schließlich untersucht, ob wir die Sicherheit verbessern können, indem wir sie als zusätzliche Eingabemodalität innerhalb unseres neuartigen Authentifizierungsmechanismus *RoomLock* [Ge20b] implementieren. Unsere Ergebnisse zeigten, dass es zwar die Sicherheit verbessert, aber weniger benutzbar ist als die zuvor getesteten Eingabemodalitäten.

3.2 Fokus 2: Kommunikation mit umstehenden Kollegen

In unserer einleitenden Forschung wurden umstehende, physische Kollegen häufig thematisiert [GSH19]. Diese Menschen tauchten unerwartet, während der HMD-Nutzung auf. Wenn die Umstehenden Fremde waren, wurde dies als negatives Erlebnis wahrgenommen. Für die Nutzung von HMDs ist es unklar wie man physische Umstehende in VR darstellt.

Des Weiteren hat die bisherige Forschung gezeigt, dass die alleinige Existenz von ubiquitären Geräten die Art und Weise verändert wie Menschen sich gegenseitig unterbrechen und kommunizieren [ML02].

Im Kontext von HMDs ist jedoch unklar, wie Umstehende einen HMD-Nutzer unterbrechen und wie die Unterbrechung in VR visualisiert werden kann. Um diese Lücken zu erforschen, untersucht der zweite Schwerpunkt meiner Arbeit die nahtlose Unterbrechung und Kommunikation zwischen HMD-Nutzern und Umstehenden.

RQ2 A: Wie kann man Unterbrechungen von HMD-Nutzern durch Umstehende unterstützen?

In einer kollaborativen Umgebung treten Unterbrechungen auf, um einen Kommunikationskanal zwischen dem Unterbrecher und dem Unterbrochenen herzustellen; in unserem Fall zwischen dem Umstehenden und dem HMD-Nutzer. Um eine nahtlose Kommunikation zu ermöglichen und die Präsenz in der VR aufrecht zu erhalten, kann der HMD-Nutzer das Headset aufgesetzt lassen, um zu kommunizieren. Dies wurde jedoch als störend empfunden und führte dazu, dass sich Umstehende ausgeschlossen fühlten [Sc18]. Eine Alternative ist, dass sich beide Gesprächspartner in der VR befinden, sei es mit einem HMD oder mit weniger immersiven Geräten wie einem Tablet oder Smartphone. Es gibt nur limitierte Forschung über die virtuelle Repräsentation eines umstehenden Kollegen in einer Mixed-Reality-Umgebung mit VR-Benutzern.

RQ2 B: Wie wirkt sich die virtuelle Repräsentation eines Umstehend auf Faktoren aus, die die Zusammenarbeit zwischen HMD-Nutzer und umstehenden Kollegen beeinflussen, wie z.B. Vertrauen, Leistung, soziale Präsenz?

Beitrag. Um RQ2a zu untersuchen, haben wir uns sowohl angesehen, wie Umstehende einen HMD-Nutzer unterbrechen, als auch ob sie allein durch die Interpretation der Gesten des HMD-Nutzers in der Lage sind, dessen Tätigkeitswechsel zu beobachten [Ge19a].

Wir fanden heraus, dass Umstehende dazu tendierten, HMD-Nutzer *während* ihren Tätigkeiten mündlich zu unterbrechen. Außerdem verbesserten Umstehende ihre Art zu unterbrechen nach mehrfachen Wiederholungen. Dies zeigt, dass sie nach dem Training in der Lage sind, in günstigen Momenten zu unterbrechen (z.B. zwischen einem Tätigkeitswechsel). Basierend auf diesen Ergebnissen schlagen wir Interaktionskonzepte vor, die es Umstehenden ermöglichen, günstige Momente für die Unterbrechung zu wählen, zum Beispiel auffälligen/sichtbaren vs. subtilen Gesten. Wir schlagen auch Konzepte vor, die das Layout der PR einbeziehen, wobei der Nutzer durch das Design von der VR Welt angeleitet wird, sich in eine bestimmte Richtung zu stellen. Abschließend haben wir eine Studie mit einer Detection-Response-Aufgabe durchgeführt, um zu verstehen, wie Unterbrechungen in der VR dargestellt werden können und inwieweit sie die Präsenz und Leistung des VR-Benutzers beeinflussen [GSH18].

Im Bezug auf RQ2b, werden in dieser Arbeit drei Publikationen vorgestellt. Wir untersuchten

den Effekt der virtuellen Repräsentation von Kollaborateuren in VR auf soziale Präsenz und Leistung [GDH18]. [Ge18] untersucht, wie wir Vertrauen in VR durch ein Vertrauensspiel messen können, während wir die Repräsentation der Spieler zwischen einem Roboter und einem Menschen ändern (Repräsentation als Roboter oder als Mensch). Schließlich untersuchten wir, ob eine Visualisierung der Echtzeit-Herzfrequenz beider Mitspieler, Präsenz und Leistung verbessert [GH19].

3.3 Fokus 3: Verwaltung der Präsenz in der physischen und virtuellen Realität

Der Konsum digitaler Medien, hat in den letzten Jahren stetig zugenommen. In Deutschland zum Beispiel verbringen Erwachsene fast vier Stunden pro Tag online [vA19] und in den USA sind es sechs Stunden [Ke19]. Wenn man davon ausgeht, dass die durchschnittliche Person acht Stunden am Tag ruht oder schläft, verbringen wir etwa ein Drittel unserer wachen Zeit in einer VR. Derzeit wird die Vermittlung dieser VR von Smartphones und PCs dominiert. Die Immersion in jede Art von VR wird überwiegend durch die physische Größe des Bildschirms beeinflusst wird. Andere Sinneskanäle einzubeziehen, zum Beispiel durch Audio, können das Gefühl des Eintauchens erhöhen. Es wurde jedoch festgestellt, dass der visuelle Sinn diese überwältigt [Wi08]. HMDs nutzen den visuellen Sinn, indem sie das Sichtfeld des Benutzers mit einem Display umschließen um ein immersiveres Erlebnis als andere ubiquitäre Systeme (z.B. Smartphones) zu bieten. Sie schaffen auf diese Weise unerforschte Herausforderungen, die wir im Folgenden zusammenfassen.

Erstens im Vergleich zwischen HMDs und Smartphones, bietet das Letztere dem Nutzer die Möglichkeit, sich der PR bewusst zu bleiben, während er in die VR eintaucht. Zum Beispiel kann ein Nutzer eine Wegbeschreibung auf Google Maps nachschlagen, während er auf der physischen Straße navigiert. Die Nutzer scheinen gelernt zu haben, ihre Präsenz zwischen der virtuellen und der physischen Realität auszubalancieren. Nutzer neigen dazu, ihre Fähigkeiten in diesem Zusammenhang zu überschätzen. Das kann sich negativ auf die körperliche Unversehrtheit und Sicherheit von ihnen selbst und von Umstehenden auswirken. Beispielsweise interagieren sie Gehens gleichzeitig auf dem Telefon [BRR12]. Diese Herausforderung ist bei HMDs prominenter, da der Nutzer vollständig von einem Bildschirm und damit nur von der VR umgeben ist.

Zweitens hat das Ausblenden von der PR einen negativen Effekt auf soziale Interaktionen, wie in Abschnitt 3.2 erwähnt.

Drittens ist die Präsenz ein Hauptmaß für die Qualität eines VR-Erlebnisses. Ein VR-Erlebnis wird als erfolgreich/gut empfunden, wenn ein hohes Maß an Präsenz erreicht wird. Wie wenig sich der Nutzer der PR bewusst ist, wird in gängigen Fragebögen als Maß genommen um diese Präsenz zu bestimmen. Es ist unklar ob ein balancieren der Präsenz möglich ist, da die existierenden Methodiken dies nicht in Betracht ziehen. Zusammenfassend wollen wir die folgende Forschungsfrage untersuchen:

RQ3: Wie können wir Nutzer dabei unterstützen, sich der PR bewusst zu sein und gleichzeitig ihre Präsenz in der VR aufrecht zu erhalten?

Beitrag. Um RQ3 zu erforschen, haben wir zuerst Experteninterviews durchgeführt, um herauszufinden, welche positiven und negativen Faktoren die Interaktion mit HMD-Nutzern beeinflussen und wie diese unterstützt bzw. überwunden werden können. Aus unserer Analyse der Antworten ergab sich der *seamless transition* (SeaT) Design-space. Der SeaT-design-space dient als Ausgangspunkt für Forscher und Praktiker, die Systeme schaffen wollen, die es Benutzern ermöglichen, nahtlos zwischen Realitäten und Zwischenzuständen von Realitäten gemäß Milgrams Reality-Virtuality-Kontinuum [Mi95] zu wechseln.

In einer zweiten Studie erstellten wir iterativ zwei exemplarische Lösungen, die aus dem SeaT-design-space hervorgehen; nämlich das *sky portal* und das *virtual phone*. Beide sind Fenster zur anderen Realität. Das Ziel war es, Lösungen zu schaffen, die alle Dimensionen des SeaT-design-space berücksichtigen, anstatt sich auf bestimmte Dimensionen zu konzentrieren. Unsere Ergebnisse zeigten, dass das *sky portal* schlechter abschnitt als die Baseline; die Kamera der HTC Vive [Co20]. Das *virtual phone* war in den Punkten Leistung und Präsenz jedoch vergleichbar mit der Baseline und die Teilnehmer empfanden es als eine Verbesserung gegenüber der Baseline. Wir argumentieren, dass die Teilnehmer eine Lösung bevorzugen, die es ihnen ermöglicht, einen Blick in die andere Realität zu werfen, ohne die Realität, in der sie sich gerade befinden, komplett zu verlassen. Die ersten beiden Studien sind in [GNH20] veröffentlicht.

In einer dritten Studie untersuchten wir die Einbeziehung von akustischen/haptischen Signalen, um Nutzer über physische Grenzen zu benachrichtigen[GTH20]. Derzeit werden Nutzer über ihren visuellen Kanal vor physischen Grenzen gewarnt, zum Beispiel durch die eingebaute Mesh-Lösung in gängigen HMDs. Basierend auf Wickens Würfel, der nahelegt, dass die Zuweisung von Sinnen zu unterschiedlichen Tätigkeiten ohne einen signifikanten Anstieg der kognitiven Anforderungen möglich ist [Wi08], haben wir untersucht, ob Sinne bestimmten Realitäten zugewiesen werden können. Während also der visuelle Sinn auf die VR fokussiert ist, dienen der auditive und der haptische Sinn als Indikatoren für physische Grenzen. Wir fanden heraus, dass die haptischen/akustischen Signale vergleichbar mit visuellen Indikatoren für physische Grenzen sind.

In einer vierten Studie übertrugen wir die oben genannten Konzepte und untersuchten die Notwendigkeit, physische Grenzen in einer begrenzten, sitzenden Umgebung, wie z. B. auf dem Rücksitz eines Autos, anzuzeigen [Li20].

4 Schlusswort

In den vorherigen Kapiteln habe ich insgesamt 12 peer-reviewed Publikationen im Forschungsbereich Mensch-Maschine Interaktion/VR vorgestellt, die meine Dissertation umfasst.

Die Publikationen in Fokus 1 sind eine der ersten und am häufigsten zitierten Arbeiten im Bereich Nutzerzentrierte Sicherheit und Privatsphäre für VR. Sie zeigen, dass Forschung in diesem Bereich notwendig ist um die Privatsphäre des Nutzers zu stärken. Neben Beobachtbarkeit und Authentifizierung, legen die Diskussionen in diesem Fokus auch offen, dass dieser Bereich noch viele ungelöste Fragen in sich birgt. Es ist unklar, wie man Interaktionsparadigmen aus existierenden virtuellen Erlebnissen (z.B. Internet-browsing auf dem Smartphone) in eine immersive VR Erfahrung mit HMDs überträgt. Fokus 2 diskutiert Probleme, die eine Zusammenarbeit zwischen HMD-Nutzern und nicht-HMD Nutzern hervorbringen würde. Außerdem stellen wir konkrete, getestete Lösungen vor, die eine Kommunikation zwischen HMD-Nutzern und Umstehenden ermöglichen. Fokus 3 fasst Arbeiten im Bereich HMD-Nutzung mit physischen Umstehenden zusammen und stellt neue Lösungen vor, die es dem HMD-Nutzer ermöglichen zwischen den Realitäten zu wechseln ohne sein HMD abzunehmen. [GTH20] ist im renommierten IEEE Transactions on Visualization and Computer Graphics Journal (TVCG) publiziert (Akzeptanzrate: 5%). Die Resultate zeigen, dass das Balancieren der Präsenz ein komplexer Prozess ist, der weiterführende Forschungsfragen hervorbringt: Ist es möglich ein sicheres Erlebnis in VR zu haben und gleichzeitig Teil der PR zu sein? Wie beeinflusst das welche Sinne und ist es überhaupt möglich eine dauerhaft, ausgeglichenes Erlebnis zu schaffen, das den Nutzer nicht mental überfordert?

Die drei Foki zeigen, dass Forschung in diesem Bereich wichtig ist aus Nutzersicht und noch Probleme in sich birgt, die ungelöst sind. Dies wird bestätigt durch die zunehmende Popularität der Forschung in diesem Bereich, messbar an der ansteigenden Zahl der Zitate, die auf die Publikationen dieser Dissertation verweisen. Ich bin zuversichtlich, mit dieser Arbeit einen wesentlichen Beitrag für ein wichtiges Gebiet der zukünftigen Entwicklung der Informatik geleistet zu haben.

Literaturverzeichnis

[AM00] Abowd, Gregory D.; Mynatt, Elizabeth D.: Charting Past, Present, and Future Research in Ubiquitous Computing. ACM Trans. Comput.-Hum. Interact., 7(1):29–58, März 2000.

[BRR12] Basacik, Dan; Reed, Nick; Robbins, Randall L.: Smartphone Use While Driving: A Simulator Study. Published project report 592, Transport Research Laboratory, 2012.

[Co20] Corporation, HTC: , Activating front-facing camera. Website, May 2020. Retrieved April 2, 2020 from `https://www.vive.com/us/support/vive/category_howto/activating-the-front-facing-camera.html`.

[Ei17] Eiband, Malin; Khamis, Mohamed; von Zezschwitz, Emanuel; Hussmann, Heinrich; Alt, Florian: Understanding Shoulder Surfing in the Wild: Stories from Users and Observers. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, S. 4254–4265, 2017.

[GDH18] George, Ceenu; Demmler, Manuel; Hussmann, Heinrich: Intelligent Interruptions for IVR: Investigating the Interplay between Presence, Workload and Attention. In: Extended

Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18). ACM, 2018.

[Ge17] George, Ceenu; Khamis, Mohamed; von Zezschwitz, Emanuel; Burger, Marinus; Schmidt, Henri; Alt, Florian; Hussmann, Heinrich: Seamless and Secure VR: Adapting and Evaluating Established Authentication Systems for Virtual Reality. In: Proceedings of the Network and Distributed System Security Symposium (NDSS 2017). USEC '17. Internet Society, 2017.

[Ge18] George, Ceenu; Eiband, Malin; Hufnagel, Michael; Hussmann, Heinrich: Trusting Strangers in Immersive Virtual Reality. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion). ACM, S. 46:1–46:2, 2018.

[Ge19a] George, Ceenu; Janssen, Philipp; Heuss, David; Alt, Florian: Should I Interrupt or Not? Understanding Interruptions in Head-Mounted Display Settings. In: Proceedings of the 2019 on Designing Interactive Systems Conference (DIS'19). ACM, S. 497–510, 2019.

[Ge19b] George, Ceenu; Khamis, M.; Buschek, D.; Hussmann, H.: Investigating the Third Dimension for Authentication in Immersive Virtual Reality and in the Real World. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). S. 277–285, March 2019.

[Ge20a] George, Ceenu: , Virtual reality interfaces for seamless interaction with the physical reality. Doctorate thesis, September 2020.

[Ge20b] George, Ceenu; Buschek, Daniel; Ngao, Andrea; Khamis, Mohamed: GazeRoomLock: Using Gaze and Head-Pose to Improve the Usability and Observation Resistance of 3D Passwords in Virtual Reality. In (De Paolis, Lucio Tommaso; Bourdot, Patrick, Hrsg.): Augmented Reality, Virtual Reality, and Computer Graphics. Springer International Publishing, Cham, S. 61–81, 2020.

[GH19] George, Ceenu; Hassib, Mariam: Towards Augmenting IVR Communication with Physiological Sensing Data. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). ACM, 2019.

[GNH20] George, Ceenu; Ngo Tien, An; Hussmann, Heinrich: Seamless, Bi-directional Transitions along the Reality-Virtuality Continuum: A Conceptualization and Prototype Exploration. In: 2020 IEEE Symposium on Mixed and Augmented Reality. IEEE Computer Society, 2020.

[GSH18] George, Ceenu; Spitzer, Michael; Hussmann, Heinrich: Training in IVR: Investigating the Effect of Instructor Design on Social Presence and Performance of the VR User. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology (VRST'18). ACM, 2018.

[GSH19] George, Ceenu; Schwuchow, Julia; Hussmann, Heinrich: Fearing Disengagement from the Real World. In: 25th ACM Symposium on Virtual Reality Software and Technology. VRST '19, Association for Computing Machinery, New York, NY, USA, 2019.

[GTH20] George, Ceenu; Tamunjoh, Patrick; Hussmann, Heinrich: Invisible Boundaries for VR: Auditory and Haptic Signals as Indicators for Real World Boundaries. IEEE Transactions on Visualization and Computer Graphics, S. 3414–3422, 2020.

[Ha14] Harbach, Marian; von Zezschwitz, Emanuel; Fichtner, Andreas; Luca, Alexander De; Smith, Matthew: It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In: 10th Symposium On Usable Privacy and Security (SOUPS 2014). USENIX Association, Menlo Park, CA, S. 213–230, Juli 2014.

[Ke19] Kemp, Simon: , Digital trends 2019 every single stat you need to know about the Internet. article, January 2019. Retrieved April 22, 2020 from https://thenextweb.com/contributors/2019/01/30/digital-trends-2019-every-single-stat-you-need-to-know-about-the-internet/.

[Li20] Li, Jingyi; George, Ceenu; Ngao, Andrea; Holländer, Kai; Mayer, Stefan; Butz, Andreas: An Exploration of Users' Thoughts on Rear-Seat Productivity in Virtual Reality. In: Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2020). ACM, S. 92–95, 2020.

[Ma99] Mazuryk, Tomasz: , Virtual Reality History, Applications, Technology and Future. Technical report, 1999.

[Mi95] Milgram, Paul; Drascic, David; Grodski, Julius J.; Restogi, Anu; Zhai, Shumin; Zhou, Chin: Merging real and virtual worlds. In: Proceedings of IMAGINA. Jgg. 95, S. 218–230, 1995.

[ML02] McFarlane, Daniel C.; Latorella, Kara A.: The Scope and Importance of Human Interruption in Human-Computer Interaction Design. Human-Computer Interaction, 17(1):1–61, 2002.

[Sc18] Schwind, Valentin; Reinhardt, Jens; Rzayev, Rufat; Henze, Niels; Wolf, Katrin: Virtual Reality on the Go? A Study on Social Acceptance of VR Glasses. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18). ACM, S. 111–118, 2018.

[Sl96] Slater, Mel; Linakis, Vasilis; Usoh, Martin; Kooper, Rob; Street, Gower: Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. In: ACM virtual reality software and technology (VRST). Jgg. 163. ACM Press New York, NY, S. 72, 1996.

[vA19] von Abrams, Karin: , Germany Time Spent with Media. report, May 2019. Retrieved April 22, 2020 from https://www.emarketer.com/content/germany-time-spent-with-media-2019.

[Wi08] Wickens, Christopher D.: Multiple resources and mental workload. Human factors, 50(3):449–455, 2008.

Ceenu George ist eine Mensch-Maschine Interaktion Forscherin, die Virtual- und Augmented Reality Systeme entwickelt. Sie hat über 6 Jahre Industrieerfahrung und 5 Jahre Forschungserfahrung in der nutzer-zentrierten Entwicklung von Produkten und Prototypen. Sie hat ihre Promotion im September 2020 an der LMU München mit summa-cum laude abgeschlossen. Sie hält derzeit eine post-doc Position an der LMU München und ist externe Beraterin bei Google.

Angriffe durch Fernzugriff auf FPGA Hardware[1]

Dennis Rolf Engelbert Gnad[2]

Abstract: Field-Programmable Gate Arrays (FPGAs) werden neben Grafikkarten immer beliebtere Rechenbeschleuniger für Cloud Umgebungen verschiedenster Anbieter. Auf rein digitaler Ebene können FPGAs bereits recht gut abgesichert werden, während bisherige Angriffe auf elektrischer Ebene physischen Zugriff zum Gerät benötigten. Diese Dissertation [Gn20a] deckt nun eine Reihe von Sicherheitsproblemen auf, die zwar auf physikalischen Effekten innerhalb integrierter Schaltkreise basieren, jedoch auch mit Fernzugriff durch Software ausnutzbar sind. Die Arbeit betrachtet FPGA Chips verschiedener Hersteller und zeigt Angriffe die alle Sicherheitsanforderungen der Vertraulichkeit, Unversehrtheit, und Verfügbarkeit verletzen können. Gleichermaßen wird mit weiteren Angriffen auf Geräte des Internets der Dinge (Internet of Things; IoT) gezeigt, dass ähnliche Probleme nicht nur FPGAs, sondern auch andere integrierte Schaltkreise betreffen können. Dadurch motiviert diese Arbeit weitere Analysen und Lösungsansätze, an denen bereits aktiv gearbeitet wird.

1 Einführung

Weltweit sind immer mehr Computersysteme miteinander verbunden und über das Internet zugänglich, was sich auf deren Sicherheitsanforderungen auswirkt. Außerdem werden zunehmend spezialisierte heterogene Systeme mit Beschleunigern verwendet, um die Effizienz zu steigern. Eine Art dieser Beschleuniger die bereits aus anderen Anwendungsbereichen bekannt sind, sind *Field-Programmable Gate Arrays* (FPGAs). FPGAs sind sehr flexible Mikrochips, die ähnlich wie Software entwickelt werden können, und damit als *programmierbare Hardware* gelten. Seit einigen Jahren kommen sie in etablierten Cloud-Platformen zum Einsatz.

Wie andere integrierte Schaltkreise, basieren FPGAs auf modernen Halbleitertechnologien, die von Fertigungstoleranzen und Laufzeitschwankungen betroffen sind. Es ist bereits bekannt, dass diese Toleranzen und Schwankungen die Zuverlässigkeit eines Systems beeinflussen können und auch deren Auswirkungen auf die Sicherheit gegenüber lokalen Angreifern sind bekannt. Seitenkanalangriffe und Fault-Angriffe auf elektrischer Ebene nutzen solche Schwankungen aus, um aktiv einzuwirken oder passiv zu messen. Seitenkanalangriffe messen Strom oder Spannung zur selben Zeit in der mit geheimen Daten gearbeitet wird. Da diese Werte zu einem gewissen Teil datenabhängig sind, können sie etwas über die geheimen Informationen aussagen die verarbeitet werden, wozu jedoch oft statistische Verfahren notwendig sind [KJJ99]. Bei elektrischen Fault-Angriffen wird aktiv die Spannungsversorgung manipuliert, um Timing-Verletzungen in der digitalen Schaltung zu verursachen, die zu Berechnungsfehlern führen, und oft auch *Glitching* genannt wird. Dies geschieht mit dem Ziel, den anschließenden Fehler analysieren zu können, um ebenfalls

[1] Englischer Titel der Dissertation: “Remote Attacks on FPGA Hardware”

[2] Karlsruher Institut für Technologie, dennis.gnad@kit.edu

Geheimnisse zu extrahieren [BDL97], oder beispielsweise Sicherheits-Überprüfungen im Bootvorgang zu überspringen. Ein Fault-Angriff kann aber auch dazu führen ein System zum Absturz zu bringen, und dadurch Denial-of-Service (DoS) zu verursachen.

In beiden Angriffsarten wird normalerweise ein Elektronik-Labor mit Zugang zu Messgeräten und Testinstrumenten benötigt, um den entsprechenden Mikrochip anzugreifen. Allerdings zählen sie auch zu den mächtigsten Angriffen, da auf unterster Systemebene Sicherheitsmechanismen des gesamten Systems ausgehebelt werden können. Was zuvor noch kaum beobachtet werden konnte, sind ähnliche Angriffe, die aus der Ferne durchgeführt werden können. Daher gehen die meisten momentanen Sicherheitsannahmen davon aus, dass derartige Angriffe nur mit lokalem physischen Zugriff möglich sind, und man sich nicht dagegen schützen muss, wenn das System ohnehin nicht physisch vom Angreifer zugänglich ist. In dieser Dissertation wird nun zum ersten Mal gezeigt, dass beide Klassen von Angriffen auch aus der Ferne möglich sind, in dem bestimmte Vorgehensweisen beim Erstellen der digitalen Schaltungen für FPGAs genutzt werden.

Um die möglichen Angriffe genauer einordnen zu können, folgt die Dissertation einem Modell eines Angreifers, was in Abbildung 1 zusammengefasst wird. Die Abbildung zeigt ein einzelnes System, welches von zwei Benutzern gemeinsam genutzt wird. Einer der Benutzer ist ein böswilliger Angreifer (Attacker), der versucht, 1) Fault-Angriffe oder 2) Seitenkanalangriffe auf das Opfer (Victim) durchzuführen, wobei diese Angriffe deswegen interessant werden, da Angreifer und Opfer im Modell logisch voneinander getrennt sind, bzw. gängigen Sicherheitsmodellen zu einer isolierten Kommunikation befolgen, wie von Huffmire et al. [Hu07] vorgestellt wurde. Weiterhin wird davon ausgegangen dass Angreifer und Opfer ein gemeinsames Stromversorgungsnetz (Power Distribution Network; PDN) – nutzen, was oft für einen realen integrierten Schaltkreis der Fall ist. Falls Angreifer und Opfer logisch auf dem Mikrochip isoliert sind, können daher dennoch Angriffe stattfinden, falls die elektrische Ebene genutzt werden kann. Da mittlerweile ein hoher Grad an Integration auf einem einzigen Mikrochip besteht, können solche Angriffe immer mehr Systeme betreffen, wie beispielsweise Systems on Chip (SoCs) die in vielen IoT Geräten

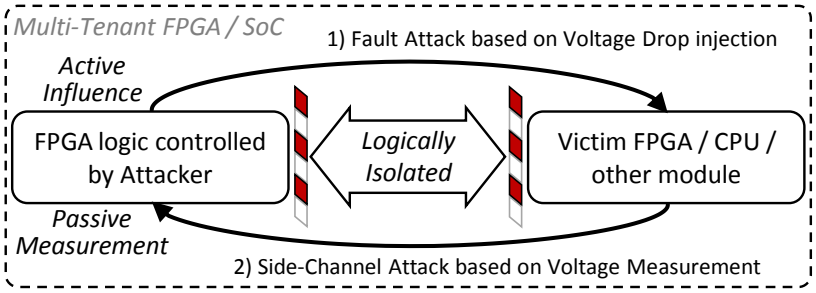


Abb. 1: Grundlegendes Modell eines Angreifers (Attacker). Auf der linken Seite befindet sich die Logik eines Angreifers, welcher durch aktiven Einfluss oder passive Messungen auf das Opfer (Victim), auf der rechten Seite wirkt, jedoch ohne eine logische Verbindung dazu haben zu müssen. Dies kann prinzipiell alle Systeme mit einer gemeinsamen Spannungsversorgung betreffen, die ein Angreifer entweder messen oder beeinflussen kann.

im Einsatz sind. Doch es kann auch ausreichend sein, dass mehrere Mikrochips in einem größeren System von derselben Stromversorgung versorgt werden, und damit potenzielle Sicherheitslücken entstehen.

Sämtliche Befürchtungen aus diesem Modell werden in der Dissertation durch praktische Experimente mit echten FPGAs und zum Teil auch IoT Geräten belegt. Dadurch zeigt diese Arbeit, dass die beschriebenen Angriffe nicht automatisch ausgeschlossen werden können, und es dringend erforderlich ist sie in einer umfassenden Sicherheitsbewertung ebenfalls zu berücksichtigen. Dies gilt besonders für FPGAs in Cloud Umgebungen, auf denen diese Angriffe nur einmal richtig entwickelt werden müssen, im Gegensatz zu Angriffen im Elektroniklabor allerdings gleich eine große Anzahl an Benutzern auf einen Schlag betreffen können. Diese Sicherheitsbedrohungen müssen zuerst gelöst werden, bevor ein sicherer Einsatz von virtualisierten FPGAs für mehrere Benutzer möglich ist.

2 Hinleitung

Vor Beginn der praktischen Arbeiten zur Dissertation gab es wenige Veröffentlichungen die sich mit schnell veränderlichen Spannungsschwankungen innerhalb integrierter Schaltkreise befasst haben. Da diese Schwankungen im Zeitbereich von Nanosekunden auftreten, sind sie schwer zu messen, und Simulationen für gesamte Mikrochips sind sehr zeitaufwändig und können nur sehr eingeschränkt benutzt werden. Zur Messung gibt es jedoch eine sehr kurze Arbeit von Zick et al. [Zi13], welche zeigt wie man mit digitaler FPGA Logik starke Spannungsschwankungen sichtbar machen kann. Zusätzlich hierzu ist allgemein bekannt, dass Ring Oszillatoren (ROs) prinzipiell Temperaturerhöhungen und Spannungsschwankungen in FPGAs erzeugen können.

Gleichzeitig bieten einige kommerzielle Cloud Anbieter FPGAs zur Nutzung an. Bisher werden die FPGAs jedoch noch komplett an einzelne Nutzer vermietet. Einige akademische und industrielle Veröffentlichungen deuten jedoch darauf hin, dass es sehr Effizienzfördernd ist, FPGAs unter mehreren Benutzern aufzuteilen [By14, FVS15, Kh18], genauso wie die Aufteilung von virtuellen CPUs und Hauptspeicherbereichen in klassischen Cloud Plattformen. Dies wird bereits grundlegend im Linux Kernel unterstützt [Ha]. Demnach sollte vorher genau untersucht werden, inwiefern sich darauf mögliche neue Sicherheitsprobleme auswirken.

3 Charakterisieren von Spannungsschwankungen innerhalb FPGAs

Die praktischen Arbeiten zur Dissertation beginnen mit einer experimentellen Analyse von Spannungsschwankungen innerhalb FPGAs, welche wir in [Gn16] zuerst veröffentlicht und in [Gn18a] erweitert haben. Hierzu wurden mehrere Sensoren die denen von Zick et al. [Zi13] ähnlich sind im FPGA verteilt. Diese Art Sensor wird in Abbildung 2 dargestellt und verwendet eine Reihe von FPGA Logikelementen die ausschließlich zur zeitlichen Signalverzögerung dienen sollen. Dies sind am Anfang LUT/Latch Elemente und später

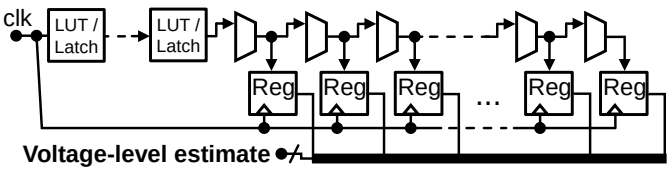


Abb. 2: Prinzip eines Sensors zum indirekten Messen von Spannung (Voltage-level estimate).

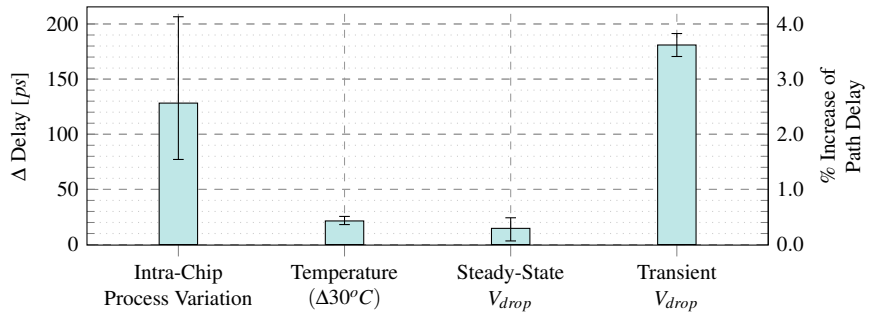


Abb. 3: Unterschiede in Verzögerungszeiten (ΔDelay) basierend auf Chip-internen Fertigungstoleranzen (Process Variation), Temperaturschwankungen, statischem Abfall der Spannung (Steady-State V_{drop}), und dynamischen Spannungsabfall (Transient V_{drop}). Gemessen mit acht Sensoren innerhalb eines Xilinx Virtex 6 FPGA (ML605 Board), in dem 8% der verfügbaren Flip-Flops verwendet wurden um Hitze oder Spannungsschwankungen zu erzeugen. Die Prozentuale Abnahme auf der rechten Ordinatenachse basiert auf der Zeitverzögerung des Pfads bei ansonsten inaktivem System.

Multiplexer, was aufgrund der verfügbaren Strukturen in FPGAs so zu den besten Ergebnissen führt. Zwischen die Multiplexer werden Register (Reg) geschaltet. Wenn nun der Takt (clk) wie abgebildet angelegt wird, kann in jedem Takt beobachtet werden, wie weit sich das Taktsignal durch die Elemente bewegt. Da dies spannungsabhängig ist, kann man so indirekt die Spannungsschwankungen aus den Registern auslesen. Mit zusätzlicher Kalibrierung, können verschiedene Einflüsse von Fertigungstoleranzen und Laufzeitschwankungen im FPGA verglichen werden. Weiterhin wurden einige oszillierene Flip-Flops im FPGA verteilt um eine Belastung im System simulieren zu können. In Abbildung 3 wird gezeigt, dass der größte Einfluss tatsächlich von Spannungsschwankungen (Transient V_{drop}) ausgeht. Diese Arbeiten enthalten zuvor unentdeckte Einblicke in das Verhalten von Spannungsschwankungen innerhalb eines integrierten Schaltkreises, welche von der Arbeitslast und räumlichen Verteilung innerhalb des FPGAs abhängig sind. Basierend auf den verwendeten Methoden in diesen Arbeiten werden im Folgenden die Sicherheitsimplikationen von Spannungsschwankungen in FPGAs analysiert.

4 Fault-Angriffe innerhalb FPGAs

Um Fault-Angriffe durchzuführen zeigt die Arbeit, wie ein böswilliger FPGA-Benutzer entweder einen DoS-Angriff auf das gesamte System durchführen kann [GOT17], oder präzise Fehler im Design eines anderen Benutzers auf dem FPGA verursachen kann [KGT18],

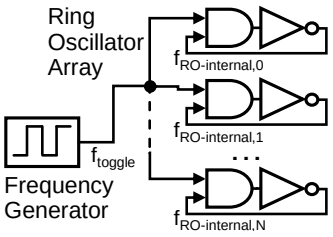


Abb. 4: Grundrinzip um mehrere Ring Oszillatoren (ROs) steuern zu können.

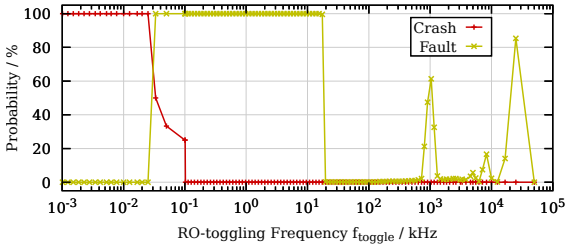


Abb. 5: Frequenzabhängige Wahrscheinlichkeit des Auftretens eines System-Crashs oder Fehlverhaltens auf einem Xilinx VCU108 Board.

was durch einen starken Spannungsabfall V_{drop} verursacht werden kann. Hierfür werden ROs in den FPGA konfiguriert wie in Abbildung 4 gezeigt ist. Die ROs haben eigene innere Frequenzen $f_{RO-internal,0..N}$, welche durch die direkte Rückkopplung jeweils so schnell oszillieren wie physikalisch möglich. Dadurch gibt es durch sehr hohe Schaltvorgänge einen sehr hohen Stromverbrauch $i(t)$. Da diese Frequenzen sehr hoch und leicht unterschiedlich sind, sind sie keine direkte Gefahr für das System. Für einen hohen Spannungsabfall im System kann dadurch gesorgt werden, wenn es spontane Belastungswechsel gibt. Eine hohe Stromveränderung in kurzer Zeit kann zu einem starken Abfall führen, und folgt dem Prinzip $V_{drop} = Ldi(t)/dt$. Nun kann f_{toggle} dafür verwendet werde alle ROs gemeinsam anzusteuern, und somit Lastwechsel auf verschiedenen Frequenzen zu erzeugen. Abhängig vom restlichen Aufbau, kann entweder Crash oder Fault alleine abhängig von der Frequenz sein, wie in Abbildung 5 gezeigt wird.

Der DoS-Angriff zielt auf den meisten getesteten Systemen auf das Netzteil der jeweiligen FPGA Karte (Board) ab. Dadurch ist ein kompletter Crash und Verlust der FPGA Konfiguration möglich. Danach kann es bei manchen FPGAs vorkommen, dass das Board vor Wiederverwendung zuerst nochmal komplett abgeschaltet werden muss, was bei einer PCIe-Beschleunigerkarte eine manuelle Abschaltung des PC-Netzteils bedeuten kann. Die präzise Beeinflussung eines anderen Benutzers im selben FPGA ist ausreichend, um Fehler in einem Modul für den Advanced Encryption Standard (AES) zu erzeugen und mit einer Differential Fault Analysis (DFA) [BDL97] die Geheimschlüssel zu extrahieren.

5 Seitenkanalangriffe innerhalb FPGAs

Um Seitenkanalangriffe innerhalb eines FPGAs auszuführen wird mit etwas Anpassung das Prinzip der zuvor vorgestellten Sensoren aus Abbildung 2 verwendet. Hierfür wird einer dieser Sensoren als Angreifer an einem Ende des FPGAs platziert, während ein AES Modul am anderen Ende des FPGAs das Opfer des Angriffs ist, wie in Abbildung 6 gezeigt wird. Während das FPGA Modul läuft, zeichnen wir nun mit dem Sensor Spannungsschwankungen auf, und speichern sie im internen FPGA Speicher (Block RAM; BRAM). Dieser Angriff wurde zusammen mit Forschern der Ruhr-Universität Bochum ausgearbeitet und in [Sc18a] publiziert.

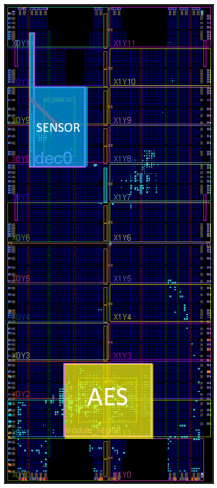


Abb. 6: Flurplan eines Spartan-6 FPGA in dem ein Angreifer einen Sensor betreibt, um die Spannungsschwankungen des AES Moduls zu messen. Ergebnisse dazu sind in Abbildung 8 gezeigt.

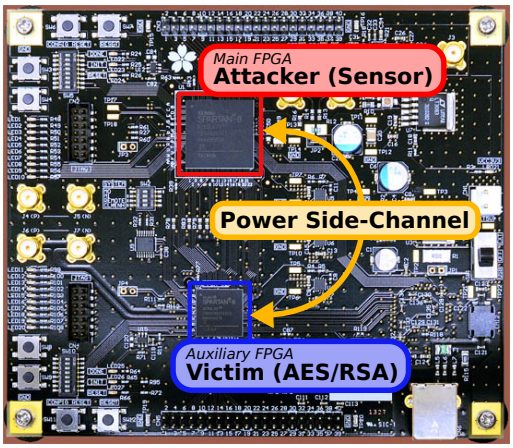


Abb. 7: SAKURA-G Platine in der die Funktionen der beiden FPGAs markiert sind.

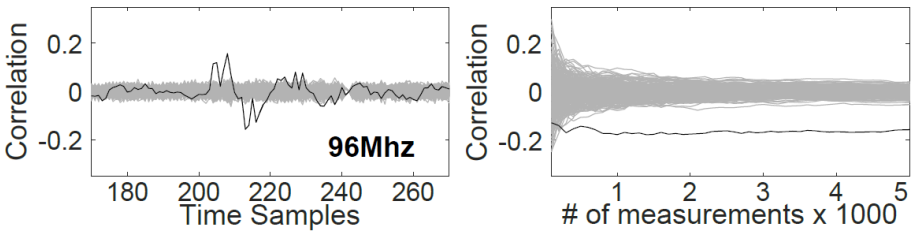


Abb. 8: Korrelation mit 5,000 Messungen einer AES Verschlüsselung (links) und der Verlauf der maximalen Korrelation über die Anzahl der Messungen (rechts). Die korrekte Key Hypothese ist in schwarz markiert. Gemessen mit einem TDC-basierten Spannungssensor in einem Spatan-6 FPGA. Der zugehörige Flurplan wird in Abbildung 6 gezeigt.

In Abbildung 8 zeigen wir eines unserer Ergebnisse einer Correlation Power Analysis (CPA) nach dem Prinzip von Brier et al. [BCO04]. Durch statistische Korrelation mehrerer Messungen mit einem Modell des erwarteten Stromverbrauchs lässt sich der geheime Schlüssel aus einem laufenden AES Modul extrahieren. Dazu bedarf es mehrerer Messungen von zufälligen Nachrichten mit demselben Schlüssel. In beiden Diagrammen zeigen die grauen Linien im Hintergrund die Korrelation mit dem Modell eines falschen Schlüssels, während die schwarze Linie der korrekte Schlüssel ist. Das linke Diagramm zeigt den zeitlichen Verlauf während einer AES Verschlüsselung, nachdem alle Messungen aufgenommen wurden, während das rechte Diagramme den Verlauf an einem bestimmten Zeitpunkt über die Anzahl der aufgenommenen Messungen zeigt.

6 Weitere Angriffe auf Platinenebene und in IoT Geräten

In den Experimenten der Dissertation wurden diese Angriffe bereits auf verschiedenen FPGA-Plattformen getestet. Eine Übersicht die wir in [Gn20b] veröffentlicht haben befindet sich in Tabelle 1. Weiterhin konnte gezeigt werden, dass ein Angriff über eine geteilte Stromversorgung auch auf einer Leiterplatte funktioniert. Hierfür wurden zwei FPGAs auf einer Platine verwendet wie in Abbildung 7 abgebildet. Einer davon implementiert AES oder RSA, während der andere den zuvor erklärten Sensor implementiert. Auch hier waren Angriffe zur Schlüssel-Extrahierung auf beide Algorithmen möglich, was wir in [Sc18b] veröffentlicht haben. Dies zeigt nun außerdem eine weitere Bedrohung: Sensoren die erst nachträglich durch Firmware-Updates in ein System eingeschleust werden.

Um die Allgemeingültigkeit von Spannungsschwankungen als Sicherheitslücke zu zeigen, zeigt die Arbeit einen weiteren Angriff, der über das PDN eines IoT-Geräts ausgeführt wird. Kostengünstige Geräte die in IoT-Anwendungen eingesetzt werden, integrieren oft analoge und digitale Komponenten auf einem einzigen Chip. Wir konnten zeigen, dass die Aktivität im digitalen Teil eines Chips sich auch auf den analogen Teil in Form von Rauschen auswirken kann. Wenn nun im nächsten Schritt etwas aus dem Analogteil digitalisiert wird, kann dies Informationen aus einem anderen Digitalteil beinhalten, auf den normalerweise nicht zugegriffen werden kann. Auch hier wurde dies an einem realen System am Beispiel einer Software AES Implementierung gezeigt, auf die ein CPA-Angriff erfolgreich ausgeführt wurde. Dies haben wir in [GKT19] veröffentlicht.

7 Zusammenfassung und Ausblick

SoCs mit integrierter FPGA-Logik sind bereits jetzt weit verbreitet. Durch den vermehrten Einsatz von FPGAs in der Cloud ändern sich auch deren Anforderungen an die Sicherheit. Das hohe Interesse am Betrieb mit mehreren Benutzern pro FPGA aus Industrie und Wissenschaft [By14, FVS15, Ha, Kh18] (*Multi-Tenant* Betrieb) verlangt, dass alle möglichen Sicherheitsrisiken untersucht werden. Diese Dissertation [Gn20a] zeigt, dass bisher unbeachtete Aspekte zu einem großen Risiko werden können, und vorige Lösungen zur rein logischen Isolation [Hu07] nicht ausreichend sind.

Diese Arbeit war die erste, die auf potenzielle Schwachstellen für Seitenkanal- und Fault-Angriffe hinwies, selbst in Szenarien, in denen eine logische Isolierung angewendet wird. Diese galten bisher als sicher gegen diese Angriffe – einfach aufgrund der Tatsache, dass ein potenzieller Angreifer keinen physischen Zugriff auf das Gerät hat, um etwa Messungen mit einem Oszilloskop durchzuführen. Die daraus resultierenden Sicherheitslücken sind keine Implementierungsfehler in der Art klassischer Software-Bugs, sondern ein allgemeines Problem, das sowohl von akademischem als auch industriellem Interesse ist. Diese Erkenntnis hat bereits jetzt die internationale Forschungsgemeinschaft dazu motiviert, mehr über vergleichbare Probleme zu erforschen, der Kürze halber können hier nur die wichtigeren Publikationen genannt werden [ZS18, Ra18, Ca18, Su19, HML19, Pr19, Al19, RGS20]. Wir und einige andere Gruppen haben bereits damit begonnen, sich auf weitere Analysen und die Entwicklung von Gegenmaßnahmen zu konzentrieren, von welchen bereits in [Gn18b, KGT19, Kr19, La20] veröffentlicht wurden. Wegen des hohen

Tab. 1: Übersicht experimentelle Ergebnisse zu Seitenkanal und Fault-Angriffen in FPGAs oder Systemen die FPGAs beinhalten, zum Stand der Veröffentlichung der Dissertation.

Board	**Erfolgreicher Angriff?**		
	Voltage Drop-basierter DoS	**Voltage Drop-basierter Fault-Angriff**	**Seitenkanal-Angriff**
Intel Terasic DE0-Nano-SoC	–	Ja, in Diss.	–
Intel Terasic DE1-SoC	Ja, in Diss.	Ja, in Diss.	–
Intel Terasic DE2-115	–	–	Ja, [Ra18][1]
Intel Terasic DE4	–	Ja, in Diss.	–
Lattice ECP5 5G Evaluation Board	–	–	Ja, in Diss.
Lattice iCE40-HX8K Breakout Board	Ja, in Diss.	Ja, in Diss.	Ja, in Diss.
Xilinx Artix-7 Basys-3	–	–	Ja, in Diss.
Xilinx Kintex-7 KC705	Ja, in Diss.	– [5]	–
Xilinx Pynq Zynq-ZC7020	Ja, in Diss.[2]	– [5]	Ja, in Diss.
Xilinx Spartan-6 SAKURA-G	–	–	Ja, in Diss.[4]
Xilinx Ultrascale VCU108	Ja, in Diss.	Ja, in Diss.	–
Xilinx Virtex-6 ML605	Ja, in Diss.	– [5]	–
Xilinx Virtex-7 VC707	–	Ja, [MS19]	–
Xilinx Zedboard Zynq-ZC7020	Ja, in Diss.[2]	– [5]	Ja, [ZS18][3]

[1] Informationslecks durch die Kopplung benachbarter Drähte im Chip, was allerdings ein weniger kritisches Problem ist, falls [Hu07] befolgt wird.

[2] Betrifft das gesamte SoC inklusive dem ARM Cortex-A9 Dual-Core.

[3] Hier wurde zusätzlich gezeigt, dass sich die integrierten ARM CPUs über den FPGA angreifen lassen.

[4] Angriff auch auf Ebene der Platine möglich.

[5] Mit mehr Aufwand vielleicht möglich.

Interesses erhielten die in der Dissertation enthaltenen Veröffentlichungen zahlreiche Nominierungen, Auszeichnungen, und werden weiterhin stark zitiert.

Neben dem stetig wachsenden akademischen Interesse ähnliche Schwachstellen zu finden oder Gegenmaßnahmen zu entwickeln, wirken sich die Ergebnisse dieser Arbeit auch auf die entsprechenden Cloud-Anbieter und FPGA-Hersteller aus. Um Sicherheit und Privatsphäre in zukünftigen Systemen zu garantieren, müssen die in dieser Dissertation gezeigten Aspekte der Stromversorgung berücksichtigt werden. Es kann davon ausgegangen werden, dass die hier gezeigten Bedrohungen in Zukunft verstärkt zu Problemen führen werden, falls sie weiterhin wenig Beachtung beim Systementwurf bekommen.

Zu guter letzt lassen sich die Ergebnisse der Dissertation auch für den Bildungsbereich nutzen. Da wir gezeigt haben, dass kostengünstige FPGA Boards alleine ausreichen können, um Fault- oder Seitenkanal-Angriffe durchzuführen, ist teure Test- und Messausrüstung in der Lehre nicht mehr unbedingt erforderlich. Dadurch kann jedem Studenten ein eigenes Board zur Verfügung gestellt werden, auf dem alle notwendigen Experimente sogar von zu Hause aus durchgeführt werden können. Zusammen mit einer begleitenden Vorlesung ist am KIT bereits ein Kurs *Praktische Einführung in die Hardwaresicherheit* etabliert worden, in dem dieser Vorteil genutzt wird. Besonders im Pandemie-Jahr 2020 war dies äußerst hilfreich um auch den praktischen Teil Online weiterführen zu können.

Literaturverzeichnis

[Al19] Alam, M; Tajik, S; Ganji, F; Tehranipoor, M; Forte, D: RAM-Jam: Remote Temperature and Voltage Fault Attack on FPGAs using Memory Collisions. FDTC, 9 2019.

[BCO04] Brier, E.; Clavier, C.; Olivier, F.: Correlation Power Analysis with a Leakage Model. CHES, 2004.

[BDL97] Boneh, D.; DeMillo, R. A.; Lipton, R. J.: On the Importance of Checking Cryptographic Protocols for Faults. In: Advances in Cryptology — EUROCRYPT. Springer Berlin Heidelberg, 1997.

[By14] Byma, S; Steffan, J G; Bannazadeh, H; Garcia, A L; Chow, P: FPGAs in the Cloud: Booting Virtualized Hardware Accelerators with Openstack. In: FCCM. IEEE, 2014.

[Ca18] Camurati, G.; Poeplau, S.; Muench, M.; Hayes, T.; Francillon, A.: Screaming Channels: When Electromagnetic Side Channels Meet Radio Transceivers. In: CCS. ACM, 2018.

[FVS15] Fahmy, S. A.; Vipin, K.; Shreejith, S.: Virtualized FPGA Accelerators for Efficient Cloud Computing. In: CloudCom. IEEE Computer Society, 2015.

[GKT19] Gnad, D. R. E.; Krautter, J.; Tahoori, M. B.: Leaky Noise: New Side-Channel Attack Vectors in Mixed-Signal IoT Devices. TCHES, 2019.

[Gn16] Gnad, D. R. E.; Oboril, F.; Kiamehr, S.; Tahoori, M. B.: Analysis of transient voltage fluctuations in FPGAs. In: FPT. IEEE, 2016.

[Gn18a] Gnad, D. R. E.; Oboril, F.; Kiamehr, S.; Tahoori, M. B.: An Experimental Evaluation and Analysis of Transient Voltage Fluctuations in FPGAs. TVLSI, 2018.

[Gn18b] Gnad, D. R. E.; Rapp, S.; Krautter, J.; Tahoori, M. B.: Checking for Electrical Level Security Threats in Bitstreams for Multi-tenant FPGAs. In: FPT. IEEE, 2018.

[Gn20a] Gnad, D. R. E.: Remote Attacks on FPGA Hardware. Dissertation, Karlsruher Institut für Technologie (KIT), 2020.

[Gn20b] Gnad, D. R. E.; Schellenberg, F.; Krautter, J.; Moradi, A.; Tahoori, M. B: Remote Electrical-level Security Threats to Multi-Tenant FPGAs. IEEE Design & Test, 2020.

[GOT17] Gnad, D. R. E.; Oboril, F.; Tahoori, M. B.: Voltage drop-based fault attacks on FPGAs using valid bitstreams. In: FPL. IEEE, 2017.

[Ha] Hao, W.: , FPGA Device Feature List (DFL) Device Drivers. https://lwn.net/Articles/757283/.

[HML19] Hunter, W.; McCarty, C.; Lerner, L.: Improved Techniques for Sensing Intra-Device Side Channel Leakage. In: FCCM. IEEE, 2019.

[Hu07] Huffmire, T.; Brotherton, B.; Wang, G.; Sherwood, T.; Kastner, R.; Levin, T. E.; Nguyen, T. D.; Irvine, C. E.: Moats and Drawbridges: An Isolation Primitive for Reconfigurable Hardware Based Systems. In: S&P. IEEE Computer Society, May 2007.

[KGT18] Krautter, J.; Gnad, D. R. E.; Tahoori, M. B.: FPGAhammer: Remote Voltage Fault Attacks on Shared FPGAs, suitable for DFA on AES. TCHES, 2018.

[KGT19] Krautter, J.; Gnad, D. R. E.; Tahoori, M. B.: Mitigating Electrical-level Attacks Towards Secure Multi-Tenant FPGAs in the Cloud. TRETS, August 2019.

[Kh18] Khawaja, A.; Landgraf, J.; Prakash, R.; Wei, M.; Schkufza, E.; Rossbach, C. J.: Sharing, Protection, and Compatibility for Reconfigurable Fabric with AmorphOS. In: USENIX OSDI. 2018.

[KJJ99] Kocher, P.; Jaffe, J.; Jun, B.: Differential Power Analysis. In: Advances in Cryptology — CRYPTO. Springer, 1999.

[Kr19] Krautter, J.; Gnad, D. R. E.; Schellenberg, F.; Moradi, A.; Tahoori, M. B.: Active Fences against Voltage-based Side Channels in Multi-Tenant FPGAs. In: ICCAD. ACM, 2019.

[La20] La, T. M.; Matas, K.; Grunchevski, N.; Pham, K. D.; Koch, D.: FPGADefender: Malicious Self-Oscillator Scanning for Xilinx UltraScale+ FPGAs. TRETS, 2020.

[MS19] Mahmoud, D.; Stojilovic, M.: Timing Violation Induced Faults in Multi-Tenant FPGAs. In: DATE. IEEE, 2019.

[Pr19] Provelengios, G.; Ramesh, C.; Patil, S. B.; Eguro, K.; Tessier, R.; Holcomb, D.: Characterization of Long Wire Data Leakage in Deep Submicron FPGAs. In: FPGA. ACM, 2019.

[Ra18] Ramesh, C.; Patil, S. B.; Dhanuskodi, S. N.; Provelengios, G.; Pillement, S.; Holcomb, D.; Tessier, R.: FPGA Side Channel Attacks without Physical Access. In: FCCM. IEEE Computer Society, 2018.

[RGS20] Rasmussen, K; Giechaskiel, I; Szefer, J: CAPSULe: Cross-FPGA covert-channel attacks through power supply unit leakage. In: S&P. IEEE, 2020.

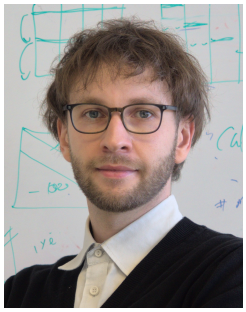
[Sc18a] Schellenberg, F.; Gnad, D. R. E.; Moradi, A.; Tahoori, M. B.: An inside job: Remote power analysis attacks on FPGAs. In: DATE. IEEE, 2018.

[Sc18b] Schellenberg, F.; Gnad, D. R. E.; Moradi, A.; Tahoori, M. B.: Remote inter-chip power analysis side-channel attacks at board-level. In: ICCAD. ACM, 2018.

[Su19] Sugawara, T.; Sakiyama, K.; Nashimoto, S.; Suzuki, D.; Nagatsuka, T.: Oscillator without a combinatorial loop and its threat to FPGA in data centre. Electronics Letters, 2019.

[Zi13] Zick, K. M.; Srivastav, M.; Zhang, W.; French, M.: Sensing Nanosecond-scale Voltage Attacks and Natural Transients in FPGAs. In: FPGA. ACM, 2013.

[ZS18] Zhao, M.; Suh, G. E.: FPGA-Based Remote Power Side-Channel Attacks. In: S&P. IEEE Computer Society, 2018.



Dennis Gnad wurde am 24. März 1988 in Pforzheim geboren. Er hat einen Bachelor in Technische Informatik (B.Eng.) von der Hochschule Pforzheim, und einen Master (M.Sc.) in Informatik vom Karlsruher Institut für Technologie (KIT). Dort promovierte er auch und schloss seinen Doktor in Informatik (Dr.-Ing.) mit Auszeichnung (Summa Cum Laude) ab. Aktuell widmet er seine Forschungsinteressen dem Gebiet der Hardware-Sicherheit. Hierbei liegt sein Fokus weiterhin auf Schwachstellen die durch eine gemeinsame Spannungsversorgung entstehen können, sowie verwandten Themen zur Sicherheit oder Verwendung von FPGAs.

Mehrebenen-Gestaltung von Dienstleistungssystemen[1]

Christian Grotherr[2]

Abstract: Die digitale Transformation verändert Geschäftsmodelle und erfordert ein Umdenken von Wertschöpfung. Wertschöpfung ist jedoch ein komplexes Phänomen, welches schwer zu beobachten und gezielt umzusetzen ist. Dieses ist für die Gestaltung von offenen Dienstleistungssystemen eine Herausforderung, die von der Mitwirkung der Akteure abhängt. Um diese Herausforderung zu adressieren wird ein Mehrebenen-Rahmenwerk entwickelt, welches die systematische Gestaltung von Dienstleistungssystemen unterstützt. Das Rahmenwerk zeigt die Zusammenhänge zwischen der Gestaltung von sozio-technischen Systemen und der Gestaltung der institutionellen Rahmenbedingungen auf. Als empirische Basis dient die Pilotierung eines IT-gestützten Crowdsourcing Systems. Durch die Evaluation des Systems wird das Mehrebenen-Rahmenwerk abgeleitet und um evidenzbasiertes Gestaltungswissen für die Einführung von internem Crowdsourcing erweitert. Die Forschungsergebnisse unterstützen in der systematischen Analyse, Gestaltung und Umsetzung von Dienstleistungssystemen, die durch sozio-technische Systeme ermöglicht werden und durch die Verzahnung der Gestaltung der institutionellen Rahmenbedingungen ihre Wirkung entfalten.

1 Motivation, Problemstellung und Zielsetzung

1.1 Digitalisierung als Treiber offener und vernetzter Dienstleistungssysteme

Die digitale Transformation ermöglicht neue Geschäftsmodelle durch Dienstleistungsinnovationen [Ba15]. Künstliche Intelligenz, soziale Plattformen und die Verfügbarkeit von Daten führen zu neuen Formen der Wertschöpfung, die zunehmend zu Öffnungseffekten und lernenden Dienstleistungssystemen führen [Sc17, BLM18]. Diese Veränderungen zeichnen sich durch (1) sozio-technische Systeme aus, die eine Wertschöpfung zwischen Akteuren ermöglichen, (2) durch Offenheit, die von der Mitwirkung der Akteure geprägt ist und (3) von den Werten, Normen und dem Verhalten der Akteure in dem Netzwerk beeinflusst werden. Schlüssel zur gemeinsamen Wertschöpfung ist somit die Mitwirkung der Akteure in einem interaktiven Prozess für einen gemeinsamen Nutzen [VL04, St16].

Die mit der digitalen Transformation einhergehende Dynamik erfordert ein Umdenken von Zusammenarbeit und Wertschöpfung und setzt sozio-technische Gestaltungsmethoden zunehmend unter Druck [Os15, Le17]. Aktuelle Methoden aus dem Service- und Softwareengineering adressieren die Dynamik durch iterative Vorgehensweisen, jedoch mit einem Fokus auf einzelne Dienstleistungen und die Gestaltung von interaktionsbezogenen Elementen zwischen einem Akteur und einem technischen System. Hierbei wird die Umgebung als Gestaltungselement nicht hinreichend betrachtet, obwohl die institutionellen Rahmenbedingungen, die Werte, Normen und Praktiken der Akteure als auch der Handlungsrahmen, in dem sich diese bewegen, stärker als bisher Wertschöpfungsinnovationen

[1] Englischer Titel der Dissertation: "Multilevel Design for Service Systems" [Gr20]

[2] Universität Hamburg, Arbeitsbereich IT-Management und -Consulting, Christian.Grotherr@uni-hamburg.de

beeinflussen [Ba15]. Zudem ist wenig darüber bekannt, wie systematisch eine gemeinsame Wertschöpfung in Dienstleistungssystemen realisiert werden kann [LN15, BLM14].

1.2 Neue Ansätze für die Gestaltung von sozio-technischen Systemen

Es sind neue Ansätze notwendig, die den Blick von der Gestaltung sozio-technischer Systeme auf die Gestaltung der Rahmenbedingungen in Dienstleistungssystemen erweitern und anstelle eines planorientierten Vorgehens eine Entscheidungsunterstützung zu den verschiedenen Gestaltungselementen bereithalten. Die Forderung nach einer methodischen Unterstützung und evidenz-basiertem Gestaltungswissen wird durch das Service Systems Engineering als Teildisziplin der Wirtschaftsinformatik getragen [BLM14]. Dafür ist jedoch eine Untersuchung der Prinzipien der Interaktion in Dienstleistungssystemen und deren systematische Gestaltung notwendig. Das Ziel der Forschungsarbeit ist es, durch ein Action-Design-Research Forschungsvorgehen im realen Umfeld einer Organisation einen methodischen Beitrag für die Gestaltung von Dienstleistungssystemen sowie evidenzbasiertem Gestaltungswissen zu generieren (s. Abbildung 1). Durch das Pilotierungsvorgehen wird eine Brücke zwischen Forschung und Praxis geschlagen, indem die Erkenntnisse aus der praktischen Anwendung mit den theoretischen Grundlagen verbunden werden.

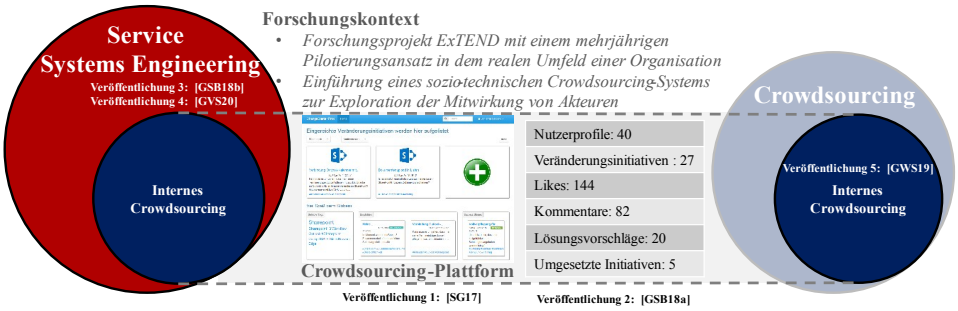


Abb. 1: Forschungskontext, angewendete Konzepte und Veröffentlichungen

Zur Exploration des Wertschöpfungsphänomens wurde das Forschungsprojekt ExTEND - Engineering von Dienstleistungssystemen durch nutzergenerierte Dienstleistungen initiiert. Der Schlüssel zur Wertschöpfung ist die Beteiligung der Akteure [St16]. Crowdsourcing stellt einen vielsprechenden, offenen Ansatz dar, um die Mitwirkung von Akteuren in Dienstleistungssystemen umzusetzen. Beide Disziplinen sind getrennt voneinander entstanden, haben aber den gemeinsamen Anspruch Akteure zu mobilisieren und zu einer gemeinsamen Wertschöpfung zu überführen. Internes Crowdsourcing bündelt Mechanismen, die Mitarbeitende dazu befähigt, ihre Ressourcen sowie ihre Fähigkeiten und ihr Wissen in einen interaktiven Wertschöpfungsprozess zu integrieren [Zu16]. Die kontinuierliche Beteiligung der Akteure stellt jedoch eine Herausforderung dar, die beim Ausbleiben der Aktivitäten zum Scheitern von Crowdsourcing-Projekten führt. Es ist somit wichtig zu verstehen, welche Faktoren die Nutzer im Organisationsumfeld zur Zusammenarbeit sowohl motivieren als auch hemmen. Über einen Zeitraum von drei Jahren wurden deshalb

Crowdsourcing-Mechanismen mithilfe einer IT-gestützten Plattform innerhalb einer Organisation mit 1.800 Mitarbeitern gestaltet und pilotiert. Aus den Beobachtungen, Interviews und Nutzungsdaten der Plattform wurden in einem iterativen Prozess Erkenntnisse für die Gestaltung und Einführung von durch sozio-technische Systeme gestützte Dienstleistungssysteme für eine gemeinsame Wertschöpfung gezogen.

2 Hintergrund

2.1 Dienstleistungssysteme als sozio-technische Systeme

Für die gemeinsame Wertschöpfung (engl. *value co-creation*) in Dienstleistungssystemen ist es von besonderer Bedeutung, das Engagement der Akteure sicherzustellen (engl. *actor engagenemt*) [St16]. Daraus ergibt sich die Notwendigkeit der Analyse von Wechselwirkungen zwischen den Akteuren, Technologien und dem Umfeld, um psychologische und verhaltensbezogene Aspekte in die Gestaltung effektiver Interaktionen voranzutreiben. Hierfür sind Gestaltungsansätze notwendig, die nicht nur die gemeinsame Wertschöpfung auf abstrakter Ebene, sondern die individuelle Interaktion zwischen den Akteuren adressieren. Um die Dynamik der Digitalisierung und die Wechselwirkungen zwischen Technologie, Menschen, Prozessen und Informationen besser zu verstehen, kann die Perspektive eines Dienstleistungssystem herangezogen werden. Dienstleistungssysteme als ”komplexe, soziotechnische Systeme, die die Wertschöpfung ermöglichen”[BLM14] helfen den Fokus von Technologie auf den Nutzungskontext und die Wirkung zu legen. Sie sind akteurszentriert, d.h. die Wertschöpfung setzt beim Akteur als Kunden und Nutzer an, welche durch Individualisierung und Kontextualisierung ermöglicht wird. Diese Anpassungen werden durch die Digitalisierung ermöglicht, die Einblicke in das Verhalten der Akteure erlaubt und zu einer Vernetzung der Akteure beiträgt. Darüber hinaus sind Dienstleistungssysteme offen, da sie basierend auf den gewonnenen Erkenntnissen fortlaufend weiterentwickelt werden und Betrieb und Entwicklung zu einer Einheit verschmelzen.

2.2 Crowdsourcing zur Mobilisierung und Integration von Akteuren

In den letzten Jahren ermöglicht die Digitalisierung neue Formen der Zusammenarbeit. Im Kern geht es um die Verschiebung der Wertschöpfung von einer güterdominierten Logik, in der ein Produkt produziert und verkauft wird, hin zu einer dienstleistungsdominierten Logik, in der der Mehrwert im Fokus steht und durch einen interaktiven Prozess erzeugt wird [VL04]. Vor diesem Hintergrund sind Ansätze wie Crowdsourcing entstanden, die klassische Dienstleistungsprozesse ergänzen und Geschäftsmodelle erweitern, aber auch neue Arbeitsorganisationen ermöglichen [Zu16]. Unternehmen haben die Möglichkeit das Wissen der Mitarbeitenden gestützt durch IT-Plattformen zeit- und ortsentkoppelt in die Wertschöpfungsprozesse einzubeziehen [LZD15]. Dies erleichtert die Integration von Ressourcen und führt zu einer Verlagerung von der Betrachtung einzelner Dienstleistungen zu offenen und lernenden Dienstleistungssystemen [BLM18].

3 Überblick der Publikationen und Forschungsergebnisse

Die Dissertation umfasst zwei Hauptergebnisse, die auf der Pilotierung einer IT-gestützten Crowdsourcing-Plattform aufsetzen: (1) ein Mehrebenen-Rahmenwerk für die systematische Gestaltung von Dienstleistungssystemen und (2) evidenz-basiertes Gestaltungswissen für die Einführung von internem Crowdsourcing. Die Ergebnisse bieten eine methodische Unterstützung bei der Gestaltung von Dienstleistungssystemen und eine Entscheidungshilfe für die Einführung von internem Crowdsourcing in Organisationen, welche in fünf Publikationen veröffentlicht wurden. Hieraus ergibt sich der kumulative Ansatz der Dissertation. Im Folgenden wird ein Einblick in die Pilotierung innerhalb der Organisation (3.1), die zentralen Beobachtungen (3.2), die Ableitung des Mehrebenen-Rahmenwerks (3.3) und die Gestaltungsprinzipien für internes Crowdsourcing (3.4) gegeben.

3.1 Pilotierung im realen Umfeld für evidenz-basiertes Gestaltungswissen

Mithilfe des Action-Design-Research Forschungsvorgehens [Se11] wurden im Projekt EXTEND Mechanismen zur Einbindung der Nutzerbasis in einen aktiven Verbesserungsprozess an dem konkreten Beispiel der Softwarenutzung exploriert. Hierfür wurde eine IT-gestützte Crowdsourcing-Plattform entwickelt und in eine Organisation eingeführt, die es den Mitarbeitenden ermöglicht, Verbesserungen für neu eingeführte Software vorzuschlagen, zu diskutieren und gemeinsam umzusetzen [SG17]. Die Plattform bildet die empirische Basis für die nachfolgenden Forschungsaktivitäten. Auf der Basis von Interviews, Workshops und Nutzungsdaten der Plattform wurden Evaluationsergebnisse gesammelt [GSB18a]. Die durch eine sozio-technische Perspektive gewonnen Erkenntnisse wurden mithilfe eines zweiten Pilotierungsfeldes angereichert und zu Gestaltungswissen für internes Crowdsourcing zusammengeführt [GWS19]. Diese Beobachtungen aus den Pilotierungen haben dabei geholfen, durch einen Generalisierungsschritt einen methodischen Beitrag für die Gestaltung solcher offener und von der Mitwirkung der Akteure abhängigen Dienstleistungssystemen zu leisten, welche in einem Mehrebenen-Rahmenwerk gebündelt werden [GSB18b]. Die Übertragbarkeit des Rahmenwerks wurde durch die Anwendung im Kontext eines Smart-Community-Forschungsprojekts zur Entwicklung eines digitalen Nachbarschaftsnetzwerkes demonstriert [GVS20].

3.2 Zusammenspiel zwischen Engagement von Akteuren, sozio-technischen Systemen und den institutionellen Rahmenbedingungen

Die Ergebnisse aus der Pilotierung zeigen, dass beteiligungsbefördernde Maßnahmen wie Community-Management nicht ausreichen, um eine kontinuierliche Mitwirkung der Akteure zu erzielen [SG17]. Nutzerzentrierte Methoden unterstützen die Gestaltung von sozio-technischen Systemen, um bspw. intuitive Benutzeroberflächen zu gestalten, berücksichtigen jedoch die institutionellen Rahmenbedingungen und die resultierende Wirkung auf die Akteure unzureichend. Die Rahmenbedingungen werden durch Werte und Normen repräsentiert, manifestieren sich durch das Verhalten der Akteure und sind geprägt durch

Regularien, Strukturen und Prozesse. Wenngleich Akteure motiviert zur Mitwirkung sind, können diese Rahmenbedingungen und das Werteversprechen des Dienstleistungssystems die Mitwirkung einschränken [GSB18a]. Am Beispiel des internen Crowdsourcings in der öffentlichen Verwaltung wird deutlich, dass die Vorgabe einer effizienten Ressourcennutzung als staatliche Institution, repräsentiert durch die Zielvorgaben der Funktionsbereiche, dem offenen und experimentierfreudigen Vorgehen beim internen Crowdsourcing entgegensteht. Deshalb müssen zusätzlich zum technischen System die Rahmenbedingungen, in dem sich die Akteure bewegen, einem aktiven Gestaltungsprozess unterzogen werden.

3.3 Mehrebenen-Rahmenwerk zur Gestaltung von Dienstleistungssystemen

Die Ergebnisse aus der Pilotierung wurden durch eine mikroökonomische Fundierung [St16] zu einem Mehrebenen-Rahmenwerk zusammengefasst [GSB18b] (s. Abbildung 2). Das Rahmenwerk hat zum Ziel, Gestaltungsaktivitäten zu systematisieren und deren Implikationen besser zu verstehen. Es zerlegt ein abstraktes Gestaltungsziel in kleinere, handhabbare und beobachtbare Gestaltungsaktivitäten durch folgende Eigenschaften:

1. Eine **Mehrebenen-Perspektive** mit Makro-Meso-Mikro-Ebenen, die das abstrakte Konzept der gemeinsamen Wertschöpfung durch gestaltbare sozio-technische Systeme mit beobachtbaren Interaktionen von individuellen Akteuren überbrückt.
2. **Iterative und validierende Gestaltungszyklen**, die miteinander verbunden sind. Hierdurch werden die Volatilität und die Unsicherheit in der Gestaltung von sozio-technischen Systemen und der umliegenden Rahmenbedingungen erfasst.

Die Betrachtung der Ebenen hilft die Gestaltungselemente zu benennen, deren Wirkungen im Dienstleistungssystem und den Wertschöpfungsbeitrag zu verstehen und Gestaltungsaktivitäten zuzuordnen. Die Auswirkungen von Gestaltungsentscheidungen auf jeder der Ebenen können analysiert werden und die Konfiguration des Dienstleistungssystems beeinflussen. Die Unterteilung in zwei verzahnte Gestaltungszyklen auf diesen Ebenen legt eine Zuordnung von Gestaltungselementen und deren Zusammenhänge wie folgt auf:

- **Institutionelles Design:** Der Gestaltungszyklus umfasst die Gestaltung der institutionellen Rahmenbedingungen. Das Ziel ist es, die Bedingungen zur Ressourcenmobilisierung und -integration auf institutioneller Ebene aufzubauen. Dieses umfasst Vereinbarungen und Verpflichtungen von Akteuren und Ressourcen, um die gemeinsame Wertschöpfung zu ermöglichen.
- **Engagement Design:** Dieser Gestaltungszyklus forciert die Entwicklung effizienter Interaktionsmuster zwischen Individuen. Das Ziel ist es, sozio-technische Systeme wie Crowdsourcing-Plattformen aufzubauen, die die Mitwirkung von unterschiedlichen Akteuren befördert.

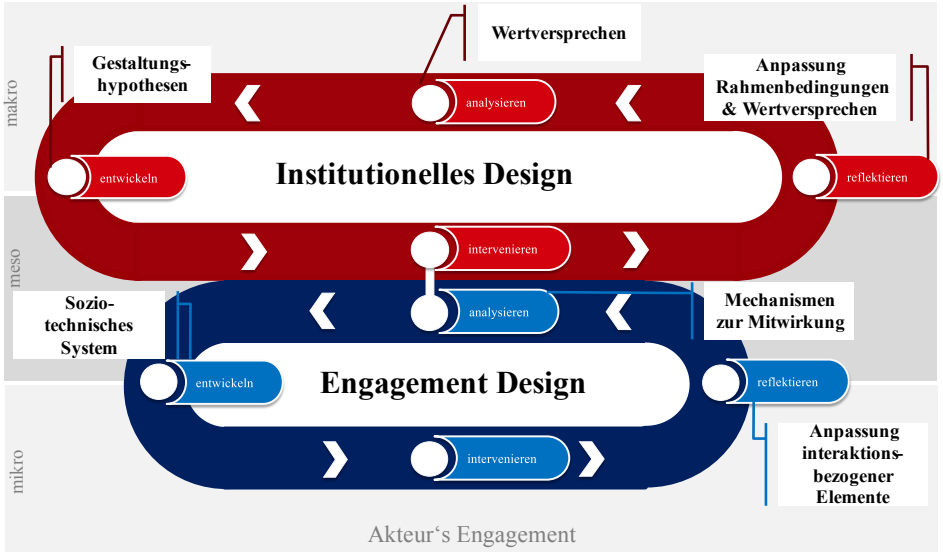


Abb. 2: Mehrebenen-Rahmenwerk für die Gestaltung von Dienstleistungssystemen

3.4 Gestaltungsprinzipien für die Einführung von internem Crowdsourcing

Während das Rahmenwerk einen methodischen Beitrag zur systematischen Gestaltung von Dienstleistungssystemen leistet, werden basierend auf den Erkenntnissen aus der naturalistischen Evaluation Gestaltungsprinzipien für internes Crowdsourcing abgeleitet. Diese Prinzipien fassen Empfehlungen zur Wahl von Gestaltungsoptionen und den korrespondierenden Effekten auf die Mitarbeitenden und Strukturen in Organisationen zusammen und leisten einen Beitrag zur Erweiterung des evidenzbasierten Gestaltungswissens. Die Gestaltungsprinzipien zur Einführung von internem Crowdsourcing als neue Form der Zusammenarbeit verdeutlichen die Gestaltung sowohl der technischen Plattform als auch des organisatorischen Aufbaus auf strategischer und operativer Ebene. Diese Aktivitäten reichen von der Veränderung von Managementpraktiken, Wertevorstellungen, Leistungsmesssystemen bis hin zu funktionsübergreifenden Kooperationsmechanismen [GWS19].

4 Theoretische Beiträge

Das Mehrebenen-Rahmenwerk zur Gestaltung von Dienstleistungssystemen leistet einen zentralen theoretischen Beitrag für die systematische Gestaltung von Dienstleistungssystemen (4.1). Die aus der Pilotierung gewonnen Gestaltungsprinzipien leisten einen Beitrag für die Gestaltung und Einführung von internen Crowdsourcing-Mechanismen (4.2).

4.1 Methodische Unterstützung zur Strukturierung von Gestaltungsaktivitäten

Das Rahmenwerk stellt durch den Mehrebenen-Ansatz einen hohen Detaillierungsgrad der Gestaltungsaktivitäten dar und hilft bei der systematischen Gestaltung von Dienstleistungssystemen, indem die Gestaltungselemente und -aktivitäten den Ebenen zugeordnet werden. Die Perspektive wird von der Gestaltung sozio-technischer Systeme mit interaktionsbezogenen Elementen auf die Gestaltung der Umgebung und der umliegenden institutionellen Rahmenbedingungen als untrennbare Einheit erweitert. Durch die Unterscheidung in *Institutionelles Design* und *Engagement Design* wird die Komplexität von Dienstleistungssystemen greifbarer und handhabbarer und zeigt gleichzeitig die Bedeutung der Verzahnung von Gestaltungszyklen. Diese Verzahnung hebt sich von bestehenden Gestaltungsansätzen ab, da die eindimensionale Perspektive auf Einzelfacetten wie der Interaktionsgestaltung, der Customer-Journey oder auch die Ermittlung eines Wertversprechens alleinig nicht den Eigenschaften von Dienstleistungssystemen gerecht wird. Hierdurch wird eine systematische Betrachtungsweise gebildet, die die Zusammenhänge zwischen individueller Beteiligung, technologischen Bausteinen und den Kontextbedingungen verdeutlicht. Hierfür wird unterschiedliches Gestaltungswissen benötigt. Diese Perspektive inkludiert psychologische, technologische und organisatorische Aspekte. Hierdurch können gezielt Domänenexperten in ein interdisziplinäres Gestaltungsteam eingebunden werden. Durch die Anwendung des Rahmenwerks auf die Domäne der Smart-Community konnten die Wirkungszusammenhänge von der Gestaltung einer digitalen Nachbarschaftsplattform als sozio-technisches System mit der Gestaltung der Rahmenbedingungen, bspw. die Einbindung von Unternehmen, Behörden und öffentlichen Institutionen, zur Skalierung des Netzwerkes und zur Verstetigung verdeutlicht werden [GVS20].

4.2 Evidenzbasiertes Gestaltungswissen für internes Crowdsourcing

Zentrale Treiber für die digitale Transformation sind neue Arbeitsformen und -prozesse. Durch die Pilotierung im realen Umfeld einer Organisation wird deutlich, dass der Erfolg digitaler Initiativen von den Umgebungsbedingungen und den Fähigkeiten der Mitarbeitenden abhängt. Sie zeigt, wie internes Crowdsourcing Mitarbeitende befähigt und bevollmächtigt, selbstständig Veränderungen voranzubringen. Das Gestaltungswissen hilft bei dem Verständnis von technischen als auch sozialen Gestaltungsoptionen. Konkret geht es um die Ausgestaltung der IT-gestützten Crowdsourcing-Plattform als sozio-technisches System, der Konstellationen von teilnehmenden Mitarbeitenden und deren Fähigkeiten, Management, Prozessstrukturen bis hin zu aufbauorganisatorischen Fragestellungen. Dieses Gestaltungswissen gibt Aufschlüsse darüber, wie auftretende organisatorische Hemmnisse überwunden werden, um die neuen Arbeitsmodelle zu verwirklichen [GWS19].

5 Praktische Beiträge

Im Folgenden wird die Rolle der Dienstleistungssysteme im Prozess der Transformation zu einer verstärkten Kunden- und Dienstleistungsorientierung dargestellt (5.1) und die Not-

wendigkeit hervorgehoben, sozio-technische Systemgestaltung unmittelbar mit der Gestaltung der institutionellen Rahmenbedingungen zu verzahnen (5.2), um einen spürbaren Nutzenbeitrag zu leisten. Das erfordert sense-and-response Fähigkeiten, um die Komplexität greifbar zu machen und die Wirkung mit einer interdisziplinären Perspektive beurteilen zu können. Durch die Forschungsergebnisse wird eine Unterstützung zur Verfügung gestellt, um diesen Prozess besser zu verstehen und beurteilen zu können.

5.1 Wechselspiel von Dienstleistungssystemen und der digitalen Transformation

Die Digitalisierung ermöglicht die Entwicklung neuartiger digitaler Geschäftsmodelle, welche durch die konsequente Wirkungs- und Nutzenorientierung von Dienstleistungssystemen im realen Umfeld skalieren können. Diese Veränderungen sind jedoch komplex. Sie beziehen sich sowohl auf sozio-technische Fragestellungen als auch auf neue Rollenverständnisse und Entscheidungskompetenzen von sowohl menschlichen als auch technischen Akteuren. Wenn das Konzept der Dienstleistungssysteme in den breiteren Kontext der Wertschöpfung gesetzt wird, wird sichtbar, dass dieser Erfolg von unterschiedlichsten Faktoren abhängt und die institutionellen Rahmenbedingungen ebenso wichtig sind wie die Kompetenzen der beteiligten Akteure oder die Gestaltung der Technologie. Das bedeutet, dass die Zusammenhänge zur vollen Wirkungsentfaltung nicht nur verstanden, sondern einem aktiven Gestaltungsprozess unterzogen werden müssen.

5.2 Erweiterung der sozio-technischen Systemgestaltung

Die Nutzenrealisierung von Dienstleistungssystemen erfordert neben den etablierten Gestaltungsmethoden für sozio-technische Systeme einen erweiterten Blick auf die Gestaltung der institutionellen Rahmenbedingungen. Das Mehrebenen-Rahmenwerk zeigt auf, wie die sozio-technische Systeme einerseits die Möglichkeiten der Interaktion und Vernetzung erweitern, andererseits die vorherrschenden, nutzerzentrierten Gestaltungsmethoden das Umfeld der Akteure nicht hinreichend berücksichtigen. Das Rahmenwerk bietet eine Unterstützung zur Koordination von Gestaltungsaktivitäten und -elementen. Es hilft verantwortliche Rollen aus unterschiedlichen Gestaltungsdomänen zuzuordnen, um geeignete Zusammenstellungen von sozio-technischen Systemen und institutionellen Rahmenbedingungen wie Governance-Strukturen oder Prozesse zusammenzustellen. Die ineinandergreifenden Gestaltungszyklen verdeutlichen die Notwendigkeit, die sozio-technische Systemgestaltung mit der Organisationsgestaltung aufeinander abzustimmen.

6 Zusammenfassung

Die digitale Transformation umfasst mehr als technologische Innovationen und organisatorische Restrukturierung. Sie greift mit neuen Geschäftsmodellen und Wertschöpfungsketten tief in die Grundlagen wirtschaftlicher und gesellschaftlicher Strukturen ein. Dieser

Wandel bezieht sich auch auf fortschreitende Öffnungseffekte, die sich in der Plattformökonomie oder neuen Formen der Zusammenarbeit wie Crowdsourcing widerspiegeln. Die einhergehende Dynamik in diesem Wandel setzt Methoden mit einem Fokus auf sozio-technische Systemgestaltung und die damit verbundene Interaktionsgestaltung zunehmend unter Druck. Die Forschungsergebnisse unterstützen in der systematischen Analyse, Entscheidung und Umsetzung von Gestaltungsoptionen bei der Entwicklung von Dienstleistungssystemen. Durch die interdisziplinäre Sichtweise auf sozio-technische Systeme und den Kontext der Akteure wird die Perspektive von technologisch fokussierten Entwicklungen um die Gestaltung von institutionellen Rahmenbedingungen erweitert. Die Gestaltung von sozio-technischen Systemen und von institutionellen Rahmenbedingungen als nicht trennbare Einheit verändert die Wertschöpfung zwischen Akteuren, dessen Wirkung auf die Organisationen und Märkte. Hierfür sind weiterführende Forschungsaktivitäten notwendig, die durch die Anwendung des Mehrebenen-Rahmenwerks in unterschiedlichen Kontexten das Wissen über das Zusammenwirken der Gestaltungszyklen erweitern.

Literaturverzeichnis

[Ba15] Barrett, Michael; Davidson, Elizabeth; Prabhu, Jaideep; Vargo, Stephen L: Service innovation in the digital age: key contributions and future directions. MIS Quarterly, 39(1):135–154, 2015.

[BLM14] Böhmann, Tilo; Leimeister, Jan Marco; Möslein, Kathrin: Service Systems Engineering. Business and Information Systems Engineering, 6(2):73–79, 2014.

[BLM18] Böhmann, Tilo; Leimeister, Jan Marco; Möslein, Kathrin: The New Frontiers of Service Systems Engineering. Business and Information Systems Engineering, 60(5):373–375, 2018.

[Gr20] Grotherrr, Christian: Multilevel Design for Service Systems. Thesis, 2020.

[GSB18a] Grotherr, Christian; Semmann, Martin; Böhmann, Tilo: Engaging Users to Co-Create - Implications for Service Systems Design by Evaluating an Engagement Platform. In: 51st Hawaii International Conference on System Sciences (HICSS). Waikoloa Village, Hawaii, USA, 2018.

[GSB18b] Grotherr, Christian; Semmann, Martin; Böhmann, Tilo: Using Microfoundations of Value Co-Creation to Guide Service Systems Design - A Multilevel Design Framework. In: International Conference on Information Systems (ICIS). San Francisco, California, USA, 2018.

[GVS20] Grotherr, Christian; Vogel, Pascal; Semmann, Martin: Multilevel Design for Smart Communities - The Case of Building a Local Online Neighborhood Social Community. In: 53rd Hawaii International Conference on System Sciences (HICSS). Grand Wailea, Maui, USA, 2020.

[GWS19] Grotherr, Christian; Wagenknecht, Thomas; Semmann, Martin: Waking Up a Sleeping Giant: Lessons from Two Extended Pilots to Transform Public Organizations by Internal Crowdsourcing. In: International Conference on Information Systems (ICIS). Munich, Germany, 2019.

[Le17] Legner, Christine; Eymann, Torsten; Hess, Thomas; Matt, Christian; Böhmann, Tilo; Drews, Paul; Mädche, Alexander; Urbach, Nils; Ahlemann, Frederik: Digitalization: opportunity and challenge for the business and information systems engineering community. Business and Information systems Engineering, 59(4):301–308, 2017.

[LN15] Lusch, Robert F; Nambisan, Satish: Service Innovation: A Service-Dominant Logic Perspective. MIS Quarterly, 39(1):155–175, 2015.

[LZD15] Leimeister, Jan Marco; Zogaj, Shkodran; Durward, David: New Forms of Employment and IT–Crowdsourcing. In (Blanpain, R.; Hendrickx, F.; Waas, B., Hrsg.): New Forms of Employment in Europe. Wolters Kluwer, S. 23–41, 2015.

[Os15] Ostrom, Amy L; Parasuraman, Ananthanarayanan; Bowen, David E; Patricio, Lia; Voss, Christopher A; Lemon, Katherine: Service research priorities in a rapidly changing context. Journal of Service Research, 18(2):127–159, 2015.

[Sc17] Schlagwein, Daniel; Conboy, Kieran; Feller, Joseph; Leimeister, Jan Marco; Morgan, Lorraine: Öpenness"with and without Information Technology: a framework and a brief history. Journal of Information Technology, 32(4):297–305, 2017.

[Se11] Sein, Maung; Henfridsson, Ola; Purao, Sandeep; Rossi, Matti; Lindgren, Rikard: Action design research. MIS Quarterly, 35(1):37–56, 2011.

[SG17] Semmann, Martin; Grotherr, Christian: How to Empower Users for Co-Creation - Conceptualizing an Engagement Platform for Benefits Realization. In: Wirtschaftsinformatik. St. Gallen, Switzerland, 2017.

[St16] Storbacka, Kaj; Brodie, Roderick J; Böhmann, Tilo; Maglio, Paul P; Nenonen, Suvi: Actor engagement as a microfoundation for value co-creation. Journal of Business Research, 69(8):3008–3017, 2016.

[VL04] Vargo, Stephen L; Lusch, Robert F: Evolving to a new dominant logic for marketing. Journal of Marketing, 68(1):1–17, 2004.

[Zu16] Zuchowski, Oliver; Posegga, Oliver; Schlagwein, Daniel; Fischbach, Kai: Internal crowdsourcing: Conceptual framework, structured review, and research agenda. Journal of Information Technology, 31(2):166–184, 2016.

Christian Grotherr ist seit 2015 wissenschaftlicher Mitarbeiter am Arbeitsbereich IT-Management und -Consulting. Sein Forschungsschwerpunkt liegt in der Gestaltung von Dienstleistungssystemen, die sich durch Offenheit und einen lernenden Charakter auszeichnen. In unterschiedlichen Anwendungsdomänen erforscht Christian Grotherr, wie digitale Dienstleistungssysteme, die sich durch ein hohes Maß an Mitwirkung von Akteuren auszeichnen, gestaltet und pilotiert werden und wie eine Skalierung erzielt werden kann. Von der Projektbeantragung bis zur -durchführung begleitet Christian Grotherr unterschiedliche Schwerpunkte: Von der Entwicklung und Pilotierung von Mitarbeiterbeteiligungsmodellen (ExTEND), Konzeption skalierbarer hybrider KI-Szenarien (INSTANT), bis hin zur Entwicklung strategischer Positionen, die Handlungsbedarfe für die künftige Dienstleistungsforschung aufzeigen (DL2030). Zuvor studierte Christian Grotherr Wirtschaftsinformatik (B.Sc.) und IT-Management und -Consulting (M.Sc.) an der Universität Hamburg.

Auf dem Weg zu robusten und interpretierbaren praktischen Anwendungen der automatischen Analyse mentaler Zustände unter Verwendung eines dynamischen und hybriden Ansatzes zur Schätzung von Gesichtsbewegungen [1]

Teena Chakkalayil Hassan[2]

Abstract: Dieses Dokument präsentiert eine Zusammenfassung der Dissertation der Autorin. In dieser Dissertation [Ha20] wurde ein neuartiger und hybrider Ansatz für die Schätzung der Intensität von Gesichtsmuskelbewegungen (Action Unit (AU)) vorgeschlagen und validiert. Dieser Ansatz basiert auf einer Gauß'schen Zustandsschätzung und kombiniert ein verformbares, AU-basiertes Gesichtsformmodell, ein viskoelastisches Modell der Gesichtsmuskelbewegung, mehrere erscheinungsbasierten AU-Klassifikatoren und eine Methode zur Erkennung von Gesichtspunkten. Es wurden mehrere Erweiterungen vorgeschlagen und in das Zustandsschätzungs-Framework integriert, um mit den personenspezifischen Eigenschaften sowie technischen und praktischen Herausforderungen umzugehen.Die mit der vorgeschlagenen Methode erzeugten AU-Intensitätsschätzungen wurden für die automatische Erkennung von Schmerzen und für die Analyse von Fahrerablenkung eingesetzt.

1 Einführung

Mentale Zustände können anhand von verschiedenen Verhaltens- und physiologischen Signale analysiert werden. Zu den Verhaltenssignalen gehören z. B. Gesichtsausdrücke, Vokalisationen und Körperbewegungen. Unter diesen hat die Mimik in psychologischen Studien eine herausragende Stellung eingenommen, vor allem aufgrund ihrer Bedeutung in der nonverbalen Kommunikation [HWM78, Pa05], die die Kommunikation von Informationen über mentale Zustände wie Emotionen und Schmerz umfasst [EF71, KL14]. In der Folge hat die automatische Analyse von Gesichtsausdrücken anhand von Bildern und Videos viel Aufmerksamkeit in der Computer-Vision-Forschung erhalten [SGC15, Ma19]. Einerseits wurden Ansätze entwickelt, um automatisch objektive Beschreibungen von Gesichtsausdrücken in Bezug auf die grundlegenden Gesichtsmuskelbewegungen zu generieren, die als Action Units (AUs) bekannt sind und im Facial Action Coding System (FACS) [EF78] definiert sind (*sign judgment*). Andererseits wurden Ansätze entwickelt, um den mentalen Zustand oder die "Botschaft"vorherzusagen, die durch den Gesichtsausdruck kommuniziert wird (*message judgment*). Das häufigste Ziel dabei ist entweder die Erkennung der von Paul Ekman und Kollegen identifizierten Basisemotionen, wie Angst, Wut,

[1] Englischer Titel der Dissertation: "Towards Robust and Interpretable Practical Applications of Automatic Mental State Analysis Using a Dynamic and Hybrid Facial Action Estimation Approach"

[2] Kognitive Systeme, Fakultät Wirtschaftsinformatik und Angewandte Informatik, Otto-Friedrich Universität Bamberg, teena.ch@gmail.com

Traurigkeit, Ekel, Freude, Verachtung und Überraschung [EC11, EF71] (kategorisches Modell) (z.B. [DZC17]) oder die Einschätzung der Valenz- und Arousal-Dimensionen von Emotionen [Ru80] (dimensionales Modell) (z.B. [Ga13]). Ansätze zur *message judgment* lernen den mentalen oder emotionalen Zielzustand entweder direkt aus visuellen Daten (*einstufige Ansätze*) (z.B. [KPM16]) oder aus den objektiven Beschreibungen, die von den Ansätzen zur *sign judgment* erzeugt werden (*zweistufige Ansätze)* (z. B. [Ba14]).

Für die automatische AU-Intensitätsschätzung haben Computer-Vision- Forscher sowohl datengetriebene Machine-Learning-Ansätze (z. B. [SSB12, KRP12]) als auch auf verformbaren Gesichtsmodellen basierende Ansätze (z. B. [Do11, DD08]) verwendet. Datengetriebene maschinelle Lernmethoden haben den Vorteil, dass sie allgemeine Modelle lernen können, die eine große Varianz in den Trainingsdaten abdecken, was zu einer sehr guten Vorhersageleistung führt. Im Gegensatz dazu könnte die Leistung von Ansätzen, die auf verformbaren Gesichtsmodellen basieren, durch die Genauigkeit und Korrektheit der Modelle und die damit verbundenen Vorannahmen begrenzt sein. Die Stärke der auf verformbaren Gesichtsmodellen basierenden Ansätze liegt jedoch in der Interpretierbarkeit der Modelle und ihrer Parameter sowie in der Möglichkeit, interdisziplinäres und menschliches Expertenwissen zu integrieren. Die Kombination der beiden Ansätze könnte helfen, sich in Richtung starker oder ultrastarker Systeme der künstlichen Intelligenz (KI) [Mi88] für die Analyse von Gesichtshandlungen und mentalen Zuständen zu bewegen, die eine gute Vorhersageleistung haben und die Nachvollziehbarkeit von Entscheidungen fördern.

Im Großen und Ganzen leistet diese Dissertation [Ha20] die folgenden Beiträge:

- Es wird eine neuartige Kombination aus datengetriebenem maschinellem Lernen und auf verformbaren Gesichtsmodellen basierenden Ansätzen entwickelt, um AU-Intensitäten zu schätzen. Ein auf Gaußscher Zustandsschätzung basierender Framework wird dafür eingesetzt. Um die Qualität der Schätzungen zu verbessern und die Robustheit in realen Anwendungen zu erhöhen, wurden mehrere Erweiterungen dieses Frameworks entworfen und untersucht. Diese Erweiterungen berücksichtigen personenspezifische, gesichtsmuskelspezifische und verformbare Gesichtsmodell-spezifische Eigenschaften und behandeln Fälle von fehlenden oder anomalen Informationen.
- Die AU-Intensitäten, die durch das oben erwähnte Framework geschätzt werden, werden dann für die automatische Analyse des mentalen Zustands verwendet: (i) zur Erkennung von Schmerzen auf der Grundlage von AU-basierten Regeln und (ii) zur Untersuchung der Unterschiede zwischen Gesichtsausdrücken unter verschiedenen Arten von Ablenkungen während des simulierten Autofahrens.
- Mehrere Herausforderungen, die angegangen werden müssen, um praktisch nützliche automatische Systeme zur Analyse des mentalen Zustands zu bauen, wurden untersucht. Es wird über erste Arbeiten berichtet, die das Sammeln von Anforderungen an Referenzdatensätze für die Analyse des mentalen Zustands, die Modellierung oder Überprüfung von zwischenmenschlichen Unterschieden in den Reaktionen auf schmerzhafte oder erregende Reize und die Generierung verschiedener Erklärungen für die automatische Erkennung von Schmerz beinhalten.

2 Schätzung der Intensitäten von Gesichtsmuskelbewegungen

In dieser Doktorarbeit [Ha20] wird eine auf Gauß'scher Zustandsschätzung basierende Methode zur Schätzung von kontinuierlichen Intensitäten von 22 AUs entwickelt. Diese Methode kombiniert (i) ein AU-basiertes verformbares Gesichtsmodell, (ii) ein viskoelastisches Gesichtsmuskelbewegungsmodell, (iii) mehrere datengetriebene AU-Klassifikationsmodelle, die auf Erscheinungsbildinformationen basieren, und (iv) ein datengetriebenes Modell zur Erkennung von Gesichtsmerkmalspunkten, das Forminformation liefert (siehe Abbildung 1). Die Parameter des 3D, verformbaren Gesichtsformmodells, insbesondere Kopfposenparameter und AU-Parameter, stellen die wichtigsten Zustandsvariablen dar. Abhängig von den verwendeten Prozessmodellen werden auch zusätzliche Parameter oder Variablen in den Zustandsvektor aufgenommen. Die Erkennung von Gesichtsmerkmalspunkten und die auf dem Aussehen basierende AU-Klassifizierung dienten als Quelle für Form- und Erscheinungsbildinformationen. Somit handelt es sich um einen hybriden Ansatz.

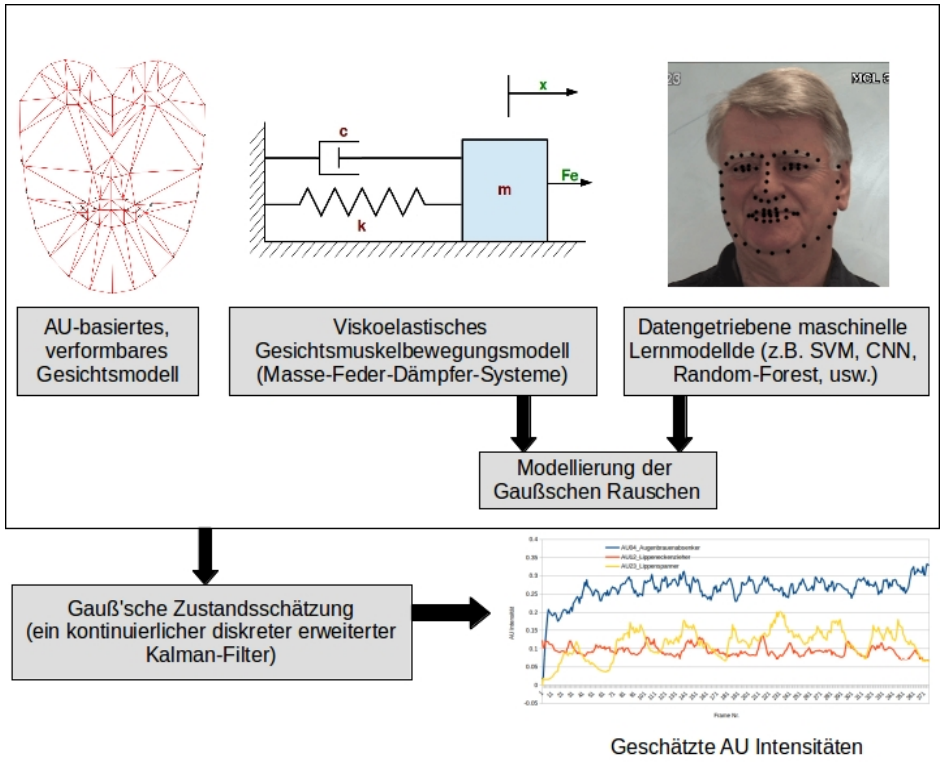


Abb. 1: Die in dieser Dissertation entwickelte neuartige Methode zur Schätzung von kontinuierlichen AU-Intensitäten. Es kombiniert ein verformbares Gesichtsmodell, viskoelastisches Gesichtsmuskelbewegungsmodell und datengetriebene maschinelle Lernmodelle innerhalb eines Gauß'schen Zustandsschätzungsframeworks, um Intensitäten von 22 AUs zu schätzen. Das Gesichtsbild stammt aus [Lu11] und ist ©Jeffrey Cohn.

Um die AU-Dynamik mit einem angetriebenen Masse-Feder-Dämpfer-System zu modellieren, wird angenommen, dass die Dehnungsrichtung der Feder identisch mit der Richtung ist, die durch den entsprechenden AU-Verformungsvektor im Modell der verformbaren Gesichtsform definiert ist. Die äußere Kraft $\vec{F}_e$ wird so modelliert, dass sie in dieser Richtung wirkt, um eine positive Auslenkung des Masse-Feder-Dämpfer-Systems zu erzeugen. Die Größe der Auslenkung $\vec{x}$ stellt somit die Intensität der entsprechenden AU dar. Um zu verhindern, dass das Masse-Feder-Dämpfer-System bei Wegnahme der äußeren Kraft über die Gleichgewichtslage hinausschießt, kann eine kritisch gedämpfte oder überdämpfte Konfiguration verwendet werden. Dies, zusammen mit einer Beschränkung von $\vec{F}_e$ auf nicht-negative Werte – wie wir in [Ha18] vorgeschlagen haben – stellt sicher, dass die AU-Intensitätsabschätzungen, die durch dieses dynamische Modell gegeben werden, nicht-negativ sind, und somit mit FACS konsistent sind.

Einige Forscher haben bereits Wahrscheinlichkeitsausgaben von datengesteuerten maschinellen Lernmethoden in diskrete Zustandsschätzungs-Frameworks integriert. Krüger et al. [Kr05] kombinierten beispielsweise Support Vector Machine (SVM)-Ausgaben innerhalb eines Hidden Markov Model (HMM) zur Spracherkennung; Valstar und Pantic [VP12] kombinierten SVM und HMM zur Erkennung der zeitlichen Phasen von AUs; Tong et al. [TLJ07] integrierte die Ausgaben von mehreren AdaBoost-Klassifikatoren als verrauschte Messungen (oder, Beobachtungen) innerhalb eines Dynamischen Bayesschen Netzes (DBN) zur Erkennung von 14 AUs; Li et al. [Li15] integrierten die Ausgaben einer Menge von SVMs als Beobachtungen in einen DBN zur Schätzung von sechs diskreten AU-Intensitätsstufen für 12 AUs. Im Gegensatz zu diesen Methoden wird in dieser Dissertation [Ha20] die Fusion der klassenweisen Wahrscheinlichkeitsausgaben von Methoden des maschinellen Lernens, wie z. B. einer SVM, innerhalb eines kontinuierlichen, gaußschen Zustandsschätzungsframework untersucht. Dazu sollen die Wahrscheinlichkeiten zunächst in Gaußsche Rauschabweichungen umgewandelt werden. Die Wahrscheinlichkeit p des Vorhandenseins einer AU A definiert eine Bernoulli-Verteilung, wobei die boolesche Zufallsvariable A den Wert 1 (Vorhandensein von AU) mit der Wahrscheinlichkeit p und den Wert 0 (Abwesenheit von AU) mit der Wahrscheinlichkeit $1 - p$ annimmt. Das zweite Moment dieser Verteilung wird als Varianz des Gaußschen Rauschens zur Integration in die Zustandsschätzungsmethode verwendet. Die Varianz ist am höchsten, wenn die Wahrscheinlichkeit $0,5$ beträgt. Zwei-Klassen-Klassifikatoren sagen die Wahrscheinlichkeit des Vorhandenseins oder Nichtvorhandenseins einer einzelnen AU voraus, und daher können diese Wahrscheinlichkeiten direkt in Bernoulli-Varianz umgewandelt werden. Im Falle von Mehrklassen-Klassifikatoren, die verschiedene boolesche Kombinationen von AUs erkennen, wird die Marginalisierungstechnik angewendet, um die Wahrscheinlichkeiten für einzelne AUs zu erhalten, die dann in Bernoulli-Varianz umgewandelt werden können. Dieser Fusionsansatz ist in [Ha16] veröffentlicht. Bei drei AUs im oberen Bereich des Gesichts, nämlich AU01, AU04 und AU06, führte der vorgeschlagene Fusionsansatz zu einer Erhöhung der Performance (Area Under receiver operating characteristic Curves (AUC)) um 5 bis 10 % im Vergleich zu den auf Erscheinungsmerkmalen wie Histogram of Oriented Gradients (HOG) oder Local Binary Patterns (LBP) trainierten Support Vector Machine (SVM) Klassifikatoren. Darüber hinaus wurde eine Steigerung von 6 bis 14 % erzielt, wenn nur die Positionen der Gesichtsmerkmale als Messung verwendet wurden.

Die vorgeschlagene Methode zur Schätzung der AU-Intensitäten wurde quantitativ und qualitativ unter Verwendung von drei verschiedenen Datensätzen bewertet, die entweder spontane oder gespielte Gesichtsausdrücke mit AU-Annotationen enthalten. Die vorgeschlagene Methode führte zu zeitlich glatteren Schätzungen, die eine feinkörnige Analyse der Gesichtsausdrücke ermöglichen. Die geschätzten AU-Intensitäten tendieren dazu, in einem relativ niedrigen Wertebereich zu bleiben, und weisen oft einen etwas verzögerten Beginn auf. Dies zeigt, dass die vorgeschlagene Methode konservativ ist. Die vorgeschlagene Methode hat eine recht gute Leistung erbracht, obwohl sie gleichzeitig Intensitäten von 22 AUs schätzt, von denen einige im Ausdruck subtil oder einander sehr ähnlich sind. Die quantitativen Ergebnisse wurden auf dem UNBC-McMaster Shoulder Pain Expression Archive Database [Lu11] generiert, das 48398 Frames aus 200 Videos von 25 Probanden enthält. Die Leistung wurde mit einem Random-Forest basierten Ansatz [DBD18] verglichen. Der Fehler in den geschätzten AU-Intensitäten ist über alle verfügbaren Frames geringer im Vergleich zu [DBD18], aber bei nur annotierten Frames ist der Fehler höher. Dies bestätigt dass die Methode konservativ ist. Um einen fairen Vergleich zu ermöglichen ist es wichtig, dass die AU intensitäten mit individualisierten Schwellwerten diskretisiert werden. Um die Leistung weiter zu verbessern, könnten die modernsten Ansätze des maschinellen Lernens zur AU-Erkennung in den vorgeschlagenen probabilistischen Framework zur Schätzung der AU-Intensitäten integriert werden.

2.1 Erweiterung des Basis Frameworks

Das Framework bietet viele Möglichkeiten fuer Verbesserung durch die Integration von interdisziplinärem und domän-spezifischem Wissen.

Um die Robustheit der AU-Intensitätsschätzung zu verbessern, wurde eine Anomalieerkennung eingeführt, um anomale Gesichtsmerkmale zu detektieren. Unbemerkten oder plötzliche Variationen der Gesichtsform und -textur, die durch Beleuchtungsänderungen, Gesichts- und Kopfbewegungen oder Gesichtsausdrücke hervorgerufen werden, können zu anomalen Erkennungen von Gesichtsmerkmalen führen, die sich deutlich von den vorhergesagten Positionen der Gesichtsmerkmale unterscheiden. Außerdem können Beleuchtungsänderungen oder Körperbewegungen die Textur der Kleidung oder des Hintergrunds beeinflussen. Dies kann zu einer plötzlichen Verschiebung der erkannten Gesichtsmerkmalen in einen völlig anderen Bereich des Bildes führen. Die Anomalien werden mit der Methode *normalisierte Innovationsquadrat* [Ni08] erkannt, die die Divergenz zwischen den tatsächlichen Messungen und den auf Basis der Apriori-Zustandsschätzungen vorhergesagten Messungen misst. Normalisierte Innovationsquadratwerte folgen der Chi-Quadrat-Verteilung. Durch geeignete Schwellenwerte können die Empfindlichkeit gegenüber Anomalien angepasst werden. Die so erkannten anomalen Gesichtsmerkmalen wurden verworfen und die Apriori-AU-Intensitätsschätzungen wurden in dem Fall nicht aktualisiert.

Abhängig von den Eigenschaften des/der Gesichtsmuskels/-muskeln könnten die Parameter des angetriebenen Masse-Feder-Dämpfer-Modells für jede AU unterschiedlich initialisiert werden. Hierfür wird in dieser Dissertation [Ha20] eine einfache Methode vorgeschlagen, die auf der Zusammensetzung der Gesichtsmuskelfasern basiert. Phasische Mus-

keln enthalten einen geringeren Anteil an Muskelfasern vom Typ I und reagieren schnell, während tonische Muskeln einen hohen Anteil an Muskelfasern vom Typ I enthalten und langsam auf Reize reagieren [Fr90]. Masse-Feder-Dämpfer-Systeme mit steiferen Federn oder stärkeren Dämpfern reagieren langsamer und milder auf die Antriebskraft, während weniger steife Federn und schwächere Dämpfer schneller und stärker auf die Antriebskraft reagieren. Basierend auf diesen Informationen wurden unterschiedliche Ausgangswerte für die Eigenfrequenz und das Dämpfungsverhältnis der Masse-Feder-Dämpfer-Systeme für verschiedene AUs konfiguriert.

3 Schmerzerkennung

Im Rahmen dieser Dissertation [Ha20] wurde eine umfassende Literaturrecherche durchgeführt [Ha19], die die Schmerzerkennungsansätze auf Basis der verwendeten Lernziele, Lernmethoden und Merkmale kategorisierte. Die Ansätze wurden auch grob in einstufige oder zweistufige Ansätze eingeteilt, je nachdem, ob schmerzbezogene Ziele durch direkte Analyse des visuellen Inputs oder auf Basis der erkannten AUs gelernt wurden. Es wurde festgestellt, dass es sich bei den Ansätzen zur automatischen Schmerzerkennung überwiegend um datengetriebene, überwachte Lernansätze handelt. Modelle, die auf Experten- und interdisziplinärem Wissen über Schmerzen und Gesichtsausdrücke basierten, beschränkten sich auf die Verwendung der Prkachin-Solomon-Pain-Intensity (PSPI)-Skala [PS08] zur AU-basierten Schmerzintensitätsschätzung in zweistufigen Ansätzen (z. B. [ZK14]). Von Experten entworfene Rechenmodelle würden es jedoch nicht nur ermöglichen, interdisziplinäres Wissen über die Eigenschaften und die Dynamik der Gesichtsmuskulatur, mögliche Verformungen der Gesichtsform und die Gesichtsausdrücke des Schmerzes mit einzubeziehen, sondern auch das menschliche Verständnis zu erleichtern.

In dieser Dissertation [Ha20] wurden die AU-Intensitätsschätzungen aus dem oben beschriebenen Framework zur Vorhersage von Schmerzen verwendet, indem einfache Regeln angewendet wurden, die auf psychologischen Erkenntnissen aus Studien von Kunz und Lautenbacher [KL14] sowie Prkachin und Solomon [PS08] basieren. Während die Arbeit von Prkachin und Solomon [PS08] einige zweistufige Ansätze inspiriert hat [ZK14, MYL17], wurden die von Kunz und Lautenbacher [KL14] identifizierten Schmerzcluster noch nicht für die automatische AU-basierte Schmerzerkennung angewendet. Basierend auf diesen AU-basierten Regeln wurden verbale Erklärungen in FACS-Terminologie erstellt, um das Potenzial des Ansatzes für den Aufbau transparenter Schmerzerkennungssysteme zu verdeutlichen. Bei der empirischen Auswertung auf dem UNBC-McMaster Shoulder Pain Expression Archive Database [Lu11] schnitten die PSPI-basierten Regeln am besten ab, da sie den Schmerzannotationen am meisten ähnelten.

4 Erkennung von Ablenkung

Es wurde eine vorläufige Analyse der Gesichtsaktivität unter dem Einfluss verschiedener Ablenkungsquellen während simulierter Fahrsitzungen durchgeführt. Dazu wurde der

Datensatz von Taamneh et al. [Ta17] verwendet, der Daten enthält, die von 68 menschlichen Probanden aufgezeichnet wurden, während sie auf vordefinierten Strecken in einem Fahrsimulator unter verschiedenen Bedingungen fuhren. Kognitive, emotionale und sensomotorische Stressoren wurden verwendet, um Ablenkung während des Fahrens zu induzieren. Die Gesichtsaktivität wurde anhand der AU-Intensitäten (und deren Zeitableitungen) bestimmt, die mit dem vorgeschlagenen, auf Gaußscher Zustandsschätzung basierenden Ansatz geschätzt wurden. Zwischen verschiedenen Ablenkungsbedingungen wurden Unterschiede in der Gesichtsaktivität gefunden. Im Großen und Ganzen ist die durchschnittliche Gesichtsaktivität während der verschiedenen Fahrsituationen nicht sehr hoch. Allerdings zeigten die während der Fahrten mit emotionalen Stressoren aufgezeichneten Gesichtsvideos eine größere Streuung der AU-Intensitäten und AU-Geschwindigkeiten als die während der Fahrten mit kognitiven Stressoren aufgezeichneten Gesichtsvideos. Dies deutet auf mehr Gesichtsaktivität während der Ablenkung durch emotionale Stressoren hin. Diese vorläufige Analyse zeigt, dass ein AU-basierter zweistufiger Ansatz das Potenzial haben könnte, Fahrerablenkung unter verschiedenen Bedingungen zu erkennen.

5 Herausforderungen und Ausblick

Drei zentrale Herausforderungen auf dem Gebiet der automatischen Analyse mentaler Zustände sind: (i) ein Mangel an Referenzdatensätzen für das Benchmarking von Algorithmen, (ii) interpersonelle Unterschiede in den Reaktionen auf Stimuli und (iii) ein Mangel an interpretierbaren Modellen für die automatische Erkennung mentaler Zustände. Im Rahmen dieser Doktorarbeit [Ha20] wurden einige grundlegende Schritte zur Bewältigung dieser Herausforderungen erforscht, hauptsächlich in Form von Forschungsarbeiten, die von Master-Studenten durchgeführt und von mir (mit-)betreut wurden. Als Ergebnis wurden (i) eine Reihe von Anforderungen für den Aufbau eines multimodalen Referenzdatensatzes zur Erkennung von menschlichem Stress entwickelt, (ii) interpersonelle Unterschiede in den Pupillenreaktionen auf einen Erregungsreiz untersucht und (iii) ein qualitativer Vergleich verschiedener erklärbarer KI Methoden durchgeführt, um tiefe neuronale Netze zur Unterscheidung von Schmerz, Freude und Ekel zu interpretieren und zu erklären.

Um das Feld der automatischen Analyse mentaler Zustände voranzubringen, sind Daten guter Qualität mit zuverlässigen Annotationen erforderlich. Angesichts der Schwierigkeit, große Datenmengen manuell zu annotieren, sollte die Verwendung von halbüberwachten und unüberwachten Methoden für die automatische Analyse mentaler Zustände in Zukunft intensiver untersucht werden. Da zwischenmenschliche Unterschiede den Ausdruck mentaler Zustände beeinflussen, sollten diese Unterschiede effektiv modelliert werden, und es sollten Systeme aufgebaut werden, die sich dynamisch an die Ausdrucksmerkmale des Individuums anpassen können. Die Generierung von Erklärungen für Schmerzwahrnehmungen ist eine sehr anspruchsvolle Aufgabe, die aus verschiedenen Perspektiven betrachtet werden sollte. All dies würde eine enge Zusammenarbeit zwischen verschiedenen Disziplinen und Forschungslabors auf der ganzen Welt erfordern.

Diese Doktorarbeit [Ha20] hat gezeigt, dass durch die Kombination von datengesteuerten maschinellen Lernansätzen mit deformierbaren Gesichtsmodellen und Zustandsschätzungs-

methoden die Vorhersageleistung für die AU-Erkennung verbessert werden kann, während gleichzeitig die Robustheit und Interpretierbarkeit bei der automatischen Analyse des mentalen Zustands auf der Grundlage von Gesichtsausdrücken gefördert wird. Zukünftige Forschung sollte sich auf die Integration verschiedener KI-Methoden und interdisziplinärem Wissen konzentrieren, um starke und ultrastarke KI-Systeme für die automatische Analyse mentaler Zustände aufzubauen.

Literaturverzeichnis

[Ba14] Bartlett, Marian Stewart; Littlewort, Gwen C.; Frank, Mark G.; Lee, Kang: Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions. Current Biology, 24(7):738–743, 2014.

[DBD18] Dapogny, A.; Bailly, K.; Dubuisson, S.: Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection. International Journal of Computer Vision, 126(2):255–271, Apr 2018.

[DD08] Dornaika, F.; Davoine, F.: Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion. International Journal of Computer Vision, 76(3):257–281, 2008.

[Do11] Dong, Yanchao; Hu, Zhencheng; Zhou, Yufeng; Uchimura, K.; Murayama, N.: A robust and efficient face tracker for driver inattention monitoring system. In: Intelligent Control and Automation (WCICA), 2011 9th World Congress on. S. 1212–1217, June 2011.

[DZC17] Ding, H.; Zhou, S. K.; Chellappa, R.: FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). S. 118–126, May 2017.

[EC11] Ekman, Paul; Cordaro, Daniel: What is Meant by Calling Emotions Basic. Emotion Review, 3(4):364–370, 2011.

[EF71] Ekman, Paul; Friesen, Wallace V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2):124–129, 1971.

[EF78] Ekman, Paul; Friesen, Wallace V.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA, 1978.

[Fr90] Freilinger, G.; Happak, W.; Burggasser, G.; Gruber, H.: Histochemical mapping and fiber size analysis of mimic muscles. Plastic and reconstructive surgery, 86(3):422–428, September 1990.

[Ga13] Garbas, J.; Ruf, T.; Unfried, M.; Dieckmann, A.: Towards Robust Real-Time Valence Recognition from Facial Expressions for Market Research Applications. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. S. 570–575, Sep. 2013.

[Ha16] Hassan, Teena; Seuss, Dominik; Wollenberg, Johannes; Garbas, Jens; Schmid, Ute: A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework. In: ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August - 2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016). Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, S. 1812–1817, 2016.

[Ha18] Hassan, T.; Seuß, D.; Ernst, A.; Garbas, J.: A Kalman filter with state constraints for model-based dynamic facial action unit estimation. In (Längle, Thomas; Puente León, Fernando; Heizmann, Michael, Hrsg.): Forum Bildverarbeitung 2018. KIT Scientific Publishing, 2018.

[Ha19] Hassan, T.; Seuß, D.; Wollenberg, J.; Weitz, K.; Kunz, M.; Lautenbacher, S.; Garbas, J.; Schmid, U.: Automatic Detection of Pain from Facial Expressions: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, S. 1–17, 2019.

[Ha20] Hassan, Teena Chakkalayil: Towards Robust and Interpretable Practical Applications of Automatic Mental State Analysis Using a Dynamic and Hybrid Facial Action Estimation Approach. Dissertation, Otto-Friedrich-Universität Bamberg, Bamberg : Otto-Friedrich-Universität, 10 2020.

[HWM78] Harper, Robert G.; Wiens, Arthur N.; Matarazzo, Joseph D.: Nonverbal communication: The state of the art. John Wiley & Sons, Oxford, England, 1978.

[KL14] Kunz, M.; Lautenbacher, S.: The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain. European Journal of Pain, 18(6):813–823, July 2014.

[KPM16] Kharghanian, R.; Peiravi, A.; Moradi, F.: Pain detection from facial images using unsupervised feature learning approach. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). S. 419–422, Aug 2016.

[Kr05] Krüger, Sven E.; Schafföner, Martin; Katz, Marcel; Andelic, Edin; Wendemuth, Andreas: Speech recognition with support vector machines in a hybrid system. In: INTERSPEECH-2005. S. 993–996, 2005.

[KRP12] Kaltwang, Sebastian; Rudovic, Ognjen; Pantic, Maja: Continuous Pain Intensity Estimation from Facial Expressions. In (Bebis, George; Boyle, Richard; Parvin, Bahram; Koracin, Darko; Fowlkes, Charless; Wang, Sen; Choi, Min-Hyung; Mantler, Stephan; Schulze, Jürgen; Acevedo, Daniel; Mueller, Klaus; Papka, Michael, Hrsg.): Advances in Visual Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, S. 368–377, 2012.

[Li15] Li, Yongqiang; Mavadati, S. Mohammad; Mahoor, Mohammad H.; Zhao, Yongping; Ji, Qiang: Measuring the intensity of spontaneous facial action units with dynamic Bayesian network. Pattern Recognition, 48(11):3417–3427, 2015.

[Lu11] Lucey, P.; Cohn, J. F.; Prkachin, K. M.; Solomon, P. E.; Matthews, I.: Painful data: the UNBC-McMaster shoulder pain expression archive database. In: Face and Gesture 2011. S. 57–64, March 2011.

[Ma19] Martinez, B.; Valstar, M. F.; Jiang, B.; Pantic, M.: Automatic Analysis of Facial Actions: A Survey. IEEE Transactions on Affective Computing, 10(3):325–347, July 2019.

[Mi88] Michie, Donald: Machine Learning in the Next Five Years. In: Proceedings of the 3rd European Conference on European Working Session on Learning. EWSL'88, Pitman Publishing, Inc., Marshfield, MA, USA, S. 107–122, 1988.

[MYL17] Meawad, Fatma; Yang, Su-Yin; Loy, Fong Ling: Automatic Detection of Pain from Spontaneous Facial Expressions. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. ICMI '17, ACM, New York, NY, USA, S. 397–401, 2017.

[Ni08] Niu, R.; Varshney, P. K.; Alford, M.; Bubalo, A.; Jones, E.; Scalzo, M.: Curvature nonlinearity measure and filter divergence detector for nonlinear tracking problems. In: 2008 11th International Conference on Information Fusion. S. 1–8, June 2008.

[Pa05] Parkinson, Brian: Do Facial Movements Express Emotions or Communicate Motives? Personality and Social Psychology Review, 9(4):278–311, 2005. PMID: 16223353.

[PS08] Prkachin, Kenneth M.; Solomon, Patricia E.: The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. PAIN, 139(2):267–274, 2008.

[Ru80] Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161–1178, 1980.

[SGC15] Sariyanidi, E.; Gunes, H.; Cavallaro, A.: Automatic Analysis of Facial Affect: a Survey of Registration, Representation, and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6):1113–1133, June 2015.

[SSB12] Savran, Arman; Sankur, Bulent; Bilge, M. Taha: Regression-based intensity estimation of facial action units. Image and Vision Computing, 30(10):774–784, 2012. 3D Facial Behaviour Analysis and Understanding.

[Ta17] Taamneh, S.; Tsiamyrtzis, P.; Dcosta, M.; Buddharaju, P.; Khatri, A.; Manser, M.; Ferris, T.; Wunderlich, R.; Pavlidis, I.: A multimodal dataset for various forms of distracted driving. Scientific Data, 2017.

[TLJ07] Tong, Y.; Liao, W.; Ji, Q.: Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(10):1683–1699, Oct 2007.

[VP12] Valstar, M. F.; Pantic, M.: Fully Automatic Recognition of the Temporal Phases of Facial Actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(1):28–43, Feb 2012.

[ZK14] Zafar, Z.; Khan, N. A.: Pain Intensity Evaluation through Facial Action Units. In: 2014 22nd International Conference on Pattern Recognition. S. 4696–4701, Aug 2014.

Teena Chakkalayil Hassan wurde 1984 in Kochi, Indien, geboren und erwarb dort 2006 ihren Bachelor in Informatik und Ingenieurwissenschaften. Anschließend war sie bis 2011 als Softwareingenieurin im Bereich Tele- und Datenkommunikation tätig. Von 2012 bis 2014 absolvierte sie an der Hochschule Bonn-Rhein-Sieg ihr Masterstudium im Fach Autonome Systeme. Darauffolgend war sie bis 2018 als Wissenschaftliche Mitarbeiterin am Fraunhofer Institut für Integrierte Schaltungen IIS in Erlangen beschäftigt und forschte dort im Bereich der automatischen Erkennung von mentalen Zuständen wie Schmerz und Stress aus inneren und äusseren Körpersignalen. Das war auch das Thema ihrer Doktorarbeit, die sie 2020 als externe Doktorandin an der Universität Bamberg in der Fakultät Wirtschaftsinformatik und Angewandte Informatik mit Auszeichnung absolvierte. Seit 2018 arbeitet sie an der Universität Bielefeld als Wissenschaftliche Mitarbeiterin und entwickelt dort im BMBF Projekt VIVA eine Software-Architektur für lebendige und soziale Mensch-Roboter-Interaktion.

Effiziente Modellierungs-, Verifikations- und Analysetechniken zur Verbesserung des Virtuellen Prototyp basierten Entwurfsablaufs für Eingebettete Systeme[1]

Vladimir Herdt[2]

Abstract: In dieser Dissertation wurden mehrere neuartige Ansätze entwickelt, die Modellierungs-, Verifikations- und Analyseaspekte abdecken, um den modernen auf *virtuellen Prototypen* (VP) basierten Entwurfsablauf stark zu verbessern. Die Beiträge sind im Wesentlichen in vier Bereiche unterteilt: Der erste Beitrag ist ein quelloffener RISC-V VP, der in SystemC TLM (engl. *Transaction Level Modeling*) implementiert ist und sowohl funktionale als auch nicht-funktionale Aspekte abdeckt. Der zweite Beitrag verbessert die Verifikation von VPs durch den Einsatz neuartiger formaler Verifikationsmethoden und fortschrittlicher automatisierter, überdeckungsgetriebener Testverfahren, die auf SystemC basierte VPs zugeschnitten sind. Der dritte Beitrag sind effiziente überdeckungsgetriebene Ansätze, die die VP-basierte SW-Verifikation und -Analyse verbessern. Der vierte und letzte Beitrag umfasst Ansätze, die eine Korrespondenzanalyse zwischen RTL (engl. *Register-Transfer Level*) und TLM durchführen, um die auf den verschiedenen Abstraktionsebenen verfügbaren Informationen gewinnbringend zu nutzen. Alle Ansätze wurden extensiv auf Basis von umfassenden Experimenten evaluiert, die ihre Effektivität für eine starke Verbesserung des VP-basierten Entwurfsablaufs eindeutig belegen.

1 Einführung

Eingebettete Systeme sind heutzutage in vielen verschiedenen Anwendungsbereichen weit verbreitet, diese reichen vom Internet der Dinge (engl. *Internet-of-Things*, IoT) über den Automobilbereich und die Produktion bis hin zu Kommunikations- und Multimediaanwendungen. Eingebettete Systeme bestehen dabei aus *Hardware* (HW) und *Software* (SW) Komponenten und sind typischerweise kleine ressourcenbeschränkte Systeme, die hoch spezialisiert sind, um anwendungsspezifische Lösungen umzusetzen. Daher erfordern Entwurfsabläufe für eingebettete Systeme effiziente und flexible Entwurfsraumexplorationstechniken, um alle anwendungsspezifischen Anforderungen, wie z.B. Energieverbrauch und Performanzanforderungen, zu erfüllen.

Darüber hinaus nimmt die Komplexität von eingebetteten Systemen kontinuierlich zu. Es werden zunehmend komplexere IP (engl. *Intellectual Property*) Komponenten (wie z.B. Prozessorkerne, domainenspezifische Beschleuniger und andere HW-Peripheriegeräte) auf der HW-Seite integriert und die Bedeutung der SW nimmt auch stetig zu. Diese erfüllt komplexe Funktionalität und ist auch für die Kontrolle und Koordination der HW-Komponenten zuständig. Insbesondere die SW-Komplexität ist in den letzten Jahren sehr stark angestiegen und daher sollte der SW im Entwurfsablauf dieselbe Priorität beigemessen werden wie

[1] Englischer Titel der Dissertation: "Efficient Modeling, Verification and Analysis Techniques to Enhance the Virtual Prototype based Design Flow for Embedded Systems"

[2] Universität Bremen, vherdt@uni-bremen.de

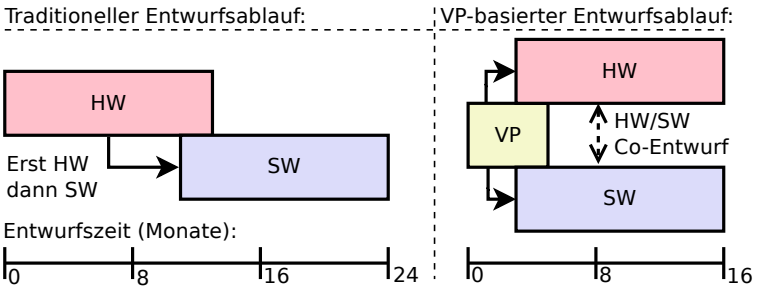


Abb. 1: Vergleich des Konzepts eines traditionellen Entwurfsablaufs (linke Seite) und eines VP-basierten Entwurfsablaufs (rechte Seite).

der HW um gute Ergebnisse zu erzielen. Heutzutage umfasst die (eingebettete) SW neben der eigentlichen Applikationslogik auch mehrere Abstraktionsschichten. Diese reichen vom Bootcode über Gerätetreiber bis hin zu vollwertigen Betriebssystemen in Kombination mit mehreren Bibliotheken (z.B. einem Netzwerkstack für die Kommunikation). Darüber hinaus wird die SW aufgrund ihrer Flexibilität zunehmend dazu genutzt, auch die Kontrolle bezüglich nicht-funktionaler Aspekte umzusetzen, wie z.B. die Implementierung von Strategien zur Steuerung des Energiemanagements von eingebetteten Systemen. Neben der Entwicklung der SW ist auch die Verifikation der SW entscheidend, um Fehler und Sicherheitslücken zu vermeiden. Ähnlich wie die SW-Entwicklungszeit nimmt auch die SW-Verifikationszeit mit steigender SW-Komplexität zu. Daher ist es sehr wichtig, so früh wie möglich mit der SW-Entwicklung und -Verifizierung zu beginnen, um die engen Vorgaben bezüglich der Entwurfszeit einzuhalten und ein qualitativ hochwertiges Produkt zu liefern.

Der traditionelle Entwurfsablauf für eingebettete Systeme ist unzureichend, um diese steigende Komplexität zu bewältigen. Der traditionelle Entwurfsablauf arbeitet im Wesentlichen sequenziell, indem zuerst die HW und dann die SW entwickelt wird (wie auf der linken Seite von Abb. 1 dargestellt). Der Grund dafür ist, dass die SW die HW benötigt, um ausgeführt zu werden, daher beginnt die SW-Entwicklung, sobald die HW-Entwicklung größtenteils abgeschlossen ist. Dies wiederum führt insbesondere aufgrund der steigenden SW-Komplexität zu erheblichen Verzögerungen im Entwurfsablauf.

Um diesem Problem zu begegnen, werden *virtuelle Prototypen* (VPs) zunehmend für die frühzeitige SW-Ausführung im Entwurfsablauf eingesetzt und ermöglichen so die parallele Entwicklung von HW und SW [DS14, Le12, Sc14, Ch19, Br15, HGD20]. Ein VP ist im Wesentlichen ein ausführbares abstraktes Modell der gesamten HW-Plattform und wird vorwiegend in SystemC TLM (engl. *Transaction Level Modeling*) [IEE11, OSC09][3] implementiert. Aus der SW-Perspektive bildet der VP die reale HW ab, d.h. der VP beschreibt die HW auf einer Ebene, die für die SW relevant ist. Zum Beispiel stellt der VP alle für die SW sichtbaren HW-Register zur Verfügung, abstrahiert aber von komplexen internen Kommunikationsprotokollen. Da es sich bei dem VP um ein abstraktes Modell der realen HW handelt, kann der VP viel schneller entwickelt werden als die HW. Sobald der VP zur

[3] Im Wesentlichen ist SystemC eine C++ Klassenbibliothek, die einen ereignisgesteuerten Simulationskernel nutzt und Bausteine für die Systementwicklung bereitstellt, um die Implementierung von VPs zu erleichtern, während TLM die Beschreibung der Kommunikation in Form von abstrakten Transaktionen ermöglicht.

Verfügung steht, kann er zur Ausführung der SW verwendet werden und ermöglicht somit eine frühzeitige Entwicklung und Testung der SW im Entwurfsablauf. Gleichzeitig dient der VP auch als ausführbares Referenzmodell für die nachfolgenden Schritte im Entwurfsablauf bei der HW-Entwicklung. Somit ermöglicht ein VP-basierter Entwurfsablauf die parallele Entwicklung von HW und SW (wie auf der rechten Seite von Abb. 1 dargestellt), was zu einer signifikanten Reduzierung der Entwurfszeit führt und auch eine agilere Entwicklung ermöglicht durch den Austausch von Feedback zwischen der HW- und SW-Ebene auf Basis des VP.

Im Folgenden wird zunächst der VP-basierte Entwurfsablauf näher beschrieben und anschließend werden die Beiträge dieser Arbeit, die den VP-basierten Entwurfsablauf signifikant verbessern, vorgestellt. Diese Arbeit ist eine Zusammenfassung der Dissertation [He20].

2 Virtueller Prototyp basierter Entwurfsablauf

Abb. 2 (auf der linken Seite) zeigt den VP-basierten Entwurfsablauf im Detail. Ausgangspunkt ist eine textuelle Spezifikation des eingebetteten Systems, die sowohl die funktionalen als auch die nicht-funktionalen Anforderungen an das eingebettete System enthält. Der VP-basierte Entwurfsablauf selbst ist im Wesentlichen in vier verschiedene Schritte unterteilt (jeweils mit einer anderen Hintergrundfarbe hervorgehoben):

1. VP-Modellierung (gelb)
2. VP-Verifikation (grün)
3. VP-basierte SW-Entwicklung (blau)
4. HW-Entwicklung (rot)

Diese vier Schritte im Entwurfsablauf werden dabei nicht strikt sequentiell nacheinander durchgeführt, sondern sind durchaus ineinander verschachtelt, da z.B. die SW und HW parallel entwickelt werden. Eine Übersicht über die einzelnen Schritte (nacheinander von oben nach unten, gruppiert durch gestrichelte Kästchen und hervorgehoben durch die jeweilige Hintergrundfarbe) ist auf der linken Seite von Abb. 2 dargestellt. Im Folgenden werden die vier Schritte detaillierter beschrieben.

Schritt 1 - VP-Modellierung Im ersten Schritt wird der SystemC-basierte VP erstellt. Der VP repräsentiert die gesamte HW-Plattform. Sie kann aus einem oder mehreren Prozessoren (potentiell mit speziellem Anwendungsfokus) sowie Peripheriegeräten in der HW bestehen. Zusätzlich zum funktionalen Verhalten werden typischerweise nicht-funktionale Verhaltensmodelle in den VP integriert, um neben der Ausführung der SW auch eine frühzeitige und genaue Abschätzungen zum nicht-funktionalen Verhalten des Systems zu erhalten. Diese Abschätzungen ermöglichen eine frühzeitige Exploration des Entwurfsraums und die Validierung nicht-funktionaler Eigenschaften wie dem Energieverbrauch und der Performanz.

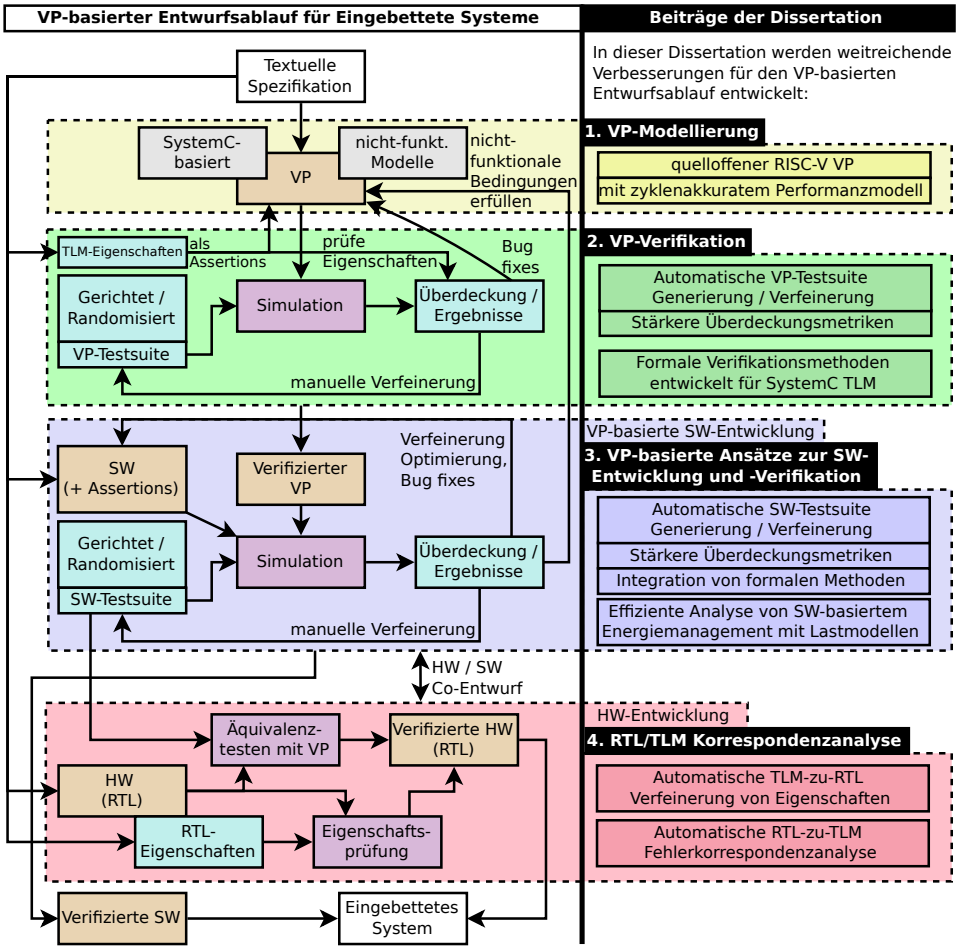


Abb. 2: Überblick zum VP-basierten Entwurfsablauf (linke Seite) sowie den korresponierenden Beiträgen dieser Arbeit (rechte Seite)

Schritt 2 - VP-Verifikation Im nächsten Schritt wird der VP verifiziert, was sehr wichtig ist, da der VP als ausführbares Referenzmodell für die nachfolgenden Schritte der HW-Entwicklung im Entwurfsablauf dient. Zur Verifikation werden dabei überwiegend simulationsbasierte Methoden eingesetzt. Sie erfordern eine umfangreiche Testsuite (d.h. eine Menge von Testfällen), um eine gründliche Verifikation zu gewährleisten. Testfälle werden entweder manuell erstellt oder auf Basis von zufallsbasierten Verfahren mit Randbedingungen generiert. Jeder Testfall repräsentiert eine bestimmte Eingabe für den VP. Für die Simulation von Testfällen wird eine Testbench bereitgestellt. Die Testbench instanziiert zunächst die zu testende VP-Konfiguration, anschließend leitet sie die Testeingaben an den VP und überprüft das Ausgabeverhalten des VP. Es ist möglich, den gesamten VP zu instanziieren oder einzelne VP-Komponenten (z.B. den Interrupt Controller oder den Prozessor) isoliert zu testen. Das erwartete Verhalten wird in Form von TLM-Eigenschaften spezifi-

ziert, die entweder direkt als (ausführbare) Assertions in den VP kodiert oder anhand der beobachteten Ergebnisse in der Testbench überprüft werden. Informationen bezüglich der Quellcodeüberdeckung werden verwendet, um die Qualität der Testsuite zu bewerten und den Prozess der Testfallgenerierung zu steuern. Ein hoher manueller Aufwand ist notwendig um eine gute Testsuite zu erstellen, die die Überdeckung maximiert. Der verifizierte VP dient dann als ausführbares Referenzmodell für die nachfolgende Schritte der SW- und HW-Entwicklung und ermöglicht die parallele Entwicklung von der SW und der HW im Entwurfsablauf. [4].

Schritt 3 - VP-basierte SW-Entwicklung Ein primärer Anwendungsfall von VPs ist die Ermöglichung einer frühzeitigen SW-Entwicklung im Entwurfsablauf. Neben der SW-Entwicklung und der Entwurfsraumexploration werden VPs auch zur Durchführung weiterer umfangreicher SW-Verifikations- und -Analyseaufgaben eingesetzt. Ähnlich wie bei der VP-Verifikation selbst, werden auch hier überwiegend simulationsbasierte Methoden eingesetzt. Allerdings arbeiten diese simulationsbasierten Methoden auf einer anderen Abstraktionsebene, da sie primär die SW und nicht den VP testen sollen. Daher repräsentiert jeder Testfall Eingaben für die SW. SW-Eigenschaften werden entweder als Assertions in der SW selbst kodiert oder anhand des beobachteten Ausgabeverhaltens des VPs überprüft. Der Prozess der Testfallgenerierung basiert wiederum größtenteils auf manuell erstellten gerichteten Testfällen sowie auf zufallsbasierten Techniken in Kombination mit Randbedingungen und wird durch die beobachteten Ergebnisse und die Quellcodeüberdeckung geleitet (wobei in diesem Fall die Überdeckung der SW berücksichtigt wird). Ähnlich wie bei der Verifikation des VPs ist ein erheblicher zeitaufwändiger und fehleranfälliger manueller Aufwand erforderlich, um eine umfangreiche Verifikation/Analyse der SW durchzuführen.

Basierend auf den erhaltenen Ergebnissen wird die SW verfeinert und optimiert, um Fehler zu beheben und die anwendungsspezifischen Anforderungen bezüglich Performanz und Energieverbrauch zu erfüllen. Insbesondere in den frühen Phasen des Entwurfsablaufs ist es auch möglich, den VP neu zu konfigurieren, um die HW für die anwendungsspezifischen Anforderungen zu optimieren.

Schritt 4 - HW-Entwicklung Die HW-Entwicklung findet parallel zur SW-Entwicklung statt. Das Ergebnis ist eine synthetisierbare Beschreibung der HW auf Registertransferebene (engl. *Register-Transfer Level*, RTL). Typischerweise werden zwei verschiedene Verifikationsmethoden in diesem Kontext eingesetzt:

- Durchführung eines Äquivalenztests der HW und des VPs, indem die in Schritt 3 erhaltene(n) SW-Testsuite(s) wiederverwendet werden.

[4] Aufgrund der immensen Komplexität in Kombination mit dem erheblichen manuellen Aufwand für die Verifikation des VP wird zu Beginn nur eine vorläufige Verifikation des VP durchgeführt. Die Verifikationsaufgabe ist damit typischerweise ein kontinuierlicher Prozess der neben der SW- und HW-Entwicklung weiterhin stattfindet. Daher sind Verbesserungen im Bereich der VP-Verifikation sehr wichtig, um die Ausbreitung von Fehlern und damit kostspielige Iterationen möglichst zu vermeiden.

- Anwendung von Verfahren der Eigenschaftsprüfung auf Basis von RTL-Eigenschaften, die auf der Grundlage der textuellen Spezifikation erstellt werden.

Abschließend werden die verifizierte SW und HW zu dem finalen eingebetteten System integriert. Die eigentliche HW-Entwicklung sowie HW-Verifikationsmethoden auf RTL und darunter stehen nicht im Fokus dieser Arbeit.

3 Beitrag der Dissertation

Der VP-basierte Entwurfsablauf verbessert den traditionellen Entwurfsablauf erheblich und hat sich auch in verschiedenen industriellen Anwendungsfällen bewährt [DS14, Oe14, Ko06, KV14]. Allerdings hat dieser moderne VP-basierte Entwurfsablauf noch immer signifikante Schwächen, insbesondere durch den hohen manuellen Aufwand für Verifikations- und Analyseaufgaben, der sowohl zeitaufwendig als auch fehleranfällig ist. Diese Dissertation schlägt mehrere neuartige Ansätze vor, um den VP-basierten Entwurfsablauf stark zu verbessern und liefert damit Beiträge zu jedem der vier Hauptschritte im VP-basierten Entwurfsablauf. Eine Übersicht zu den Beiträgen der Dissertation ist auf der rechten Seite von Abb. 2 zu finden. Die Beiträge sind in vier Bereiche gruppiert, die (größtenteils) den vier Schritten des VP-basierten Entwurfsablaufs entsprechen und im Folgenden detaillierter vorgestellt werden:

1. VP-Modellierung
2. VP-Verifikation
3. VP-basierte Ansätze für die SW-Verifikation und -Analyse
4. RTL/TLM Korrespondenzanalyse

Beitragbereich 1 - VP-Modellierung Der erste Beitrag dieser Arbeit ist ein in SystemC TLM implementierter quelloffener RISC-V VP.

RISC-V ist eine offene und freie *Instruktionssatzarchitektur* (ISA) [WA19a, WA19b], die lizenzgebührenfrei verwendet werden kann. Ähnlich wie im Bereich der quelloffenen SW zu beobachten, gewinnt auch die RISC-V ISA eine große Bedeutung sowohl im akademischen Umfeld als auch in der industriellen Anwendung. Insbesondere für eingebettete Systeme, z.B. im IoT-Bereich, bietet RISC-V weitreichende Vorteile und erfährt sehr große Popularität. Rund um RISC-V ensteht ein großes und ständig wachsendes Ökosystem, dessen Umfang von verschiedenen HW-Implementierungen (d.h. RISC-V Prozessoren) bis zu SW-Bibliotheken, Betriebssystemen und Compilern reicht. Darüber hinaus sind mehrere *Instruktionssatzsimulatoren* (ISS) verfügbar die auf eine möglichst hohe Simulationsperformanz ausgelegt sind. Der starke fokus der ISSs auf eine hohe Simulationsleistung führt jedoch dazu, dass diese nur sehr schwer erweitert werden können um weitere Anwendungsfälle auf Systemebene zu unterstützen, wie z.B. die Entwurfsraumexploration, Validierung von Anforderungen bezüglich der Performanz und des Energieverbrauch, oder auch die Analyse von komplexen HW/SW-Interaktionen. Das Ziel des RISC-V VPs dieser Dissertation ist es,

diese Lücke im RISC-V Ökosystem zu schließen und weitere Forschung und Entwicklung in diesem Bereich anzuregen.

Der RISC-V VP stellt eine 32/64 Bit ISS zusammen mit einem Satz an wesentlicher Peripherie bereit und bietet Unterstützung für die Simulation von Mehrkernsystemen. Darüber hinaus bietet der VP auch SW-Debugging (über die Eclipse-IDE) und Methoden zur Messung der Überdeckung an. Zudem werden die Betriebssysteme FreeRTOS, Zephyr und Linux unterstützt. Der VP ist als erweiterbare und konfigurierbare Plattform (als ein Beispiel wird auch eine Konfiguration bereitgestellt, die dem RISC-V HiFive1 Board von SiFive entspricht) mit einem generischen Bussystem konzipiert und im standardkonformem SystemC TLM implementiert. Der letzte Punkt ist sehr wichtig, da er die Nutzung modernster SystemC-basierter Modellierungstechniken ermöglicht, die für die vorher genannten Anwendungsfälle auf Systemebene essenziell sind. Schließlich ermöglicht der VP eine deutlich höhere Simulationsperformanz im Vergleich zu RTL und ist dabei akkurater als bestehende ISSs.

Darüber hinaus integriert der VP ein effizientes Prozessorperformanzmodell in der ISS, um eine schnelle und genaue Performanzanalyse für RISC-V basierte Systeme zu ermöglichen. Das Performanzmodell ist mit dem ISS über eine Reihe von wohldefinierten Schnittstellen verbunden, die die funktionalen von den nicht-funktionalen Aspekten entkoppeln und eine einfache Rekonfiguration des Modells ermöglichen. Als Beispiel wird eine Konfiguration des Performanzmodells passend für das RISC-V HiFive1 Board von SiFive bereitgestellt.

Beitragbereich 2 - VP-Verifikation Diese Dissertation verbessert den VP-Verifikationsablauf durch Entwicklung neuartiger formaler Verifikationsmethoden und verbesserter automatisierter Testverfahren auf Basis von Überdeckungsdaten. Die Verfahren sind dabei auf SystemC-basierte Designs zugeschnitten.

Formale Verifikationsmethoden können die Korrektheit eines SystemC Designs in Bezug auf eine Menge von Eigenschaften beweisen. Die formale Verifikation ist jedoch sehr schwierig, da alle möglichen Eingaben sowie Prozessausführungsfolgen berücksichtigen werden müssen, was schnell zu einer Zustandsraumexplosion führen kann. In dieser Dissertation wurden effiziente symbolische Simulationstechniken für SystemC entwickelt, um der Komplexitätsproblematik entgegenzuwirken. Neben der grundlegenden symbolischen Simulationstechnik, die die Basis für eine effiziente Zustandsraumexploration bildet, wurden mehrere wichtige Erweiterungen und Optimierungen entwickelt, wie SSR (engl. *State Subsumption Reduction*) und CSS (engl. *Compiled Symbolic Simulation*). SSR bietet Unterstützung für die Verifikation von zyklischen Zustandsräumen, indem die Reexploration symbolischer Zustände verhindert wird. Damit wird eine vollständige Verifikation ermöglicht. CSS ist eine starke Optimierungstechnik, die die symbolische Simulationsengine eng mit dem zu verifizierenden SystemC Design integriert, um die Verifikationsleistung durch Einsatz von nativer Ausführung drastisch zu steigern. Die entwickelten Verfahren haben den Stand der Technik bezüglich der formalen Verifikation von SystemC signifikant verbessert. So konnten einzelne VP-Komponenten, wie ein Interrupt Controller, effizient formal untersucht werden. Betrachtet man aber die gesamte VP-Platform in Kombination, besteht immer noch die Gefahr in eine Zustandsraumexplosion zu geraten.

Daher wurden in dieser Dissertation auch fortschrittliche überdeckungsgetriebene Methoden betrachtet, die auf Testfallgenerierung und Simulation beruhen. Insbesondere betrachtet wurden die beiden sogenannten Verfahren *Data Flow Testing* (DFT) und *Coverage-guided Fuzzing* (CGF), die sich beide in der SW-Domäne als sehr effektiv erwiesen haben, zugeschnitten auf die Verifikation von VPs. Für DFT wurden dabei SystemC spezifische Überdeckungsmetriken entwickelt, die die SystemC Semantik berücksichtigen. Dazu gehört das nicht-präemptive Prozessscheduling und eine ereignisbasierte Synchronisation. CGF wurde speziell für die Verifikation von RISC-V basierten ISSs optimiert, durch angepasste funktionale Überdeckungsmetriken sowie neu entwickelten Mutationsverfahren, und angewendet. Es wurden Fehler in verschiedenen RISC-V Simulatoren gefunden.

Beitragsbereich 3 - VP-basierte Ansätze für die SW-Verifikation and -Analyse
Der erste Beitrag in diesem Bereich sind neuartige VP-basierte Ansätze für die SW-Verifikation. Dabei werden im Vergleich zum normalen VP-basierten Entwurfsablauf stärkere Überdeckungsmetriken betrachet sowie automatisierte Verfahren zur Testgenerierung und formale Methoden eingesetzt. Umfassende SW-Verifikation ist sehr wichtig um Fehler und Sicherheitslücken zu vermeiden.

Der erste Ansatz kombiniert konkolisches Testen mit einer VP-basierten Simulation. Konkolisches Testen ist im Wesentlichen eine automatisierte Technik, die sukzessive neue Pfade durch die SW erkundet, indem symbolische Randbedingungen gelöst werden, die parallel zur konkreten Ausführung mitprotokolliert werden. Eine VP-basierte Integration birgt verschiedene Herausforderungen aufgrund der komplexen HW/SW-Interaktionen und der Peripherie. Die entwickelte Lösung ermöglicht eine hohe Simulationsleistung mit akkuraten Ergebnissen und einem vergleichsweise geringem Aufwand für die Integration von Peripherie mit konkolischer Ausführung. Auf dieser Basis konnten Fehler im TCP/IP Stack des FreeRTOS Betriebssystems aufgezeigt werden, was die Skalierbarkeit der Methodik auf realistische Softwaresysteme demonstriert.

Obwohl im Normalfall sehr effizient, beruht konkolisches Testen auf symbolischen Randbedingungen und ist daher anfällig für Skalierbarkeitsprobleme. Daher wurde ein zweiter Verifikationsansatz entwickelt, der modernes CGF mit VPs kombiniert für die effiziente und skalierbare Verifikation eingebetteter SW. Der Fuzzingprozess wird dabei auf Basis einer Feedbackschleife über die kombinierte Überdeckung der SW und VP-basierten Peripherie gesteuert. Dies ermöglicht einen sehr effizienten Testgenerierungsprozess.

Neben dem korrekten funktionalen Verhalten, ist eine hohe Performanz in Kombination mit einem geringen Energieverbrauch eine wichtige Anforderung für eingebettete Systeme. Aufgrund der einfachen Handhabung und Flexibilität werden Strategien für das Energiemanagement oft in SW implementiert. Diese werden analysiert, indem das nicht-funktionale Verhalten des VP neben der SW-Ausführung beobachtet wird.

Dafür wurden in dieser Dissertation zwei neuartige VP-basierte Verfahren zur Validierung von Energiemanagementstrategien entwickelt. Erstens, eine auf Randbedingungen basierte Enkodierung zur Spezifikation von abstrakten Anwendungslasten, die die Generierung einer umfassenden Testsuite ermöglicht mit anwendungsspezifischen Auslastungsprofilen. Und zweitens, ein automatisiertes Verfahren zur Generierung von Anwendungslasten auf

Basis von Testfällen, die die Überdeckung in Bezug auf die Energiemanagementstrategien maximieren.

Beitragsbereich 4 - RTL/TLM Korrespondenzanalyse Der finale Beitrag dieser Dissertation sind zwei Ansätze, die eine Korrespondenzanalyse zwischen TLM und RTL durchführen.

Der erste entwickelte Ansatz ermöglicht eine automatisierte TLM-zu-RTL Eigenschaftsverfeinerung. Dabei werden TLM-Eigenschaften transformiert zu RTL-Eigenschaften, die als Ausgangspunkt für die Eigenschaftsprüfung der HW dienen. Dies vermeidet eine fehleranfällige und zeitaufwändige manuelle Transformation.

Der zweite entwickelte Ansatz führt eine RTL-zu-TLM Fehlerkorrespondenzanalyse durch. Dabei werden korresponierende TLM-Fehler für transiente Bitkipper auf RTL identifiziert. Die erzielten Ergebnisse können die Genauigkeit einer VP-basierten Fehlereffektsimulation deutlich verbessern. Eine Fehlereffektsimulation funktioniert im Wesentlichen durch Injektion von Fehlern in den VP während der SW-Ausführung, um die Robustheit der SW gegenüber verschiedenen HW-Fehlern zu prüfen. Eine solche Analyse ist sehr wichtig für eingebettete Systeme, die in empfindlichen Umgebungen arbeiten oder sicherheitskritische Aufgaben ausführen, um sich gegen Auswirkungen von z.B. Strahlung und Alterung zu schützen.

Fazit Zusammenfassend lässt sich sagen, dass die Beiträge der Dissertation den VP-basierten Entwurfsablauf stark verbessern, wie von den ausführlichen Experimenten der Dissertation belegt wird [He20]. Einer der Hauptvorteile ist die drastisch verbesserte Verifikationsqualität in Kombination mit einem deutlich geringeren Verifikationsaufwand aufgrund der umfangreichen Automatisierung. Dadurch wird einerseits die Anzahl der unentdeckten Bugs reduziert und die Gesamtqualität des eingebetteten Systems signifikant verbessert. Zum anderen verkürzt sich dadurch die Entwicklungszeit und damit die Markteinführung.

Literaturverzeichnis

[Br15] Bringmann, O.; Ecker, W.; Gerstlauer, A.; Goyal, A.; Mueller-Gritschneder, D.; Sasidharan, P.; Singh, S.: The next generation of virtual prototyping: Ultra-fast yet accurate simulation of HW/SW systems. In: DATE. S. 1698–1707, 2015.

[Ch19] Charif, Amir; Busnot, Gabriel; Mameesh, Rania; Sassolas, Tanguy; Ventroux, Nicolas: Fast Virtual Prototyping for Embedded Computing Systems Design and Exploration. In: RAPIDO Workshop. S. 3:1–3:8, 2019.

[DS14] De Schutter, Tom: Better Software. Faster!: Best Practices in Virtual Prototyping. Synopsys Press, March 2014.

[He20] Herdt, Vladimir: Efficient Modeling, Verification and Analysis Techniques to Enhance the Virtual Prototype based Design Flow for Embedded Systems. Dissertation, University of Bremen, 2020.

[HGD20] Herdt, Vladimir; Große, Daniel; Drechsler, Rolf: Enhanced Virtual Prototyping: Featuring RISC-V Case Studies. Springer, 2020.

[IEE11] IEEE Std. 1666. IEEE Standard SystemC Language Reference Manual, 2011.

[Ko06] Kong, J.; Yoo, B.; Song, D.; Nam, H. J.; Hwang, J.; Kim, J.; Lee, S.; Eo, S.; Yoo, S.; Choi, K.; Jin, H.; Kim, J.; Lee, S.; Hong, S.: Creation and utilization of a virtual platform for embedded software optimization:: an industrial case study. In: CODES+ISSS. S. 235–240, 2006.

[KV14] Kramer, Angela; Vaupel, Martin: , Virtual Platforms for Automotive: Use Cases, Benefits and Challenges. https://dvcon-europe.org/sites/dvcon-europe.org/files/archive/2014/proceedings/T01_tutorial_part3.pdf, 2014.

[Le12] Leupers, R.; Schirrmeister, F.; Martin, G.; Kogel, T.; Plyaskin, R.; Herkersdorf, A.; Vaupel, M.: Virtual platforms: Breaking new grounds. In: DATE. S. 685–690, 2012.

[Oe14] Oetjens, J. H.; Bannow, N.; Becker, M.; Bringmann, O.; Burger, A.; Chaari, M.; Chakraborty, S.; Drechsler, R.; Ecker, W.; Grüttner, K.; Kruse, T.; Kuznik, C.; Le, H. M.; Mauderer, A.; Müller, W.; Müller-Gritschneder, D.; Poppen, F.; Post, H.; Reiter, S.; Rosenstiel, W.; Roth, S.; Schlichtmann, U.; von Schwerin, A.; Tabacaru, B. A.; Viehl, A.: Safety evaluation of automotive electronics using Virtual Prototypes: State of the art and research challenges. In: DAC. S. 1–6, 2014.

[OSC09] OSCI. OSCI TLM-2.0 Language Reference Manual, 2009.

[Sc14] Schuster, Thomas; Meyer, Rolf; Buchty, Rainer; Fossati, Luca; Berekovic, Mladen: SoCRocket - A virtual platform for the European Space Agency's SoC development. In: ReCoSoC. S. 1–7, 2014. Available at http://github.com/socrocket.

[WA19a] Waterman, Andrew; Asanović, Krste, Hrsg. The RISC-V Instruction Set Manual; Volume I: Unprivileged ISA. 2019.

[WA19b] Waterman, Andrew; Asanović, Krste, Hrsg. The RISC-V Instruction Set Manual; Volume II: Privileged Architecture. 2019.

Vladimir Herdt erhielt 2014 den M.Sc. in Informatik von der Universität Bremen, Deutschland. Danach begann er als Doktorand in der Arbeitsgruppe Rechnerarchitektur unter der Betreuung von Prof. Rolf Drechsler. Im Jahr 2020 erhielt er den Dr.-Ing. Titel in Informatik von der Universität Bremen. Seit 2020 ist er als Senior Researcher an der Universität Bremen und am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) tätig. Seine aktuellen Forschungsinteressen umfassen Virtual Prototyping sowie Verifikations- und Analysetechniken mit einem besonderen Fokus auf RISC-V. In diesen Bereichen veröffentlichte er mehr als 30 peer-reviewed Journal- und Konferenzbeiträge mit zwei Best-Paper-Candidates und einem Best-Paper-Award bei der FDL. Er erhielt den Springer BestMasters Award und den DAC Young Fellow Award.

Über Diskriminierung durch Künstliche Intelligenz[1]

Vasileios Iosifidis[2]

Abstract: Heutzutage gibt es eine Fülle von Daten, was aber nicht gleichbedeutend mit mehr Informationen ist. Die Datenqualität ist ein wichtiges Thema, da Algorithmen für maschinelles Lernen auf Daten basieren. Wir untersuchen das Class-Imbalance-Problem, das Algorithmen für maschinelles Lernen dramatisch beeinträchtigt. Es führt dazu, dass KI-Modelle effektiv eine bestimmte Klasse lernen, während sie andere Klassen aufgrund von schiefen Label-Verteilungen ignorieren. Dies führt dazu, dass Modelle des maschinellen Lernens, die in Bereichen mit großer gesellschaftlicher Bedeutung eingesetzt werden, Gruppen von Menschen oder Individuen, die in den Daten nicht gut repräsentiert sind, voreingenommen gegenüberstehen. In dieser Arbeit wird die Klassenungleichheit beim fairnessbasierten Lernen untersucht. Unsere Methoden bekämpfen die Klassenungleichheit und liefern faire Ergebnisse für unterrepräsentierte Personen, die von Algorithmen diskriminiert werden.

1 Einführung

Der Aufstieg des Web 2.0 und die rasanten technologischen Fortschritte haben eine Datenexplosion verursacht, die für Entscheidungsunterstützungssysteme von Vorteil ist. Solche Systeme stützen sich auf historische Daten, um neue Hypothesen abzuleiten, die zu den Daten passen ohne explizit kodiert zu werden; dadurch sind solche Systeme in einer Vielzahl von Bereichen anwendbar. Entscheidungsunterstützungssysteme profitieren von großen Datenmengen, was die Datenverfügbarkeit extrem wertvoll macht. Dennoch bedeuten mehr Daten nicht unbedingt mehr Information, vielmehr spielt die Qualität der Daten eine entscheidende Rolle. Bei vielen realen Problemen sind die Daten unvollständig, enthalten Rauschen bzw. Ausreißer, haben kodierte Verzerrungen bzw. schiefe Klassenverteilungen, d.h. Klassenungleichheit. Es ist von entscheidender Bedeutung, mit solchen Problemen umzugehen, da Algorithmen für maschinelles Lernen, die auf Daten aus der realen Welt trainiert werden, dazu neigen, bestehende Fehler zu verstärken.

Das Problem der Klassenungleichheit ist recht alt und wird häufig in allen Volumenbereichen von Daten beobachtet. Class-Imbalance bezieht sich auf ungleiche Klassenverteilungen, d. h., wenn eine Klasse (Mehrheit) die anderen Klassen (Minderheitsklassen) in Bezug auf die Anzahl der Instanzen bei weitem übertrifft. Dieses Problem kann zu dis-

[1] Englischer Titel der Dissertation: Semi-supervised Learning and Fairness-aware Learning under Class Imbalance

[2] Institut für Verteilte Systeme – Leibniz Universität Hannover, iosifidis@kbs.uni-hannover.de

proportionalen Fehlerraten zwischen verschiedenen Klassen führen.

Konventionelle Algorithmen für maschinelles Lernen berücksichtigen dieses Problem nicht und übertragen es auf ihre Ergebnisse. Ein weiteres Problem, das in unausgewogenen Datensätzen liegt, ist das Problem der seltenen Fälle. Seltene Fälle (auch klasseninternes Ungleichgewicht oder Gruppenungleichgewicht genannt) werden oft als Ausreißer oder Rauschen behandelt, da sie der Mehrheit der Instanzen innerhalb einer Klasse nicht sehr ähnlich sind, z. B. kann ein medizinischer Datensatz, der gesunde und kranke Patienten enthält, kranke Patienten enthalten, die an einer sehr seltenen Krankheit leiden.

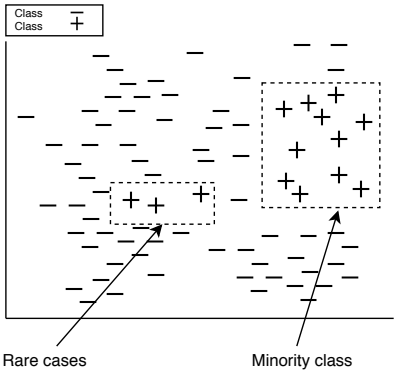


Abb. 1: Seltene Fälle in ungleichgewichtigen Daten

Ein weiteres häufiges Problem mit Daten ist, dass sie oft nicht beschriftet sind. Um qualitative Beschriftungen für die Daten zu erhalten, muss man sich auf menschliche Annotatoren verlassen; die Aufgabe der menschlichen Annotation ist jedoch zeitaufwändig und teuer. Hinzu kommt, dass die Datengeschwindigkeit heutzutage Petabytes pro Tag übersteigt (Geschwindigkeit bezieht sich auf die Datenerzeugungsrate), was es unmöglich macht, all diese Daten zu annotieren, indem man sich auf menschliche Annotationen verlässt. In vielen Bereichen sind die beschrifteten Daten begrenzt, und die nicht beschrifteten Daten sind unverhältnismäßig mehr als die beschrifteten. Ein ganzes Forschungsgebiet im Bereich des maschinellen Lernens, das so genannte halbüberwachten Lernen, konzentriert sich auf die Kombination von beschrifteten und unbeschrifteten Daten, um Labels für die unbeschrifteten Daten zu schätzen, die anschließend von konventionellen Entscheidungsunterstützungssystemen verwendet werden können.

Bewältigung des Klassenungleichgewichts beim halbüberwachten Lernen. Der erste Aspekt dieser Arbeit [Io20] konzentriert sich auf das Problem der Ausbreitung von Klassenungleichgewichten in der halbüberwachten Umgebung. Halbüberwachte Methoden berücksichtigen nicht die ungleiche Klassenverteilung; daher neigen sie dazu, klassenungleiche Ergebnisse zu propagieren, was die Gesamtleistung verschlechtern kann. Indem wir die Klassenverteilungen ausgleichen, zwingen wir diese Methoden dazu, alle Klassen effektiv zu lernen, um eine solche Verschlechterung zu vermeiden.

Schließlich werfen Daten und die Nutzung von maschinellem Lernen in Bereichen mit großer gesellschaftlicher Bedeutung Bedenken hinsichtlich Datenschutz, Fairness sowie rechtlicher und ethischer Richtlinien auf. Daten enthalten oft Proxy-Attribute zu anderen Attributen wie Rasse oder Geschlecht (auch geschützte Attribute genannt), was dazu führen kann, dass ein Entscheidungsunterstützungssystem diskriminierende Ergebnisse produziert. Beispielsweise kann ein Entscheidungsunterstützungssystem Leistungen für Personen oder Personengruppen ablehnen, die bestimmte Merkmale (geschützte Gruppe) aufweisen, und

andere (nicht geschützte Gruppe) bevorzugen. Darüber hinaus spiegeln die Daten gesellschaftliche Verzerrungen wider und sind nicht repräsentativ für die gesamte Bevölkerung (Stichprobenverzerrung). Darüber hinaus können Systemverzerrungen dazu führen, dass verzerrte Daten generiert werden, die zu Modellen führen, die solche diskriminierenden Maßnahmen weiter verstärken, wie z. B. bei der vorausschauenden Polizeiarbeit. Schließlich haben Modelle des maschinellen Lernens ihre eigenen Annahmen und Verzerrungen (Model Bias), die ihre Generalisierungsleistung deutlich beeinflussen. Aufgrund der Komplexität von Big Data und maschinellen Lernalgorithmen und deren komplexen Wechselwirkungen ist die Integration von fairnessfördernden Eingriffen in den Lernprozess unerlässlich.

Abschwächen von Klassenungleichheit und Unfairness in überwachten Modellen. Der zweite Aspekt dieser Arbeit [Io20] konzentriert sich auf das kombinierte Problem der Klassenungleichheit und diskriminierender Ergebnisse beim überwachten Lernen. Obwohl es eine Fülle von Arbeiten gibt, die darauf abzielen, diskriminierende Ergebnisse von überwachten Modellen abzuschwächen, berücksichtigen sie nicht das Problem der Klassenungleichheit; daher verwerfen diese Methoden die große Mehrheit der qualifizierten Instanzen, die zur Minderheitenklasse gehören, aufgrund ihrer Unfähigkeit, die Minderheitenklasse effektiv zu lernen.

2 Klassenungleichheit beim halb-überwachten Lernen

Halbüberwachte Methoden werden häufig verwendet, um die Klassifizierungsleistung zu verbessern, indem große Mengen an unbeschrifteten Daten genutzt werden. Diese Methoden sind jedoch anfällig für die Ausbreitung von Klassenungleichheit, wenn die verwendeten Daten schiefe Klassenverteilungen haben.

Modelle, die auf unausgewogenen Daten trainiert werden, lernen hauptsächlich die Mehrheitsklasse, während die Minderheit ignoriert wird [HM13]. In unserem Fall wird das Problem durch die Propagierung der vorhergesagten Labels in den nächsten Runden des halbüberwachten Lernprozesses verschärft. Infolgedessen ist die Tendenz der Modelle zur Mehrheitsklasse in den endgültigen Modellen viel höher.

In dieser Arbeit werden halbüberwachte Methoden wie Co-Training [BM98] und Self-Learning [Fr67] ausführlich an der Aufgabe der Sentimentanalyse (Textklassifikation) evaluiert. Wir verwenden halbüberwachten Methoden, um ein großes Korpus von meinungsbetonten Kurztexten zu annotieren, genannt *TSentiment15* [IN17], zu annotieren und der Community öffentlich zur Verfügung zu stellen. TSentiment15[3] besteht aus mehr als 200 Millionen englischen Kurztexten und ist damit der erste groß angelegte Datensatz von meinungsbetonten Kurztexten überhaupt. Um mit der Ausbreitung von Klassenungleichheit umzugehen, kombinieren wir halbüberwachten Methoden mit Augmentationstechniken wie Over-Sampling, Under-Sampling, Verzerrung und semantischer Ähnlichkeit und analysieren die Vor- und Nachteile der einzelnen Methoden. Unsere Experimente zeigen,

[3] `https://www.l3s.de/~iosifidis/TSentiment15/`

dass die Kopplung von halbüberwachten und Augmentierungsmethoden die standardmäßigen halbüberwachten Methoden deutlich übertrifft.

Traditionell wird ein Klassenungleichgewicht durch Oversampling (von der Minderheitsklasse) und/oder Under-Sampling (von der Mehrheitsklasse) behandelt. Beide Ansätze sind jedoch mit Einschränkungen verbunden; beim Under-Sampling kann man nicht kontrollieren, welche Informationen über die Mehrheitsklasse weggeworfen werden, während beim Oversampling keine neuen Informationen zum Trainingssatz hinzugefügt werden, sondern einige Instanzen aus der Minderheitsklasse repliziert werden, was einen stärkeren Effekt auf den Klassifikator hat. In dieser Arbeit entwickeln wir neben dem Oversampling und dem Under-Sampling auch Datenaugmentierungstechniken zur Generierung plausibler Pseudo-Instanzen für die Minderheitenklasse, um das Klassenungleichgewicht zu korrigieren.

Die Augmentation wird in den halbüberwachten Lernprozess integriert [IN19b], um das Problem des Klassenungleichgewichts über die Trainingsiterationen zu kontrollieren. Auf diese Weise werden in jeder Iteration die Datenverteilungen vor dem Trainingsprozess der Co-Training- oder Self-Learning-Methoden ausgeglichen. Die von uns vorgeschlagene Vorverarbeitungsarchitektur kann auf alle bestehenden halbüberwachten Lernmethoden angewendet werden.

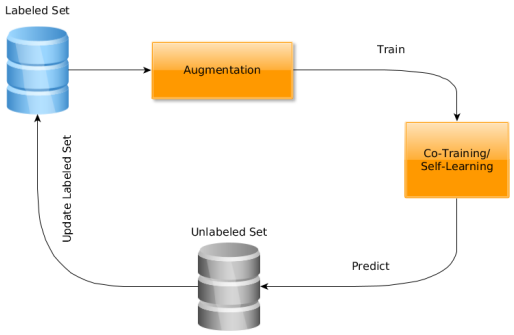


Abb. 2: Augmentierungs-unterstütztes halb-überwachtes Lernen

Wir schlagen außerdem zwei Augmentationstechniken vor: die semantische Augmentation und die Blankout-Korruption zum Ausgleich der Klassen, jenseits der bekannten Under-Sampling- und Over-Sampling-Ansätze. Ersteres nutzt die semantische Ähnlichkeit zwischen Wörtern (über word-embeddings [Mi13]), um semantisch ähnliche Instanzen aus den ursprünglichen Instanzen zu erzeugen. Letztere korrumpiert die Trainingsinstanzen, indem sie Informationen (Wörter) aus den Originalinstanzen entfernt und somit korrumpierte Versionen der Originalinstanzen erzeugt. Beide Techniken sind Feature-Transformationstechniken, d.h. sie verändern die einzelnen Wörter. Um sicherzustellen, dass die augmentierten Instanzen ihre Labels behalten, *transformieren* wir nur nicht-sentimentale Wörter.

Semantische Ähnlichkeit: Um Pseudo-Instanzen der gleichen Klasse zu generieren, verwenden wir die semantische Ähnlichkeit von Wörtern, wie sie durch Word-Embeddings [Mi13] erfasst wird. Unsere Idee ist es, Pseudo-Instanzen zu erzeugen, indem wir Wörter im Originaldokument durch semantisch ähnliche Wörter ersetzen. Insbesondere generieren wir für ein ausgewähltes Wort w, das in einem Originaldokument d vorkommt, seine ähnlichen Wörter basierend auf ihren Einbettungsvektoren und wählen zufällig eines der Top-k ähnlichen Wörter w' aus, um das Original w in der Pseudo-Instanz d' zu ersetzen. Wir berücksichtigen nur Wörter, die sentimental sind. Als Beispiel könnte der Text “I love

this car very much" den Text "I like this car very much" erzeugen. Es gibt verschiedene Word-Embedding-Varianten, die auf den für ihr Training verwendeten Daten basieren. Für jedes sentimentale Wort in unserem Korpus generieren wir die Top-k ähnlichsten Wörter basierend auf Glove-Einbettungen (in unseren Experimenten setzen wir $k = 10$). Eine Liste der Top-k ähnlichen Wörter, genannt *similarity-list*, wird aus dem oben genannten Prozess generiert, die 33.037 Begriffe enthält.

Blankout-Korruption: Die Korruption von Bildern durch Rauschen ist eine gängige Transformation in der Bilddomäne und zielt darauf ab, robustere Modelle für maschinelles Lernen zu erstellen. Wir verfolgen eine ähnliche Idee für Text: Wir erzeugen Pseudo-Instanzen, indem wir ein (zufällig ausgewähltes) Wort aus dem Originaldokument löschen. Um sicherzustellen, dass das Klassenlabel erhalten bleibt, entfernen wir keine Negationen oder sentimentale Wörter. Um sicherzustellen, dass das resultierende Dokument noch plausibel ist, wenden wir Korruption auf ausreichend lange Dokumente an, nämlich auf Dokumente mit mindestens vier Wörtern. Als Beispiel könnte der Text "I don't like the morning traffic" in "I don't like the traffic" umgewandelt werden. Da unser Ziel ein Klassenungleichgewicht ist, erzeugen wir Pseudo-Instanzen nur für die Minderheitenklasse.

Vor- und Nachteile: Augmentierungsverfahren wie semantische Ähnlichkeit und Ausblendung haben ein ähnliches Ziel wie Oversampling/Undersampling, nämlich ein Gleichgewicht der Population der beiden Klassen. Es gibt jedoch fundamentale Unterschiede zwischen den verschiedenen Ansätzen und jeder bringt seine eigenen Annahmen und Einschränkungen mit sich. Einerseits fügt Oversampling dem Prozess keine neuen Informationen hinzu und kann auch vorhandenes Rauschen verstärken, indem verrauschte Instanzen dupliziert werden. Auf der anderen Seite kann ein Undersampling dazu führen, dass wertvolle Informationen aus dem Datensatz entfernt werden, was die Gesamtleistung des Modells verschlechtern kann. Die semantische Augmentation hängt von der Qualität der vortrainierten Wort-Embeddings ab. Wenn die verwendeten Wort-Embeddings aus einem anderen Korpus als dem verwendeten stammen, ist die Wörterbuchüberschneidung zwischen den Wort-Embeddings und dem Korpus begrenzt. Außerdem können polyseme Wörter den Kontext eines Satzes völlig verändern; zum Beispiel kann "I like apple products", das sich auf die berühmte Firma bezieht, in "I like vegetable products" umgewandelt werden. Korruption kann auch die Stimmung eines Satzes verändern; z. B. kann "I support banning smoking in public areas" in "I support smoking in public areas" umgewandelt werden. Schließlich besteht ein gemeinsamer Fallstrick bei allen Augmentierungsmethoden darin, dass durch die Augmentierung von bereits verrauschten und/oder verzerrten Instanzen die Verstärkung von Rauschen und Fehlern unvermeidlich ist und somit die Gesamtdatenqualität verschlechtert wird.

Einblicke: Das Ziel des Augmentierungsprozesses ist es, mehr Trainingsdaten aus den vorhandenen Trainingsdaten zu erzeugen, indem Variation durch domänenspezifische und fundierte Transformation hinzugefügt wird. In beiden Fällen war es unsere Absicht, die Klassenlabels zu erhalten und gleichzeitig die Tweets plausibel zu machen; natürlich ist dies, wie bereits im Text besprochen, nicht garantiert und daher kann die Augmentation eine weitere Verschlechterung der Datenqualität verursachen. Basierend auf den Experimenten haben wir festgestellt, dass die Augmentierungsverfahren das Problem des Klasse-

nungleichgewichts angehen und dabei eine sehr hohe Leistung beibehalten. Im Vergleich zu den ursprünglichen Self-Learning- und Co-Training-Methoden wurde ein hochsignifikanter Unterschied erzielt, wenn diese mit Augmentierungsmethoden ausgestattet wurden. Obwohl Augmentierungsmethoden essentiell sind, um effektiv mit Klassenungleichgewicht umzugehen, können sie bestehende Fehler verstärken, wenn sie nicht mit Bedacht eingesetzt werden. Wenn sie mit halbüberwachten Methoden gekoppelt werden, werden diese Fehler auf die Gesamtvorhersagen übertragen, wodurch die Kombination solcher Methoden nicht trivial ist. Es sollte ein Filtermechanismus untersucht werden, um Pseudo-Instanzen herauszufiltern, die den ursprünglichen Instanzen semantisch nicht ähnlich sind (für die Aufgabe der Sentiment-Klassifikation).

3 Abschwächen von KI-Diskriminierung

Direkte und indirekte Diskriminierung ist durch internationale Gesetze verboten [DI98], was es zum Gebot der Stunde macht, unfaire Ergebnisse des maschinellen Lernens abzuschwächen. Es gibt eine Vielzahl von Gründen, die dazu führen, dass Algorithmen für maschinelles Lernen diskriminierend werden [Nt20], z. B. können Daten gesellschaftliche Vorurteile kodieren, Daten können Rückkopplungsschleifen enthalten, Daten können unterschiedliche Datenverteilungen für verschiedene Segmente enthalten und so weiter und so fort.

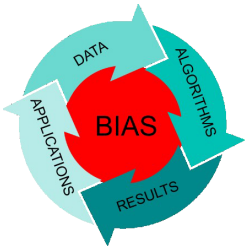


Abb. 3: Bias Loop

In dieser Arbeit entschärfen wir unfaires Verhalten von überwachten Machine-Learning-Modellen, indem wir Eingriffe vor, während und nach der Verarbeitung vornehmen. Insbesondere untersuchen wir, wie Klassen- und klasseninterne Ungleichgewichte die Entscheidungen eines Modells beeinflussen. Wir schlagen ein Fairness-Aware Ensemble-Framework vor, genannt FAE [IFN19], das Klassen- und klasseninternes Ungleichgewicht bekämpft, um unfaire Ergebnisse abzuschwächen. FAE führt zwei fairnessbewusste Eingriffe durch: i) den Vorverarbeitungsschritt, bei dem die Daten partitioniert und neu ausbalanciert werden, um das Ungleichgewicht zwischen den Klassen und innerhalb der Klassen zu mildern, und ii) den Nachverarbeitungsschritt, bei dem die Entscheidungsgrenze des Ensembles verschoben wird, um unfaire Ergebnisse abzuschwächen. Unsere Experimente zeigen, dass Modelle, die auf Daten trainiert werden, die disproportionale Verteilungen für jedes Segment enthalten, im Gegensatz zu unserem Ansatz sehr diskriminierend sind.

Außerdem untersuchen wir Fairness in sequentiellen Modellen wie AdaBoost [Sc99]. Wir schlagen den Begriff *cumulative fairness* vor, der von sequentiellen Modellen verwendet wird, um unfaire Ergebnisse zu mildern, indem die Datenverteilungen im Hinblick auf die Fairness modifiziert werden. Darüber hinaus optimiert unsere Methode die ausgeglichene Fehlerrate, indem sie eine Sequenz von Modellen auswählt, die eine Verlustfunktion minimiert, die die ausgeglichene Fehlerrate und das diskriminierende Verhalten kombiniert. Der kumulative Fairness-Begriff bewertet das Fairness-Verhalten eines sequenziellen Modells vom Anfang bis zur aktuellen Runde. Wir zeigen, dass das induzierte Modell in der Lage ist, faire Ergebnisse zu produzieren, indem es falsch klassifizierten Instanzen fairnessbezogene Gewichte zuweist, die auf unserem kumulativen

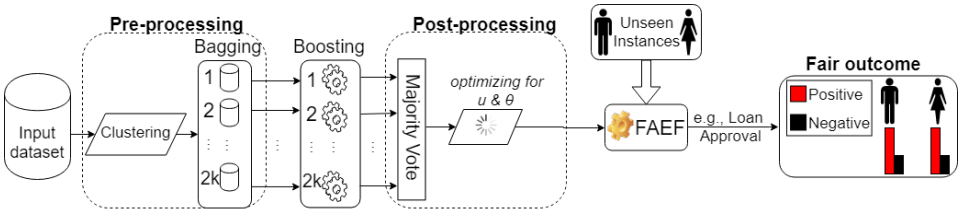


Abb. 4: Fairness-Aware Ensemble Architektur

Fairness-Begriff basieren. Das von uns vorgeschlagene Modell mit dem Namen AdaFair [IN19a] ist in der Lage, Diskriminierung abzuschwächen und das Problem des Klassenungleichgewichts zu lösen, und übertrifft die neuesten fairnessbewussten Ansätze.

3.1 Fairness-Aware Ensemble framework

Diskriminierung ist oft ein Artefakt komplexer Interaktionen zwischen großen, komplexen Daten und Algorithmen, und daher ist ein ganzheitlicherer Ansatz erforderlich. Wie wir in diesem Abschnitt zeigen, haben sowohl klassen- als auch klasseninterne (seltene Fälle) Ungleichgewichte einen Einfluss auf das diskriminierende Verhalten eines Klassifikators. Zu diesem Zweck schlagen wir ein FAE-Framework (Fairness-Aware Ensemble[4]) vor, das Fairness-bezogene Eingriffe sowohl im Pre- als auch im Post-Processing-Schritt kombiniert. Im Pre-Processing-Schritt gehen wir das Problem der Unterrepräsentation der geschützten Gruppe an, das im Folgenden als Gruppenungleichgewicht bezeichnet wird, sowie das Problem des Klassenungleichgewichts, wobei die Zielklasse die Minderheitenklasse ist. Im Nachbearbeitungsschritt gehen wir das Problem der Klassenüberschneidung im Merkmalsraum an, indem wir die Entscheidungsgrenze in Richtung der Fairness verändern.

Abbildung 4 zeigt eine Übersicht von FAE, vom Training (linke Seite) bis zur Vorhersage neuer Instanzen (rechte Seite). FAE kombiniert fairnessbezogene Vor- und Nachbearbeitungseingriffe wie folgt: i) **Fairness-aware ensemble lernen**: in *pre-processing* gehen wir die Probleme der Gruppen- und Klassen-Unausgewogenheit an. Insbesondere verwenden wir *bagging*, um die Gruppen in jedem Beutel auszugleichen, indem wir die geschützte positive Gruppe und eine repräsentative Stichprobe aus den anderen Gruppen berücksichtigen. Danach wird *boosting* [Sc99] auf jeden Beutel angewendet, so dass am Ende ein Ensemble von Ensembles gelernt wird. ii) **Fairness-aware Verschiebung der Entscheidungsgrenze**: Im *post-processing* verschieben wir die Entscheidungsgrenze des Lerners in Richtung Fairness basierend auf einem einstellbaren Parameter θ, bis die Diskriminationsbewertung den benutzerdefinierten Schwellenwert ε erfüllt.

Einblicke: Unsere Experimente zeigen, dass durch die Berücksichtigung von Klassenungleichheit und Gruppenungleichheit die diskriminierenden Ergebnisse des Modells deutlich reduziert werden. Allerdings diskriminiert das Modell immer noch aufgrund gesellschaftlicher Verzerrungen, die in den Daten kodiert sind. Daher verschieben wir die Entscheidungsgrenze und wählen zusätzlich Hypothesen aus den Ensemble-Lernern für nahezu

[4] `https://github.com/iosifidisvasileios/Fairness-Aware-Ensemble-Framework`

ideale Diskriminierungsergebnisse aus. Solche Schritte stellen sicher, dass eine Reduzierung der Diskriminationswerte nicht auf Kosten der Fähigkeit des Modells geht, Instanzen korrekt in die entsprechenden Klassen zu klassifizieren. FAE kann eine maximale Klassifizierungsleistung erreichen und wichtige Faktoren wie die Diskriminierung für eine bestimmte Metrik berücksichtigen.

3.2 Adaptive Fairness-aware Boosting

In diesem Abschnitt schlagen wir AdaFair [IN19a] vor, einen fairness-bewussten Klassifikator, der auf AdaBoost [Sc99] basiert. AdaFair[5] (Abbildung 5) aktualisiert die Gewichte der Instanzen in jeder Boosting-Runde unter Berücksichtigung einer kumulativen Vorstellung von Fairness, die auf allen aktuellen Ensemble-Mitgliedern basiert, und geht dabei explizit gegen Klassenungleichheit vor, indem die Anzahl der Ensemble-Mitglieder für einen ausgeglichenen Klassifikationsfehler optimiert wird.

Wir passen AdaBoost für Fairness an, indem wir seinen Neugewichtungsprozess anpassen. Insbesondere: i) berücksichtigen wir das Fairnessverhalten des Modells direkt im Gewichtungsprozess, indem wir den Begriff des kumulativen Fairnessverhaltens des Modells bis zur aktuellen Boosting-Runde einführen. Darüber hinaus verwenden wir, anders als Vanilla AdaBoost, ii) Konfidenzwerte im Neugewichtungsprozess, um eine Differenzierung der Instanzgewichtung zu ermöglichen, die darauf basiert, wie sicher das Modell bezüglich ihrer Klasse ist.

Schließlich optimieren wir die Anzahl der schwachen Lerner im endgültigen Ensemble, indem wir die ausgeglichene Fehlerrate berücksichtigen und somit das Klassenungleichgewicht direkt in die Auswahl des besten Modells einbeziehen.

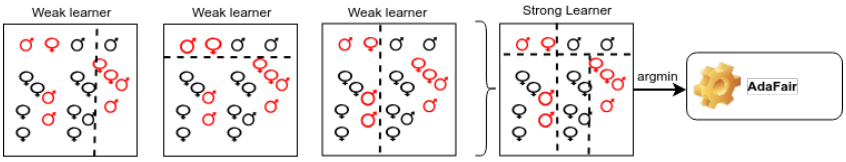


Abb. 5: AdaFair Methode

Einblicke: Unsere Experimente an vier realen Datensätzen zeigen einen erheblichen Unterschied in der Wiedererkennung im Vergleich zu aktuellen fairnessbewussten State-of-the-Art-Ansätzen. Wir haben gesehen, dass AdaFair in Fällen schwerer und extremer Klassenungleichheit Fairness zwischen geschützten und nicht-geschützten Gruppen erreichen und gleichzeitig eine deutlich bessere Klassifikationsleistung im Vergleich zu anderen Ansätzen beibehalten kann. Außerdem ist AdaFair bei extremer Klassenungleichheit in der Lage, Methoden zu übertreffen, die sich nur auf die Klassenungleichheit konzentrieren. In Fällen von klassenbalancierten Datensätzen erreicht AdaFair eine ähnliche Fairness wie andere Fairness-bewusste Methoden.

[5] `https://github.com/iosifidisvasileios/AdaFair`

4 Abschluss

In dieser Arbeit haben wir uns auf Klassenungleichheit und ihre Auswirkungen auf zwei populäre Bereiche konzentriert: i) halbüberwachtes Lernen, bei dem Label-Knappheit besteht, und ii) Fairness-bewusstes Lernen, bei dem Bevölkerungssegmente ungleich behandelt werden. In diesem Abschnitt fassen wir unsere wichtigsten Ergebnisse zusammen und diskutieren zukünftige Richtungen für jedes Kapitel.

Wir haben halbüberwachte Methoden wie Co-Training und Selbstlernen unter dem Prisma der Klassenungleichheit untersucht. Wir haben gezeigt, dass halb-überwachte Methoden Fehler und Klassenungleichheit in jeder Iteration propagieren. Um mit Klassenungleichheit umzugehen, haben wir halbüberwachten Methoden mit verschiedenen Augmentierungsmethoden gekoppelt, wie z.B.: Over-Sampling, Under-Sampling, Verzerrung und semantische Ähnlichkeit (über Wort-Embeddings). Durch den Einsatz von Augmentierungsmethoden haben wir mehr Trainingsdaten erzeugt und auch mehr Variation durch domänenspezifische und klangliche Transformationen hinzugefügt. Unsere Experimente zeigen, dass eine solche Kombination von Methoden (halbüberwacht mit Augmentierungsmethoden) die Ausbreitung der Klassenungleichheit in den Griff bekommt. Darüber hinaus haben wir große Mengen an unbeschrifteten Textdaten annotiert.

Wir haben das Problem der unfairen Ergebnisse in überwachten Lernmodellen durch das Prisma der Klassen- und klasseninternen Unausgewogenheit untersucht. Wir haben gezeigt, dass schiefe Datenverteilungen die überwachten Modelle in Bezug auf das Unterscheidungsverhalten beeinflussen. Wir haben gesehen, dass eine Unausgewogenheit innerhalb einer Klasse (Gruppen-Unausgewogenheit) die Modelle dazu zwingt, Minderheitensegmente im Vergleich zu anderen Segmenten überproportional falsch zu klassifizieren, da sie nicht in der Lage sind, alle Bevölkerungssegmente effektiv zu lernen. Darüber hinaus haben wir faires maschinelles Lernen in sequenziellen Modellen untersucht und den Begriff der kumulativen Fairness eingeführt. Kumulative Fairness zwingt das Modell dazu, unfaire Ergebnisse über die Iterationen hinweg abzuschwächen. Daher verhält sich das Modell über die Iterationen hinweg fair.

Zusammenfassend lässt sich sagen, dass wir die Auswirkungen von Klassenungleichheit in zwei verschiedenen Bereichen untersucht haben, z. B. beim halbüberwachten Lernen und beim fairen maschinellen Lernen; Klassenungleichheit betrifft jedoch verschiedene Bereiche des maschinellen Lernens. Obwohl Standard-Klassenungleichheitstechniken leistungsstarke Methoden sind, um mit dem eigenständigen Klassenungleichheitsproblem umzugehen (z. B. Klassifizierungsleistung), können sie ineffektiv werden, wenn sie mit einem gemeinsamen Problem konfrontiert werden. Daher sind domänenspezifische oder heuristische Ansätze erforderlich, um kombinierte Probleme zu bewältigen.

Literaturverzeichnis

[BM98] Blum, Avrim; Mitchell, Tom: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. ACM, pp. 92–100, 1998.

[DI98] DIRECTIVE, HAS ADOPTED THIS: COUNCIL DIRECTIVE 97/80/EC of 15 December 1997 on the burden of proof in cases of discrimination based on sex. Official Journal L, 14(20/01):0006–0008, 1998.

[Fr67] Fralick, S: Learning to recognize patterns without a teacher. IEEE Transactions on Information Theory, 13(1):57–64, January 1967.

[HM13] He, Haibo; Ma, Yunqian: Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, 2013.

[IFN19] Iosifidis, Vasileios; Fetahu, Besnik; Ntoutsi, Eirini: FAE: A Fairness-Aware Ensemble Framework. In: IEEE International Conference on Big Data. pp. 1375–1380, 2019.

[IN17] Iosifidis, Vasileios; Ntoutsi, Eirini: Large scale sentiment learning with limited labels. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp. 1823–1832, 2017.

[IN19a] Iosifidis, Vasileios; Ntoutsi, Eirini: AdaFair: Cumulative Fairness Adaptive Boosting. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 781–790, 2019.

[IN19b] Iosifidis, Vasileios; Ntoutsi, Eirini: Sentiment analysis on big sparse data streams with limited labels. Knowledge and Information Systems, pp. 1–40, 2019.

[Io20] Iosifidis, Vasileios: Semi-supervised learning and fairness-aware learning under class imbalance. PhD thesis, Hannover: Institutionelles Repositorium der Leibniz Universität Hannover, 2020.

[Mi13] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S; Dean, Jeff: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119, 2013.

[Nt20] Ntoutsi, Eirini; Fafalios, Pavlos; Gadiraju, Ujwal; Iosifidis, Vasileios; Nejdl, Wolfgang; Vidal, Maria-Esther; Ruggieri, Salvatore; Turini, Franco; Papadopoulos, Symeon; Krasanakis, Emmanouil et al.: Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. e1356, 2020.

[Sc99] Schapire, Robert E: A brief introduction to boosting. In: IJCAI. volume 99, pp. 1401–1406, 1999.

Vasileios Iosifidis ist Post-Doc am L3S Forschungszentrum an der Leibniz Universität Hannover (LUH). Er erhielt sein Diplom (2014) in Computer Engineering, seinen Master (2016) in Software Engineering von der Universität Patras, Griechenland und seinen PhD (2020) in Maschinellem Lernen von der Leibniz Universität Hannover. Seine Forschung liegt in den Bereichen Data Mining und Maschinelles Lernen, wo er Methoden für das Lernen über komplexe Daten und Datenströmen sowie Methoden für faires maschinelles Lernen entwickelt. Vor seiner Tätigkeit an der LUH war er als wissenschaftlicher Mitarbeiter am Institut für Computertechnik und beim Institut *Diophantus* tätig.

Verbesserung der Qualität von automobilen Testfallspezifikationen[1]

Katharina Juhnke[2]

Abstract: Üblicherweise werden automobile Testfallspezifikationen für das Testen kundenerlebbarer Funktionen in Fahrzeugprototypen überwiegend in natürlicher Sprache von mehreren Testdesignern geschrieben und von verschiedenen Testern ausgeführt. Dadurch wirken sich Qualitätsmängel, wie z.B. mehrdeutige, unvollständige oder inkonsistente Testfälle, negativ auf den Kosten- und Zeitaufwand des Testens aus. Um die Qualität von Testfallspezifikationen zu verbessern, werden in dieser Dissertation zunächst die Qualität beeinflussende *Herausforderungen* identifiziert. Darauf aufbauend wird zum einen ein *Qualitätsmodell* vorgestellt, das die Grundlage für die entwickelten perspektivenbasierten *Review-Checklisten* als analytische Qualitätssicherungsmaßnahme bildet. Zum anderen unterstützt die entwickelte *Testfallspezifikations-orientierte Domänenanalysemethode* die Ableitung von systemspezifischen Schablonen auf Basis bestehender Testfallspezifikationen und ermöglicht damit die Definition von *Testing DSLs* als konstruktive Qualitätssicherungsmaßnahme. Des Weiteren wurde erstmalig anhand der Ergebnisse eines kontrollierten Experiments nachgewiesen, dass durch die Verwendung von Testing DSLs die Qualität von Testfällen signifikant verbessert wird.

1 Einführung

Innovationen in Fahrzeugen entstehen zunehmend durch neue Funktionen und Technologien, die durch Software und elektronische Systeme realisiert werden. Neuartige, aber auch etablierte Funktionen müssen ausreichend getestet werden, um Ausfälle zu vermeiden. Andernfalls kann dies nicht nur zu unzufriedenen Kunden führen, sondern sogar deren Leben gefährden. Daher ist ein solider Testprozess einer der wichtigsten Bestandteile bei der Entwicklung elektrischer und elektronischer Systeme im Automobilbereich. Normen wie ISO 26262 [In11] oder Automotive SPICE [VD17] definieren die erforderlichen Aktivitäten und Arbeitsprodukte für das Testen sowie die anzuwendenden Testmethoden. Ein wesentliches Arbeitsprodukt für die Entwicklung einer Funktion oder eines Systems ist dabei die *Testfallspezifikation*.

Eine Testfallspezifikation enthält die Menge relevanter, aus einer *Testbasis* abgeleiteter Testfälle für ein bestimmtes *Testobjekt* [In19]. Beispielsweise ist PRE-SAFE[3] ein System (*Testobjekt*), das im Gefahrenfall Schutz bietet. Um die Systemanforderungen für dieses System anhand eines Fahrzeugprototyps zu validieren, spezifiziert ein Testdesigner eine Reihe von Testfällen. Diese Testfälle enthalten detaillierte Beschreibungen der von einem Testfahrer auszuführenden Fahrmanöver und die dazugehörigen Reaktionsbeschreibungen (z.B. Aktivierung der Gurtstraffung oder automatische Schließfunktion für Seitenfenster), die aus der Systemanforderungsspezifikation (*Testbasis*) abgeleitet wurden.

[1] Englischer Titel der Dissertation: „Improving the Quality of Automotive Test Case Specifications“ [Ju21]

[2] Institut für Softwaretechnik und Programmiersprachen, Universität Ulm, katharina.juhnke@uni-ulm.de

[3] Mercedes-Benz PRE-SAFE®: `https://www.youtube.com/watch?v=vTmLYY-Z2rc`

Eine qualitativ hochwertige Testfallspezifikation ist notwendig, um Fehler oder Fehlinterpretationen zu vermeiden. Dies ist besonders wichtig, wenn die Autoren (*Testdesigner*) einer Testfallspezifikation (z.B. Systemverantwortliche oder externe Dienstleister) nicht die letztendlichen *Tester* (z.B. HiL-Tester, Testfahrer, Produktionsmitarbeiter) sind, welche die spezifizierten Testfälle ausführen. Dies ist der Regelfall im Automobilumfeld, zumal eine Testfallspezifikation verschiedene Anwendungsbereiche und Teststufen abdeckt. So dienen die Testfälle beispielsweise als Grundlage für die Implementierung von Testskripten für Hardware-in-the-Loop (HiL) Tests, für die Durchführung von manuellen Tests in einem Fahrzeugprototypen oder als Abnahmetests am Ende der Produktion. Dementsprechend stellt eine qualitativ hochwertige Testfallspezifikation sicher, dass die Tester die Testfälle genau so verstehen, implementieren und ausführen, wie es der Testdesigner beabsichtigt hat. Dies kann eine Herausforderung sein, da alle Beteiligten in der Regel unterschiedliche Kenntnisse oder Annahmen über das Testobjekt und unterschiedliche Erfahrungen mit Testtechniken, Testprozessen, Testskriptsprachen oder formalen Notationen im Allgemeinen haben.

Eine zusätzliche Herausforderung besteht darin, dass die in dieser Dissertation betrachteten Testfälle meist in natürlicher Sprache spezifiziert sind. Dies gilt insbesondere für (kundenerlebbare) Akzeptanztestfälle im Automobilumfeld, die manuell von einem menschlichen Tester in einem Fahrzeugprototypen ausgeführt werden. Aber auch *logische Testfälle*, die als Grundlage für die Implementierung von automatisch ausführbaren Testskripten (*konkrete Testfälle*) dienen, werden in natürlicher Sprache spezifiziert. Dies ist darauf zurückzuführen, dass Testdesigner nicht immer über ein umfassendes Wissen über verschiedene Testtechnologien (z.B. HiL-Tests) verfügen, da dies in der Regel nicht in ihrem Verantwortungsbereich liegt. Daher sind sie oft nicht qualifiziert, typische Testskriptsprachen zu verwenden, weshalb sie Testfallabläufe in natürlicher Sprache beschreiben und es dann die Aufgabe dedizierter Tester ist, für bestimmte Testtechnologien Testskripte zu entwickeln.

1.1 Problemstellung

Wie bereits von natürlichsprachlichen Anforderungen bekannt [DBK03, Ba15, Fe17], haben auch natürlichsprachliche Testfälle Probleme mit Mehrdeutigkeit, Inkonsistenz, Verständlichkeit und Unvollständigkeit [Ha13]. Dementsprechend beeinflussen solche Probleme die Qualität von Testfallspezifikationen, was wiederum zu einem Mehraufwand für Testdesigner und Tester führt. Im schlimmsten Fall können unentdeckte Mängel in Testfallspezifikationen ein Risiko für den Endverbraucher darstellen, wenn z.B. eine Fahrzeugfunktion unzureichend getestet wurde und dies zu einem Unfall führt.

Tatsächlich berichten Praktiker aus der Automobilindustrie, mit denen im Rahmen dieser Dissertation ein Austausch stattfand, dass fehlerhafte Testfälle existieren und die Qualität der Testfallspezifikationen schlecht ist [JTH18c, JTH20]. Dies äußert sich z.B. durch einen hohen Kommunikationsaufwand aufgrund von Rückfragen der Tester bei Mehrdeutigkeiten in Testfallspezifikationen oder durch fehlerhaft implementierte Testfälle. Eine fehlerhafte Testfallspezifikation (z.B. inkonsistente Testfälle, fehlende Testfälle für bestimmte Testplattformen, Zuordnung von Testfällen zur falschen Testplattform) führt dazu, dass

das Testen zu viel Zeit in Anspruch nimmt (z.B. bis Unklarheiten geklärt sind), zu teuer ist (z.B. durch redundante Testfälle) oder in manchen Fällen das Testen keine Wirkung hat und keine Fehler entdeckt werden (z.B. wenn Testfälle falsch implementiert sind oder fehlen). Um diese Folgen zu vermeiden, ist das Ziel dieser Dissertation einen Beitrag zur Verbesserung der Qualität von automobilen Testfallspezifikationen zu leisten. Dies erfolgt am Beispiel des Automobilherstellers Daimler und einigen seiner Entwicklungsdienstleister, die auch für andere Automobilhersteller tätig sind.

1.2 Forschungsmethode, Forschungsfragen und Beiträge der Dissertation

Dieser Dissertation [Ju21] liegt die Forschungsmethodik Design Science Research (DSR) zugrunde, da die vorgestellte Forschung in Kooperation mit dem Automobilkonzern Daimler durchgeführt wurde. Daher war es von besonderem Interesse einerseits am Ende des Forschungsprozesses Artefakte zu erhalten, die einen echten Nutzen haben und eine effektive Lösung für die identifizierten Probleme der realen Welt darstellen. Andererseits sollte hierfür gleichermaßen ein stringenter und wissenschaftlich fundierter Forschungsprozess zur Anwendung kommen. Abbildung 1 zeigt die einzelnen Aktivitäten dieses in der Dissertation angewandten Forschungsprozesses zusammen mit den entwickelten, demonstrierten und evaluierten Lösungsartefakten, welche die Beiträge dieser Dissertation darstellen.

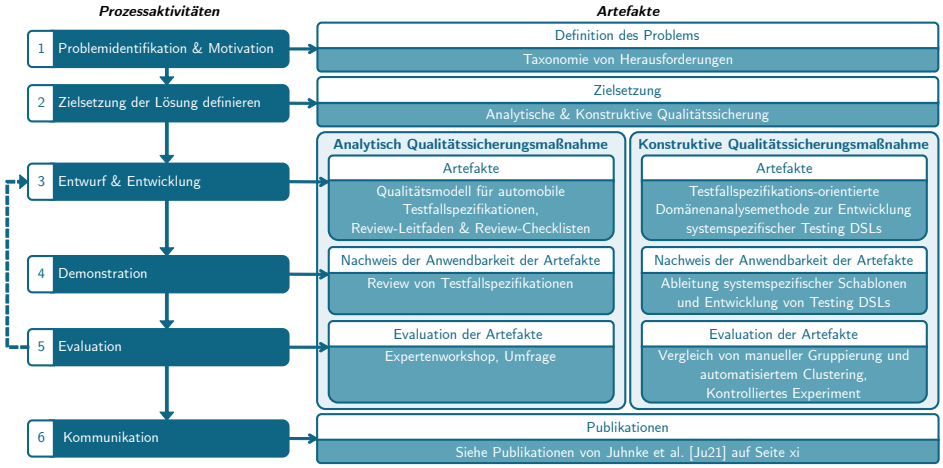


Abb. 1: Design Science Research Prozessmodell nach Peffers et al. [Pe07] mit definierten Ergebnisartefakten für die in der Dissertation durchgeführten Forschungsaktivitäten

Um die Qualität von automobilen Testfallspezifikationen zu verbessern, werden zunächst Herausforderungen untersucht, die aus Sicht von Praktikern die Qualität beeinträchtigen. Dazu gehört auch die Untersuchung, welche Ursachen und Folgen diese Herausforderungen haben und wie relevant sie für Praktiker tatsächlich sind. Diesen Aspekten widmet sich die erste Hauptforschungsfrage:

RQ1 *Über welche Herausforderungen, die die Qualität von automobilen Testfallspezifikationen beeinflussen, sind sich Praktiker bewusst?*

Das Ergebnis dieser Untersuchung ist eine *Taxonomie von Herausforderungen bei der Erstellung und Weiterverarbeitung von Testfallspezifikationen* (siehe [JTH18a, JTH18b,

JTH20]). Die Taxonomie ergänzt verwandte Arbeiten [Gr03, SPL11, KPM13], die zum Beispiel auf allgemeine Herausforderungen im automobilen Testprozess, aber nicht explizit auf Testfallspezifikationen eingehen. Darüber hinaus liefert die Taxonomie einen wichtigen Beitrag für die Definition von Qualität und die Entwicklung möglicher Ansätze zur Verbesserung der Qualität von Testfallspezifikationen, wie sie konkret in den beiden folgenden Hauptforschungsfragen adressiert werden.

Die zweite Hauptforschungsfrage fokussiert die Entwicklung einer analytischen Qualitätssicherungsmethode, die notwendig ist, um die Qualität von automobilen Testfallspezifikationen zu beurteilen:

RQ2 *Wie kann die Qualität von Testfallspezifikationen beurteilt werden?*

Das Ergebnis ist ein *Qualitätsmodell für automobile Testfallspezifikationen*. Die enthaltenen Qualitätskriterien werden insbesondere aus der Taxonomie der Herausforderungen, verwandten Arbeiten, dem Standard für Software Testing ISO 29119 [In13] und aus von Praktikern hervorgehobenen Kriterien abgeleitet, die ihrer Meinung nach zu einer qualitativ hochwertigen Testfallspezifikation beitragen. Basierend auf dem Qualitätsmodell werden ein *Reviewleitfaden* und *Review-Checklisten* entwickelt. Der Reviewleitfaden beschreibt die Durchführung eines mehrdimensionalen Reviews, das verschiedene Perspektiven auf die Qualität einer Testfallspezifikation berücksichtigt. So können beispielsweise die Sichtweisen von Testdesignern und Testern berücksichtigt werden, die unterschiedliche Anforderungen an eine qualitative Testfallspezifikation haben. Die Review-Checklisten unterstützen Inspektoren beim Verständnis von Qualität und der Durchführung von multidimensionalen Testfallspezifikationsreviews. Um jedoch dazu ergänzend die Entstehung von Fehlern in Testfallspezifikationen proaktiv zu verhindern oder zumindest zu reduzieren, ist eine konstruktive Qualitätssicherungsmaßnahme notwendig. Daher adressiert die dritte Hauptforschungsfrage die Entwicklung eines schablonenbasierten Ansatzes mittels domänenspezifischer Sprachen (DSLs) für die systemspezifische Spezifikation von automobilen Testfällen als konstruktive Qualitätssicherungsmaßnahme:

RQ3 *Wie kann die Entwicklung von systemspezifischen Testing DSLs unterstützt werden und wie können sie die Qualität von Testfallspezifikationen verbessern?*

Auch in der Automobilindustrie werden typische Elemente eines Testfalls [In13] (z.B. ID, Testfallname, Vor- und Nachbedingung, Testschritte bestehend aus Aktionen und Reaktionen, etc.) beispielsweise in Form von MS Excel-Vorlagen strukturiert erfasst. Der entwickelte schablonenbasierte Ansatz und dessen Umsetzung mittels systemspezifischer Testing DSLs geht dabei jedoch einen Schritt weiter, indem er sich auf die Inhalte der natürlichsprachigen Aktions- und Reaktionsbeschreibungen stützt, diese formalisiert und gleichzeitig systemspezifische Eigenschaften berücksichtigt (z.B. Fachterminologie, Parameter- und Signalnamen, etc.). Um für die über 100 Systeme in einem Premiumfahrzeug geeignete systemspezifische Testing DSLs bereitzustellen wird eine eigens entwickelte *Testfallspezifikations-orientierte Domänenanalysemethode* vorgestellt, die Praktiker in die Lage versetzt, auf Basis bestehender Testfallspezifikationen individuelle und systemspezifische Schablonen für die Spezifikation ihrer Testfälle abzuleiten. Diese Schablonen sind für Praktiker mit unterschiedlichen Qualifikationen einfach zu handhaben und leicht zu erlernen, da sie typische Formulierungen und Fachbegriffe aus der jeweiligen Fahrzeugdomäne enthalten. Außerdem sind sie flexibel an

individuelle Bedürfnisse (z.B. Testplattform, Testobjekt, projektspezifische Anforderungen) angepasst. Mittels einer Kombination aus unüberwachten maschinellen Lernverfahren in Form eines automatisierten Clusterings von Testfallbeschreibungen und der Ableitung von Testing DSLs mittels eines multiplen Sequenzalignment (MSA) Ansatzes aus dem Bereich der Bioinformatik sowie einer entwickelten Heuristik wird gezeigt, wie Testing DSLs aus realen automobilen Testfallspezifikationen semi-automatisiert generiert werden. Darüber hinaus zeigt der experimentelle Vergleich des Testing DSL Ansatzes mit dem konventionellen natürlichsprachlichen Ansatz, dass Testing DSLs Fehler in Testfällen (z.B. unvollständige und fehlerhafte Beschreibungen, strukturelle Mängel, Mehrdeutigkeiten, Schreib- und Tippfehler) signifikant reduzieren und dass auch Praktiker von diesen Vorteilen überzeugt sind.

Im Folgenden wird auf die entwickelte Testfallspezifikations-orientierten Domänenanalysemethode, deren Automatisierung und die Evaluation dieser Methode sowie des Testing DSL Ansatzes für die Testfallspezifikation näher eingegangen.

2 Systemspezifische Testing DSLs zur Spezifikation von Testfällen

Um von Praktikern akzeptiert zu werden, müssen sich systemspezifische Schablonen und eine daraus abgeleitete Testing DSL an der üblicherweise von Domänenexperten verwendeten Grammatik und dem etablierten Vokabular, beispielsweise aus früheren Testfallspezifikationen, orientieren. Dazu zählt auch, dass Testfallbeschreibungen je nach Funktion oder System, beabsichtigter Zieltestplattform oder gar je nach verantwortlicher Fachabteilung spezifische Eigenschaften aufweisen und sich teils stark voneinander unterschieden. Eine einzige Testing DSL für die über 100 Systeme in einem Premiumfahrzeug bereitzustellen würde zu einem immensen und unüberschaubaren Sprachumfang führen. Die Notwendigkeit eine neue Sprache zu erlernen gepaart mit einer übermäßigen Fülle oder dem Fehlen notwendiger Ausdrucksmittel kann jedoch die Benutzerakzeptanz verringern.

2.1 Testfallspezifikations-orientierte Domänenanalysemethode

Um eine hohe Benutzerakzeptanz zu gewährleisten wurde basierend auf den grundlegenden Schritten der Domänenanalyse nach Prieto-Díaz [PD90] die Testfallspezifikations-orientierte Domänenanalysemethode als Mechanismus zur Extraktion, Abstraktion und Klassifizierung von Informationen aus bestehenden Testfallspezifikationen entwickelt. Kern dieser Methode ist das Extrahieren von Phrasen aus Aktions- und Reaktionsbeschreibungen existierender Testfallspezifikationen und deren Zerlegung in *elementare Phrasen*. Dabei handelt es sich um einen Teil der Aktions- oder Reaktionsbeschreibung, der auch in anderen Beschreibungen vorkommen kann (d.h. eine Teilaktion, die z.B. durch „und“, Komma oder „–“ separiert ist). Anschließend werden ähnliche elementare Phrasen gruppiert. Aus diesen Gruppen lassen sich unter Einbeziehung von Domänenanalysten letztlich *konzeptionelle Templates* ableiten, wie in Abbildung 2 dargestellt.

Konzeptionelle Templates bieten eine übersichtliche Darstellung der in bestehenden Testfallbeschreibungen verwendeten Strukturen und dienen als Unterstützung für die Kommunikation mit Domänenexperten. Es werden statische und variable Teile einer Phrase

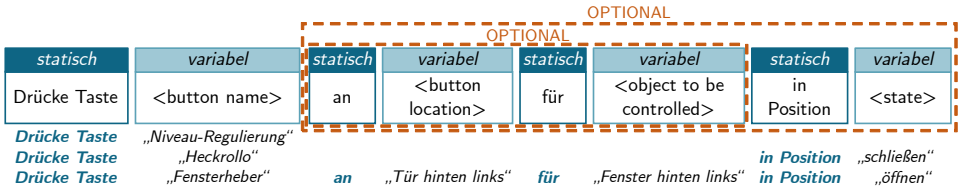


Abb. 2: Beispiel für ein konzeptionelles Template [JT19]

definiert. Die statischen Teile entsprechen dem Vokabular der konzeptionellen Templates. Die variablen Teile beschreiben systemspezifische Begriffe, Parameter, Werte oder Steuerungskonzepte. Weiterhin werden optionale Teile eines konzeptionellen Templates und Positionen bzw. Zusammenhänge zwischen den einzelnen Teilen definiert, was anhand der grafischen Darstellung in Abbildung 2 beispielhaft gezeigt wird.

Die Anwendbarkeit der entwickelten Domänenanalysemethode wurde anhand von fünf realen automobilen Testfallspezifikationen (bestehend aus 66 bis zu 4 458 Testfallbeschreibungen) demonstriert und zeigt, dass große Teile der Testfallspezifikationen (70% bis 95%) durch eine geringe Anzahl von Templates (11 bis 35) beschrieben werden können [JT19]. Damit leistet die Dissertation einen Beitrag zur formalen Domänenanalyse im Kontext der DSL-Entwicklung, die von Kosar et al. [KBM16] als Forschungslücke beschrieben wurde. Darüber hinaus zeigen die Ergebnisse, dass aufgrund der hohen Wiederverwendung von Phrasen in den betrachteten automobilen Testfallspezifikationen ein deutliches Potenzial für den Einsatz von derartigen Templates besteht.

2.2 Automatisierung der Entwicklung systemspezifischer Testing DSLs

Erfahrungen in einer Testabteilung bei Mercedes-Benz Cars Development haben gezeigt, dass ein Domänenanalytiker ca. zwei Tage benötigt, um aus einer kleineren Testfallspezifikation mit ca. 250 Testschritten konzeptionelle Templates für Aktions- und Reaktionsbeschreibungen *manuell* abzuleiten. Sehr große Spezifikationen können bis zu 16 000 Testfälle enthalten, was einen immensen Zeitaufwand bedeutet. Daher ist es sinnvoll und notwendig, die Aktivitäten bezüglich der Gruppierung ähnlicher elementarer Phrasen und der Ableitung von konzeptionellen Templates bzw. Testing DSLs zu automatisieren.

Ein zusätzlicher Beitrag der Dissertation ist daher die Automatisierung dieser Domänenanalyseaktivitäten auf Basis bestehender und bereits manuell qualitätsgeprüfter Testspezifikationen. Dieser erstmals vorgestellte semi-automatisierte Ansatz erlaubt es, die in Aktions- und Reaktionsbeschreibungen verwendeten elementaren Phrasen durch Clustering zu gruppieren, um so die Grundlage für die Ableitung konzeptioneller Templates zu schaffen. Hierfür wird der DBSCAN-Algorithmus [Es96] in Kombination mit einem entwickelten Ähnlichkeitsmaß namens *Run Length Similarity* verwendet. Außerdem wird gezeigt, wie der DBSCAN-Algorithmus für die Anwendung auf automobile Testfallspezifikationen geeignet parametrisiert werden kann (z.B. $MinPts = 2$, $Eps = 0\,333$). Eine Besonderheit des Automatisierungsansatzes ist die Anwendung auf unbekannte und unmarkierte Daten, d.h., es sind keine Kenntnisse über die zu analysierenden Testfallspezifikationen oder eine spezielle Vorverarbeitung erforderlich, die z.B. domänenspezifische

Wörterbücher oder eine Voranalyse durch Domänenexperten erfordern würde. Außerdem werden durch die Verwendung des DBSCAN-Algorithmus elementare Phrasen eindeutig einem Cluster zugeordnet und Ausreißer sind erlaubt, so dass eine unpassende Zuordnung zu einem Cluster nicht zwangsläufig auftritt und der automatisierte Ansatz somit die Grundlage für die Ableitung konzeptioneller Templates und somit auch für die Entwicklung von systemspezifischen Testing DSLs darstellt.

Darüber hinaus zeigt die Evaluation des automatisierten Clustering-Ansatz anhand von drei realen automobilen Testfallspezifikationen, dass es eine nahezu perfekte Übereinstimmung zwischen den Ergebnissen des automatisierten Clustering und den zuvor manuell erstellten Gruppen gibt (*Cohen's Kappa* $\kappa > 0{,}81$). Das heißt, es konnte gezeigt werden, dass elementare Phrasen sinnvoll automatisiert gruppiert werden können.

Diese resultierenden Gruppen werden verwendet, um Vorschläge für konzeptionelle Templates automatisch abzuleiten, die als Vorstufe für die Definition einer endgültigen systemspezifischen Testing DSL dienen. In der Dissertation [Ju21] wird demonstriert, wie die geclusterten elementaren Phrasen zur Ableitung konzeptioneller Templates verwendet werden. Hierfür wird das multiple Sequenzalignment (MSA) verwendet, welches einen etablierten Ansatz aus der Bioinformatik zur Analyse von Strukturen in RNA- und DNA-Sequenzen darstellt, und anschließend durch eine Heuristik ergänzt.

Der Vorteil des automatisierten Clustering besteht darin, dass die Ergebnisse personenunabhängig und damit leichter reproduzierbar sind, was bei manuell erstellten Gruppen nicht unbedingt der Fall ist. Insgesamt kann dadurch der manuelle Aufwand für die Analyse bestehender Testfallspezifikationen deutlich reduziert und Domänenanalysten in der Entwicklung Testing DSLs unterstützt werden.

3 Evaluation des systemspezifischen Testing DSL Ansatzes

Die Anwendbarkeit von systemspezifischen Testing DSLs, um Testfälle für ein bestimmtes System oder passend zu den Anforderungen einer bestimmten Zieltestplattform zu spezifizieren, wurde bereits gezeigt (siehe Abschnitt 2.1 und [JT19]). Allerdings gibt es bisher keine empirischen Belege dafür, dass derartige Testfälle tatsächlich weniger Probleme enthalten als natürlichsprachliche Testfälle. Außerdem gibt es keine Belege dafür, dass sich Testdesigner sicherer fühlen, wenn sie Testfälle mit einer systemspezifischen Testing DSL erstellen und dass ihre Selbsteinschätzung der angenommenen Testfallqualität höher ist. Diese Aspekte sind jedoch ein wichtiges Indiz für eine reibungslose Einführung des Konzepts der systemspezifischen Testing DSL in die industrielle Praxis.

Deshalb wurde ein kontrolliertes Experiment mit einem 2 (Ansatz) $\times$ 2 (System) Mixed Design mit Messwiederholung auf dem ersten Faktor mit 20 zufällig ausgewählten Studenten durchgeführt. Die Teilnehmer mussten Testfälle unter Verwendung des natürlichsprachlichen Ansatzes (*NL*) oder des systemspezifischen Testing DSL Ansatzes (*DSL*) für zwei Systeme (*MW* und *CM*) erstellen und dokumentieren. Die erstellten Testfälle wurden anschließend hinsichtlich enthaltener Mängel analysiert und zusätzlich wurden Daten mit Hilfe von Fragebögen erhoben. Basierend darauf wurden unter anderem die folgenden Hypothesen getestet:

H_1 : Die Gesamtanzahl enthaltener Qualitätsprobleme ist für Testfälle, die mit einer systemspezifischen Testing DSL spezifiziert werden, geringer als für Testfälle, die mit natürlicher Sprache spezifiziert werden, d.h.,
$H0_1 : Mean_{NL}(\text{Problemanzahl}) \leq Mean_{DSL}(\text{Problemanzahl})$
$H1_1 : Mean_{NL}(\text{Problemanzahl}) > Mean_{DSL}(\text{Problemanzahl})$.

$H_2 - H_4$: Wenn die Testfälle unter Verwendung einer systemspezifischen Testing DSL spezifiziert werden, schätzen Testdesigner die Konsistenz (H_2), Vollständigkeit (H_3) und Verständlichkeit (H_4) der beschriebenen Testfälle besser ein als bei der Spezifikation von Testfällen in natürlicher Sprache.

H_5 : Wenn die Testfallerstellung mit einer systemspezifischen Testing DSL erfolgt, fühlen sich Testdesigner sicherer als bei der Erstellung von natürlichsprachlichen Testfällen.

H_6 : Der Spaßfaktor wird bei der Verwendung einer systemspezifischen Testing DSL höher bewertet als bei Testfällen, die in natürlicher Sprache spezifiziert werden.

$H0_2 - H0_6 : Median_{NL}(\text{Einschätzung}) \leq Median_{DSL}(\text{Einschätzung})$
$H1_2 - H1_6 : Median_{NL}(\text{Einschätzung}) > Median_{DSL}(\text{Einschätzung})$

Zum Testen dieser Hypothesen wurde der Mann-Whitney-Test (H_1) und der Wilcoxon-Signed-Rank-Test ($H_2 - H_6$) unter Prüfung der geltenden Voraussetzungen verwendet. Die Analyse der während des Experiments erstellten Testfälle zeigt beispielsweise für das System *MW* eine signifikante Reduktion von Problemen in Testfällen, die mit einer Testing DSL erstellt wurden ($M = 3{,}90$, $Std = 3{,}11$), als bei der Verwendung natürlicher Sprache ($M = 37{,}30$, $Std = 12{,}62$), $U = 0{,}00$, $Z = 3{,}797$, $p < 0{,}001$, $r = 0{,}85$. Auch Testfälle des Systems *CM* enthielten insgesamt signifikant weniger Probleme, wenn sie mit der Testing DSL erstellt wurden ($M = 3{,}60$, $Std = 3{,}596$) als bei der Erstellung mit natürlichen Sprache ($M = 28{,}70$, $Std = 9{,}190$), $U = 0{,}00$, $Z = 3{,}790$, $p < 0{,}001$, $r = 0{,}85$. In beiden Fällen deutet die Effektgröße auf einen großen Effekt hin, da $r > 0{,}50$ ist. Somit ist die Nullhypothese $H0_1$ für beide Systeme zu verwerfen.

Auch die Selbsteinschätzung der Teilnehmer und deren Bewertung des schablonenbasierten Ansatzes mittels einer systemspezifischen Testing DSL zeigen signifikante Ergebnisse. So bewerteten die Teilnehmer die Testfälle hinsichtlich ihrer Konsistenz ($Z = 3{,}449$, $p < 0{,}001$, $r = 0{,}55$), Vollständigkeit ($Z = 2{,}077$, $p = 0{,}024$, $r = 0{,}33$) und Verständlichkeit ($Z = 2{,}693$, $p = 0{,}004$, $r = 0{,}43$) signifikant besser als die zuvor in natürlicher Sprache dokumentierten Testfälle, womit die Nullhypothesen $H0_2 - H0_4$ zu verwerfen sind. Die Teilnehmer wurden zudem gefragt, inwieweit sie glauben, dass der Einsatz einer systemspezifischen Testing DSL das Potenzial hat, die Qualität von Testfällen positiv zu beeinflussen. 45% (9) der Teilnehmer glauben dass eine „eher bessere Qualität“ und 55% (11) eine „deutlich bessere Qualität“ durch Einsatz von Testing DSLs erzielt werden kann. Darüber hinaus fühlen sich die Teilnehmer signifikant sicherer bei der Erstellung von Testfällen mit einer systemspezifischen Testing DSLs ($H_5 : Z = 3{,}038$, $p = 0{,}001$, $r = 0{,}48$), haben dabei mehr Spaß ($H_6 : Z = 3{,}471$, $p < 0{,}001$, $r = 0{,}55$) und 85% (17) der Teilnehmer glauben, dass dies die Testfallerstellung beschleunigen würde. Die Benutzerfreundlichkeit des schablonenbasierten Konzeptes, das in einem Werkzeug umgesetzt wurde, wurde als gut bewertet (SUS-Score von $M = 78{,}88$). All diese Ergebnisse zusammen unterstreichen die Vorteile der Spezifikation von Testfällen mittels einer

systemspezifischen Testing DSL und fördern somit eine reibungslose Einführung des schablonenbasierten Ansatzes in die industrielle Praxis.

4 Zusammenfassung

Zusammenfassend wurden in dieser Dissertation in Kooperation mit Praktikern aus dem Automobilbereich Herausforderungen identifiziert und bewertet, die die Qualität von automobilen Testfallspezifikationen beeinflussen. Darauf aufbauend wurden zwei konkrete Qualitätssicherungsmaßnahmen entwickelt, um Qualitätsmängel in Testfällen zu adressieren und Inspektoren beim Review von automobilen Testfallspezifikationen zu unterstützen. Ein Schwerpunkt liegt dabei vor allem auf der entwickelten Testfallspezifikationsorientierten Domänenanalysemethode und deren Automatisierung, um die Entwicklung von systemspezifischen Testing DSLs als konstruktive Qualitätssicherungsmaßnahme zu unterstützen. Insbesondere konnte anhand der Ergebnisse des kontrollierten Experiments ein positiver Effekt der Verwendung von Testing DSLs auf die Qualität von Testfällen bestätigt werden. So enthalten diese Testfälle weniger Probleme hinsichtlich Unvollständigkeit, falscher Verwendung von Parametern, struktureller Mängel und falscher Beschreibung des Testablaufs. Auch die subjektive Bewertung durch die Teilnehmer ergab, dass die Qualität der Testfälle hinsichtlich Konsistenz, Vollständigkeit und Verständlichkeit deutlich höher eingeschätzt wird. Außerdem fühlen sie sich bei der Verwendung der Testing DSL sicherer als bei der Spezifikation von Testfällen in natürlicher Sprache, was ein Indiz für die reibungslose Einführung des Testing DSL Ansatzes in die industrielle Praxis ist.

Literaturverzeichnis

[Ba15] Bano, Muneera: Addressing the Challenges of Requirements Ambiguity: A Review of Empirical Literature. In: Proc. of EmpiRE'15. S. 21–24, 2015.

[DBK03] Denger, Christian; Berry, Daniel M.; Kamsties, Erik: Higher Quality Requirements Specifications through Natural Language Patterns. In: Proc. of SwSTE'03. S. 80–90, 2003.

[Es96] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. of KDD'96. S. 226–231, 1996.

[Fe17] Femmer, Henning; Méndez Fernández, Daniel; Wagner, Stefan; Eder, Sebastian: Rapid Quality Assurance with Requirements Smells. Journal of Systems and Software, 123:190–213, 2017.

[Gr03] Grimm, Klaus: Software Technology in an Automotive Company: Major Challenges. In: Proc. of ICSE'03. S. 498–503, 2003.

[Ha13] Hauptmann, Benedikt; Heinemann, Lars; Vaas, Rudolf; Braun, Peter: Hunting for Smells in Natural Language Tests. In: Proc. of ICSE'13. S. 1217–1220, 2013.

[In11] International Organization for Standardization (ISO): Road Vehicles – Functional Safety – Part 4: Product Development at the System Level. ISO 26262-4:2011, 2011.

[In13] International Organization for Standardization (ISO): Software and Systems Engineering – Software Testing – Part 3: Test Documentation. ISO/IEC/IEEE 29119-3:2013, 2013.

[In19] International Software Testing Qualifications Board (ISTQB): Standard Glossary of Terms used in Software Testing. Version 3.3, 2019.

[JT19] Juhnke, Katharina; Tichy, Matthias: A Tailored Domain Analysis Method for Developing System-Specific Testing DSLs Enabling their Smooth Introduction in Automotive Practice. In: Proc. of SEAA'19. S. 10–18, 2019.

[JTH18a] Juhnke, Katharina; Tichy, Matthias; Houdek, Frank: Challenges Concerning Test Case Specifications in Automotive Software Testing. In: Proc. of SEAA'18. S. 33–40, 2018.

[JTH18b] Juhnke, Katharina; Tichy, Matthias; Houdek, Frank: Challenges with Automotive Test Case Specifications. In: Proc. of ICSE'18. S. 131–132, 2018.

[JTH18c] Juhnke, Katharina; Tichy, Matthias; Houdek, Frank: Quality Indicators for Automotive Test Case Specifications. In: Workshop on Software Engineering for Applied Embedded RealTime Systems (SE'18), S. 96–100. 2018.

[JTH20] Juhnke, Katharina; Tichy, Matthias; Houdek, Frank: Challenges Concerning Test Case Specifications in Automotive Software Testing: Assessment of Frequency and Criticality. Software Quality Journal, S. 1–57, 2020.

[Ju21] Juhnke, Katharina: Improving the Quality of Automotive Test Case Specifications. Dissertation, Universität Ulm, Institut für Softwaretechnik und Programmiersprachen, 2021. Im Druck.

[KBM16] Kosar, Tomaž; Bohra, Sudev; Mernik, Marjan: Domain-Specific Languages: A Systematic Mapping Study. Information and Software Technology, 71:77–91, 2016.

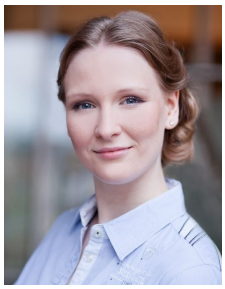
[KPM13] Kasoju, Abhinaya; Petersen, Kai; Mäntylä, Mika V.: Analyzing an Automotive Testing Process with Evidence-based Software Engineering. Information and Software Technology, 55(7):1237–1259, 2013.

[PD90] Prieto-Díaz, Rubén: Domain Analysis: An Introduction. ACM SIGSOFT Software Engineering Notes, 15(2):47–54, 1990.

[Pe07] Peffers, Ken; Tuunanen, Tuure; Rothenberger, Marcus A.; Chatterjee, Samir: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3):45–77, 2007.

[SPL11] Sundmark, Daniel; Petersen, Kai; Larsson, Stig: An Exploratory Case Study of Testing in an Automotive Electrical System Release Process. In: Proc. of SIES'11. IEEE, S. 166–175, 2011.

[VD17] VDA QMC Working Group 13 / Automotive SIG: Automotive SPICE Process Assessment / Reference Model. Version 3.1, 2017.



Katharina Juhnke erhielt ihren M. Sc. in Informatik an der HTWK Leipzig. Sie arbeitete als IT-Beraterin, Requirements & Usability Engineer bevor sie sich bei Mercedes-Benz Cars Development auf die Verbesserung der Qualität von Testfallspezifikationen fokussierte. Diesbezüglich entstand auch ihre Dissertation in Kooperation mit der Daimler AG als Industriepartner. Zurzeit arbeitet sie an der Universität Ulm als Postdoktorandin. Ihre Forschungsinteressen sind Embedded Software Testing, domänenspezifische Sprachen sowie Usability Engineering und empirische Forschungsmethoden.

Verständnis und Weiterentwicklung der Programmiersprache Rust[1]

Ralf Jung[2]

Abstract: *Rust* ist eine junge systemnahe Programmiersprache. Sie vereint die Sicherheit und das Abstraktionsniveau von Sprachen wie Java und Haskell mit der Kontrolle von Systemressourcen, wie C und C++ sie bieten. Meine Dissertation [Ju20a] untersucht die Sicherheitsgarantien von Rust erstmals formell und trägt somit entscheidend zum besseren *Verständnis* und zur *Entwicklung* dieser zunehmend bedeutsamen Sprache bei. Dafür habe ich drei Systeme entwickelt und im Beweisassistenten Coq verifiziert: RustBelt, Iris, und Stacked Borrows.

RustBelt ist ein formelles Modell des Typsystems von Rust einschließlich eines Korrektheitsbeweises, welcher die Sicherheit von Speicherzugriffen und Nebenläufigkeit zeigt. RustBelt ist in der Lage, einige komplexe Komponenten der Standardbibliothek von Rust zu verifizieren, obwohl die Implementierung dieser Komponenten intern *unsichere* Sprachkonstrukte verwendet.

RustBelt ist nur möglich dank der Entwicklung von *Iris*, einem Framework zur Konstruktion von Separationslogiken zur Programmverifikation von beliebigen Programmiersprachen. Die Stärke von Iris liegt in der Möglichkeit, neue Beweismethoden mit Hilfe weniger einfacher Bausteine herzuleiten.

Stacked Borrows ist eine Erweiterung der Spezifikation von Rust, die es dem Compiler erlaubt, den Quelltext mit Hilfe der im Typsystem kodierten Alias-Informationen besser zu analysieren. So werden neue mächtige intraprozedurale Optimierungen ermöglicht.

1 Einführung

Im Bereich der Systemprogrammierung genießen Sprachen ohne starke Typ- und Speichersicherheit nach wie vor eine große Verbreitung. Ein Großteil der Software, die das Fundament moderner Computer bildet, ist in C oder C++ geschrieben – Sprachen, die sich über die Jahre deutlich weiterentwickelt haben, aber nach wie vor die Verantwortung für grundlegende Speichersicherheit dem Programmierer überlassen. Programmierer jedoch machen unvermeidlich Fehler, und das mit handfesten Konsequenzen: sowohl Microsoft als auch die Entwickler von Google Chrome geben an, dass ca. 70% der Sicherheitslücken in ihren Produkten durch Verletzungen der Speichersicherheit entstehen [Th19, Ch20].

Viele Sprachen erzielen Speichersicherheit, indem sie zu einem gewissen Grad dem Programmierer die Kontrolle darüber entziehen, wie das Programm mit dem Speicher interagiert. Das gilt insbesondere für das automatische Bereinigen des Speichers durch einen "garbage collector", wobei sowohl die Struktur der Daten im Speicher als auch die Deallokation von nicht mehr benötigtem Speicher aus der Hand des Programmierers genommen werden. Bei der Programmierung von fundamentalen Systemkomponenten ist dies jedoch

[1] Englischer Titel der Dissertation: "Understanding and Evolving the Rust Programming Language"

[2] MPI-SWS, jung@mpi-sws.org

keine Option. Hier sind minimaler Speicher- und Rechenzeitverbrauch oberste Priorität, und der Programmierer muss diese Aspekte direkt unter Kontrolle haben. Daher wird eine Sprache benötigt, die in dieser Hinsicht mit C und C++ auf einer Ebene steht, während sie gleichzeitig Speichersicherheit und Typsicherheit garantiert.

Rust beansprucht für sich, solch eine Sprache zu sein. Seinen Ursprüngen bei Mozilla entwachsen, lebt Rust inzwischen von einer großen Open-Source-Gemeinschaft und wird zunehmend auch industriell eingesetzt: in Mozillas Firefox, aber auch in vielen anderen Firmen vom kleinen Start-Up bis zu Tech-Riesen wie Amazon und Microsoft.[3]

Ähnlich wie C++ hat der Programmierer bei Rust volle Kontrolle über die Struktur und Deallokation von Daten im Speicher. Eine weitere Parallele ist der Fokus auf "zero-cost" Abstraktionen im Sinne von Stroustrup [St94]: "What you don't use, you don't pay for. And further: What you do use, you couldn't hand code any better."

Im großen Gegensatz zu C++ jedoch verspricht Rust Typsicherheit und Speichersicherheit. Rust will außerdem Probleme aus der Welt schaffen, unter denen auch viele Sprachen mit Speichersicherheit leiden, zum Beispiel Iteratoren, die ihre Gültigkeit durch gleichzeitige Veränderung der zugrundeliegenden Datenstruktur verlieren. Darüber hinaus nimmt Rust für sich in Anspruch, Nebenläufigkeitsfehler zu vermeiden: Rust-Programme haben keine "data races", d.h. es gibt keine *unbeabsichtigte* Kommunikation zwischen mehreren Threads durch das Verwenden desselben Speicherbereichs. Damit geht Rust über die Sicherheitsgarantien der meisten "sicheren" Sprachen hinaus.[4]

Soweit die Behauptungen, mit denen Rust von sich Reden macht. Doch ist Rust wirklich in der Lage, diesen Ansprüchen gerecht zu werden? Im Rahmen meiner Dissertation [Ju20a] habe ich das erste logische Framework entwickelt, das in der Lage ist, diese Ansprüche formell zu beweisen. Im Folgenden werde ich den grundlegenden Ansatz dieses Beweises erläutern; zuerst jedoch will ich darlegen, was diesen Beweis so anspruchsvoll macht.

2 Mehr Sicherheit trotz "unsafe" Code?

Das folgende Beispiel zeigt repräsentativ, welche Art von Problemen in C++-Programmen auftreten können:

```
1  std::vector<int> v { 10, 11 };
2  int *vptr = &v[1]; // Zeigt auf den *Inhalt* von ‘v‘.
3  v.push_back(12); // Verschiebt den Inhalt von ‘v‘ an eine andere Stelle.
4  std::cout << *vptr; // Alter Inhalt von ‘v‘ nach Deallokation verwendet.
```

In der ersten Zeile wird ein `std::vector<int>`, also ein vergrößerbares Array von Ganzzahlen, angelegt. Der Inhalt von `v`, die beiden Elemente `10` und `11`, werden in einem dafür

[3] Unter `https://www.rust-lang.org/production/users` sind Firmen aufgelistet, die Rust in ihren Produkten einsetzen; unter `https://foundation.rust-lang.org/members` sind die Mitglieder der kürzlich gegründeten Rust Foundation zu sehen.

[4] In Java zum Beispiel sind data races möglich, und es gibt nur schwache Garantien, was in diesem Fall passiert. In Go und Swift können data races sogar die Speichersicherheit verletzen.

allozierten Bereich im Speicher abgelegt. In der zweiten Zeile wird ein Zeiger `vptr` erstellt, der *in* diesen Bereich zeigt; genau genommen zeigt er auf das zweite Element (welches aktuell den Wert `11` hat). `v` und `vptr` zeigen jetzt beide auf überlappende Teile desselben Speicherbereichs, es besteht also *Aliasing* zwischen diesen Zeigern.

In Zeile 3 wird `v` um ein neues Element am Ende verlängert: `12` wird nach der `11` in den Speicherbereich mit dem Inhalt von `v` abgelegt. Falls dafür nicht mehr genug Platz ist, wird ein neuer Speicherbereich alloziert, die vorhandenen Elemente werden dorthin verschoben und der alte Bereich wird dealloziert. Für dieses Beispiel gehen wir davon aus, dass genau das passiert. Dieser Fall ist besonders interessant, weil `vptr` noch auf den alten Speicherbereich zeigt! Mit anderen Worten, durch das Hinzufügen eines neuen Elements zu `v` wurde der Zeiger `vptr` ungültig. Das wird in der letzten Zeile zum Problem. Hier greift das Programm mit Hilfe des Zeigers auf den alten, deallozierten Speicher zu: ein klassischer "use-after-free"-Fehler.

Das Beispiel mag künstlich wirken, aber in der Praxis wird der Aufruf von `push_back` an einer völlig anderen Stelle im Code sein als `vptr`. Statt eines expliziten Zeigers ist `vptr` oft ein Iterator; man spricht dann auch von "iterator invalidation". Insbesondere bei der Wartung von vorhandenem Code ist es oft quasi unmöglich festzustellen, ob ein `push_back` irgendwelche wichtigen Zeiger an anderer Stelle im Programm ungültig macht.

In Rust werden solche Probleme statisch erkannt – an Stelle eines Laufzeitfehlers oder einer Sicherheitslücke gibt es eine Fehlermeldung vom Compiler. Die Rust-Version des C++-Programms sieht wie folgt aus;

```
1  let mut v: Vec<i32> = vec![10, 11];
2  let vptr = &mut v[1]; // Zeigt auf den *Inhalt* von ‘v‘.
3  v.push(12); // Verschiebt den Inhalt von ‘v‘ an eine andere Stelle.
4  println!("{}", *vptr); // Fehlermeldung durch Compiler.
```

Wie gehabt gibt es hier einen getrennten Bereich im Speicher, in dem die Elemente von `v` gespeichert werden; und wie gehabt kann `push` diesen Speicherbereich verschieben, was `vptr` ungültig macht und in der letzten Zeile zu einem Problem würde – wenn nicht der Compiler das Programm mit dem Fehler ablehnte, dass `v` nicht "mehrfach zur selben Zeit veränderlich ausgeliehen" werden kann. In meiner Dissertation erkläre ich im Detail, wie der Rust-Compiler dieses Problem erkennt. Für diese Einführung genügt es zu wissen, dass das Typsystem von Rust komplizierter ist als bei anderen Sprachen üblich und Ideen wie *Eigentum* (eine Form von linearen Typen) involviert sowie eine Datenflussanalyse, die *Leihgaben* ("borrows") von Zeigern für gewisse *Lebenszeiten* ("lifetimes") ermöglicht.

Allerdings hat dieser Ansatz eine erhebliche Einschränkung: Datenstrukturen wie `Vec` verwenden Code, den auch das komplizierte Typsystem von Rust nicht vollständig prüfen kann. Statt dessen gibt es in Rust das Schlüsselwort `unsafe`, welches syntaktisch Bereiche des Programms markiert, in denen gefährliche Operationen durchgeführt werden können. So wird sichergestellt, dass Programmierer nicht versehentlich den sicheren Bereich von Rust verlassen. `Vec` jedoch nutzt `unsafe`, um mit Hilfe ungeschützter Zeiger das Hinzufügen von Elementen in amortisiert konstanter Zeit zu realisieren (genau wie `std::vector` in C++).

Damit stellt sich natürlich die Frage: macht dieser `unsafe` Code nicht die Sicherheitsgarantien von Rust zunichte? Um das zu verhindern, bedient Rust sich des Konzepts der *Kapselung*. Das Typsystem kann sicherstellen, dass auch `unsafe` Code "nach außen", also bei Benutzung seiner öffentlichen Schnittstelle, sicher zu verwenden ist. Für die Sicherheit der Implementierung ist jedoch der Programmierer selbst verantwortlich.

Dieser Ansatz kann gut am Beispiel von `Vec` erläutert werden. In C++ gibt es für Typen wie `std::vector` ausführliche Dokumentation, welche erklärt, wie man den Typ korrekt benutzt und welche auf Probleme wie die in dem obigen Beispiel hinweist. Allerdings sind dies alles nur *Kommentare* – für den Compiler gibt es keine Möglichkeit, den Programmierer beim Einhalten dieser Regeln zu unterstützen. Im Gegensatz dazu sind bei `Vec` in Rust die *Typen* der beteiligten Funktionen detailliert genug, dass der Compiler prüfen kann, ob der Nutzer sich an die angegebenen Regeln hält. Die Autoren von `Vec` versprechen, dass die internen `unsafe` Operationen von diesen Typen "gekapselt" werden: Programmierer, die nur die öffentliche Schnittstelle zu `Vec` verwenden und selber keinen Gebrauch von `unsafe` machen, genießen weiterhin die volle Typ- und Speichersicherheit von Rust. Das Typsystem von Rust ist nicht stark genug, um korrekte "Kapselung" zu prüfen, aber es *ist* stark genug, um zu prüfen, ob die öffentliche Schnittstelle *korrekt verwendet* wird.

Dabei bleibt jedoch ein Problem: Die öffentliche Schnittstelle von `Vec` basiert auf den Konzepten von Eigentum und Leihgaben, die dem Rust-Typsystem zugrunde liegen. Intuitiv sind diese Konzepte gut verstanden, aber der Teufel steckt wie üblich im Detail: Sind die Invarianten des Typsystem tatsächlich stark genug, um die korrekte Verwendung von `Vec` zu garantieren? Bei `Vec` ist das relativ unstrittig; die Interaktion mit dem Typsystem ist hier nicht kompliziert. Andere Komponenten der Standardbibliothek, wie zum Beispiel `Mutex`, verwenden das Typsystem jedoch auf deutlich interessantere Weise. `Mutex` erlaubt es Rust-Code, veränderliche Daten zwischen mehreren Threads zu teilen, wobei durch ein Lock sichergestellt wird, dass immer nur ein Thread gleichzeitig auf den Daten arbeitet. Im Allgemeinen geht der Rust-Compiler davon aus, dass geteilte Daten nicht verändert werden können, aber Typen wie `Mutex` verwenden eine subtile Kombination von getypter Schnittstelle und Laufzeitkontrollen, um diese Einschränkung zu umgehen und so Rust-Code das sichere Arbeiten mit geteilten veränderlichen Daten zu ermöglichen. Es ist alles andere als offensichtlich, dass `Mutex` auf diese Art nicht die Sicherheitsgarantien von Rust verletzt. Um Rusts Versprechen von sicherer Systemprogrammierung einzulösen, ist es also wichtig, dies formell prüfen zu können.

Allerdings ist der übliche *syntaktische* Ansatz zum Beweis von Typsicherheit [WF94] für Rust nicht geeignet. Bei diesem Ansatz geht man von einer geschlossenen Welt aus, man nimmt also an, dass dem Programm nur einen fester Satz von Primitiven mit ihren Typregeln zur Verfügung steht. Typsicherheit in Rust kann so nur für Programme gezeigt werden, die keinerlei `unsafe` Code verwenden, auch nicht indirekt über eingebundene Bibliotheken. Natürlich gilt diese Einschränkung auch für andere Programmiersprachen, deren Typsystem man mit einem Konstrukt wie `unsafe` umgehen kann, wie zum Beispiel OCaml mit `Obj.magic`, Haskell mit `unsafePerformIO`, oder jede beliebige Sprache von der aus man C-Bibliotheken aufrufen kann. Allerdings werden solche Konstrukte in anderen Sprachen bei formeller Betrachtung der Typsicherheit üblicherweise ignoriert. Diese Lücke in

den Beweisen ist nicht zufriedenstellend, und sie wäre in Rust noch deutlich größer als bei den meisten anderen Sprachen: Um eine mit C oder C++ vergleichbare Performance zu erreichen, sind bei grundlegenden Datenstrukturen keine Kompromisse möglich. `unsafe` wird daher in den unteren Schichten des Rust-Ökosystems sehr viel eingesetzt, und jede realistische Betrachtung von Rust muss sich auch mit `unsafe` auseinandersetzen.

3 RustBelt: Ein tieferes Verständnis von Rust

In meiner Dissertation beschreibe ich RustBelt [Ju18a], das erste formelle (und maschinengeprüfte) Modell von Rust, welches in der Lage ist, Typsicherheit bei Verwendung von `unsafe` zu beweisen. Dieser Beweis ist *erweiterbar* für eine neue Bibliothek mit interner Verwendung von `unsafe`, und zwar in dem Sinne, dass RustBelt klar definiert, welche Aussage bewiesen werden muss, damit alle Nutzer dieser Bibliothek weiterhin Speichersicherheit und Threadsicherheit genießen. Der Beweis ist außerdem *modular* in dem Sinne, dass mehrere Bibliotheken unabhängig voneinander verifiziert werden können und der Beweis auch bei beliebiger Kombination ihrer Schnittstellen seine Gültigkeit behält.

Die zentrale Idee hinter diesem Beweis ist die, ein *semantisches Modell* von Rusts Typsystem zu definieren. Dies ist ein durchaus bewährter Ansatz, der bereits für den allerersten Typsicherheitsbeweis von Milner [Mi78] für einen polymorphen λ-Kalkül im Stile von ML eingesetzt wurde. Milners Ansatz basiert auf früheren Arbeiten an *logischen Relationen* [Ta67]. Logische Relationen definieren die *Bedeutung* eines Typs als die Menge der Werte, die das gewünschte *beobachtbare Verhalten* an den Tag legen. Insbesondere ist für den Typ einer Funktion nur ihr Eingabe-Ausgabe-Verhalten relevant, im Gegensatz zum syntaktischen Ansatz, wo der Quelltext der Funktion gewissen Regeln genügen muss. Beim semantischen Ansatz darf die Funktion durchaus potentiell unsichere Dinge tun, solange die Garantien und Invarianten des Typsystems dabei nicht verletzt werden.

Es gelang zunächst nicht, diesen Ansatz auf mächtigere Sprachen mit Seiteneffekten und Funktionen höherer Ordnung anzuwenden, weshalb sich der einfachere (aber weniger mächtige) syntaktische Ansatz von “progress and preservation” durchsetzte. Allerdings gab es in den letzten zwei Jahrzehnten große Fortschritte auf dem Gebiet der logischen Relationen [Ah04], sodass es inzwischen prinzipiell möglich ist, den semantischen Ansatz mit den ausdrucksstarken Typsystemen moderner Sprachen zu verwenden. In meiner Arbeit wende ich diesen Ansatz nun erstmals auf ein Typsystem wie das von Rust an.

Nach der Definition des semantischen Modells sind drei Schritte nötig, um den Beweis der Typsicherheit in RustBelt zu vervollständigen:

1. Das *fundamentale Theorem der logischen Relation* stellt sicher, dass alle syntaktischen Typregeln korrekt sind, wenn man sie semantisch interpretiert. Dafür wird nicht nur den Typen, sondern auch allen anderen Komponenten der Typregeln eine semantische “Bedeutung” zugeordnet, welche (intuitiv gesprochen) die durch das Typsystem kodierten Invarianten explizit macht.
2. Zudem muss man beweisen, dass ein *semantisch* wohlgetyptes Programm in der Tat speichersicher und threadsicher ist.

3. Schließlich müssen die Bibliotheken, die intern `unsafe` verwenden, korrekt bewiesen werden. Das semantische Modell ermöglicht es, die Typen der öffentlichen Schnittstelle dieser Bibliotheken in einen formellen Vertrag umzuwandeln. Die Herausforderung besteht nun darin, zu beweisen, dass die Implementierung der Schnittstelle dem Vertrag genügt.

Zusammengenommen zeigen diese Schritte, dass ein Programm speicher- und threadsicher ist, wenn aller `unsafe` Code sich innerhalb von korrekt bewiesenen Bibliotheken befindet.

Für den dritten Schritt habe ich viele komplexe Komponenten der Standardbibliothek korrekt bewiesen, die eine zentrale Rolle im Rust-Ökosystem spielen, insbesondere solche zum Arbeiten mit geteilten veränderlichen Daten: `Arc`, `Rc`, `Cell`, `RefCell`, `Mutex` (hier wurde durch meine Arbeit ein Fehler aufgedeckt und behoben), `RwLock`, `mem::swap` und `thread::spawn`; sowie `rayon::join` und `take_mut`, welche weitere interessante Aspekte des Typsystems aufzeigen. Alle diese Beweise habe ich mit dem Beweisassistenten Coq durchgeführt; sie wurden also von Coq auf Korrektheit geprüft.

Die Entwicklung eines semantischen Modells für ein Typsystem wie das von Rust stellte mich vor einige größere technische Herausforderungen. In dieser Zusammenfassung möchte ich kurz auf die beiden größten Hürden eingehen: die *Wahl der richtigen Logik* und die Entwicklung eines Modells für *Leihgaben und Lebenszeiten*.

Die Wahl der Logik. "Welche Logik wird verwendet" mag nach einer seltsamen Frage klingen, aber für das semantische Modell von RustBelt war dies in der Tat eine der wichtigsten Entscheidungen. In Rust drücken Typen nicht nur aus, dass ein Wert eine bestimmte Form hat, sondern decken auch Aspekte des *Eigentums* an den verwendeten Ressourcen ab, zum Beispiel die exklusive Kontrolle über einen bestimmten Speicherbereich. Eigentum kann explizit beim Bau des semantischen Modells in Betracht gezogen werden, aber dieser Ansatz ist mühselig und fehleranfällig, vergleichbar mit dem Schreiben eines Programms in Assemblersprache. RustBelt verwendet Separationslogik ("separation logic") [Re02], um auf einem höheren Abstraktionsniveau arbeiten zu können. Separationslogik kann direkt logische Aussagen über Eigentum von Speicherbereichen treffen und eignet sich daher gut dafür, das Eigentum von Typen wie `Vec` zu definieren.

Allerdings ist Eigentum von Speicherbereichen allein nicht ausreichend. Um die Korrektheit von Rust-Typen wie `Mutex` zu beweisen, sind flexiblere Formen von Eigentum notwendig. Für solche Zwecke haben wir *Iris* [Ju15, Ju16, Kr17, Ju18b] entwickelt, eine hochgradig flexible Separationslogik wo Nutzer ihre eigenen Formen von "Eigentum" definieren können. Iris unterstützt außerdem "step-indexing" [DAB11], eine wichtige Komponente moderner logischer Relationen, und nimmt dem Nutzer weitgehend die üblicherweise damit verbundene Buchführung ab. Zu guter Letzt ermöglicht Iris interaktive maschinengeprüfte Beweise in Coq [Kr18].

Leihgaben und Lebenszeiten. Für ein vollständiges Modell des Typsystems von Rust wird eine Logik benötigt, die in der Lage ist, nicht nur über Eigentum zu argumentieren, sondern auch über *Leihgaben* von Eigentum für eine gewisse *Lebenszeit*. Hier stellte es

sich als entscheidend heraus, dass Iris von Anfang an darauf ausgelegt war, dem Nutzer das Herleiten neuer Beweismethoden in der Logik möglichst einfach zu machen.

Unter Verwendung aller wichtigen Komponenten von Iris, insbesondere *impredikativer Invarianten* [SB14] und *Ressourcen höherer Ordnung* [Ju16], habe ich eine solche "Lebenszeitlogik" in Iris definiert und ihre Korrektheit beweisen. Die Lebenszeitlogik ermöglicht es, bei der Definition des semantischen Modells von Rust-Typen direkten Gebrauch vom Konzept einer Leihgabe zu machen, und auch der Korrektheitsbeweis des Typsystems kann auf diesem hohen Abstraktionsniveau geführt werden.

4 Stacked Borrows: Mehr Optimierungen für Rust

Typsysteme wie das von Rust sind nicht nur nützlich, weil sie Programme sicherer und zuverlässiger machen, sie können auch dabei helfen, effizienteren Code zu erzeugen. Zum Beispiel muss eine Sprache mit einem starken Typsystem nicht Rechenzeit und Speicher darauf aufwenden, dynamische Typinformationen zu verwalten. In Rust erzwingt das Typsystem eine strikte Aliasing-Disziplin auf Zeigern. Es wäre daher sehr interessant, diese statisch bekannte Information für Optimierungen auszunutzen.

So sind zum Beispiel *veränderliche Referenzen*, geschrieben `&mut T`, in Rust immer exklusiv in dem Sinne, dass aktuell keine anderen Zeiger auf dieselben Daten verwendet werden können. Das sollte es uns erlauben, die folgende Funktion zu optimieren:

```
1  fn example1(x: &mut i32, y: &mut i32) -> i32 {
2    *x = 42;
3    *y = 13;
4    return *x; // Hier wird 42 gelesen, weil x und y nicht aliasen!
5  }
```

`x` und `y` sind als veränderliche Referenzen beide exklusiv und können daher nicht aliasen, d.h. die Speicherbereiche, auf die sie zeigen, können nicht überlappen. Daher sollte dem Compiler die Annahme gestattet sein, dass in Zeile 4 immer `42` gelesen wird; der Speicherzugriff kann also durch eine Konstante ersetzt werden.

In Rust wird die Situation jedoch durch `unsafe` Code verkompliziert, welcher mittels direkter Zeigermanipulation die üblichen Alias-Regeln umgehen kann. Es ist nicht schwer, `unsafe` Code zu schreiben, welcher dazu führt, dass die obige Funktion `13` zurückgibt:[5]

```
1  fn main() {
2    let mut local = 5;
3    let raw_pointer = &mut local as *mut i32;
4    let result = unsafe {
5      example1(&mut *raw_pointer, &mut *raw_pointer)
6    };
7    println!("{}", result); // Ausgabe: "13".
8  }
```

[5] Diese Tests wurden mit Rust 1.35.0 im "release mode" durchgeführt.

In Zeile 3 wird die *Referenz* vom Typ `&mut i32` in einen *ungeschützten Zeiger* ("raw pointer") vom Typ `*mut i32` konvertiert. Wie bei Zeigern in C kann man ungeschützte Zeiger in Rust in Ganzzahlen konvertieren und umgekehrt, und auch Zeigerarithmetik ist möglich. Um Speichersicherheit nicht zu gefährden, ist das *Dereferenzieren* solcher Zeiger nur innerhalb von `unsafe`-Blöcken erlaubt; der Programmierer muss also explizit angeben, potentiell gefährliche Operationen zu verwenden, und haftet an dieser Stelle selber für die Typsicherheit. Ungeschützte Zeiger sind bei der Interaktion mit C-Bibliotheken notwendig und für eine effiziente Implementierung von Datenstrukturen wie `Vec`.

Dieses Beispiel jedoch nutzt ungeschützte Zeiger, um das Typsystem gezielt zu untergraben. In Zeile 5 wird der Zeiger zurück in eine Referenz umgewandelt, indem man ihn dereferenziert und dann direkt eine neue Referenz erstellt (`&mut *raw_pointer`). Und weil das Typsystem diese Zeiger kaum kontrolliert, kann dies auch zweimal geschehen! Im Endeffekt wird also `example1` mit zwei Referenzen auf dieselbe Variable aufgerufen: es besteht Aliasing zwischen `x` und `y`, was eigentlich unmöglich sein sollte. `example1` gibt dementsprechend `13` zurück, und wenn das Programm so optimiert würde, dass es immer `42` zurückgibt, würde sich das beobachtbare Verhalten des Programms verändern. Damit ist die Optimierung in diesem Fall also inkorrekt.

Es ist an dieser Stelle verlockend, das Problem zu ignorieren, da es ja "nur" `unsafe` Code betrifft. Dies wird jedoch der (oben bereits angesprochenen) wichtigen Rolle von `unsafe` Code im Rust-Ökosystem nicht gerecht. Damit die Optimierung auch bei Verwendung von `unsafe` durchgeführt werden kann, muss vom Programmierer verlangt werden, dass `unsafe` Code das Typsystem nicht wie oben geschehen untergräbt. Doch was genau sind die Bedingungen, die `unsafe` Code dafür erfüllen muss?

Als Antwort auf diese Frage beschreibe ich im dritten Teil meiner Dissertation *Stacked Borrows* [Ju20b], eine operationale Semantik für das Aliasing von Zeigern in Rust. Gemäß dieser Semantik hat das Beispielprogramm *undefiniertes Verhalten*, es gilt also als ungültig.[6] Der Compiler muss für ungültige Programme keine korrekte Ausführung sicherstellen, sodass hier also kein Gegenbeispiel mehr vorliegt. Gleichzeitig gilt: gültiger `unsafe` Code mit voll definiertem Verhalten wird durch die Optimierung nicht beeinflusst.

Im Vergleich zu einer naiven Semantik, wie sie in RustBelt verwendet wird, fügt Stacked Borrows eine neue Form von undefiniertem Verhalten hinzu: die Verletzung der Aliasing-Regeln. Bei undefiniertem Verhalten ist allerdings Speichersicherheit nicht mehr gewährleistet, daher darf es nicht "zu viel" undefiniertes Verhalten geben. Alle Programme im typsicheren Fragment von Rust (ohne `unsafe`) müssen also weiterhin voll definiert sein, und es muss auch weiterhin möglich sein, Datenstrukturen wie `Vec` mit Hilfe von ungeschützten Zeigern zu definieren. Stacked Borrows wurde daher auf zwei Arten validiert:

- Um sicherzustellen, dass nicht zu viel undefiniertes Verhalten eingeführt wurde, habe ich Miri,[7] einen bereits vorhandenen Interpreter für Rust-Programme, mit einer direkten Implementierung der operationalen Semantik von Stacked Borrows ausgestattet. Anschließend habe ich große Teile der Test-Suite der Rust-Standardbibliothek

[6] Dies ist vergleichbar mit einem C-Programm, das z.B. einen Null-Zeiger dereferenziert.
[7] Mehr Informationen zu Miri gibt es online unter `https://github.com/rust-lang/miri/`.

in diesem Interpreter ausgeführt. So konnte ich prüfen, ob all diese Tests den strikten Regeln von Stacked Borrows genügen. Die überwiegende Mehrzahl der Tests brauchte dazu keinerlei Anpassungen. Ich habe jedoch auch einige Verletzungen der Alias-Regeln gefunden, von denen fast alle durch die Rust-Entwickler als Fehler anerkannt und inzwischen behoben wurden.

- Um sicherzustellen, dass alle potentiellen Gegenbeispiele als ungültig erklärt wurden, zeige ich in meiner Dissertation einige Beweisskizzen, welche die Korrektheit von Optimierungen wie der in `example1` für alle gültigen Programme belegen. Diese Beweisskizzen sind von maschinegeprüften Beweisen in Coq untermauert.

5 Schlussfolgerungen und Anwendungen

In meiner Dissertation beschreibe ich drei Projekte, die signifikant zur Entwicklung von interaktiver Programmverifikation allgemein und von Rust im Besonderen beitragen.

Iris, ein Framework zur Entwicklung von flexiblen Separationslogiken, erfährt bereits erste industrielle Nutzung und diente als Grundlage für mehr als 30 weitere Veröffentlichungen von Forschern an zehn verschiedenen Einrichtungen.[8]

RustBelt ist der erste formelle Beweis der Typsicherheit von Rust und Teilen seiner Standardbibliothek. Damit belegt es die Praktikabilität von semantischen Modellen für realistische Sprachen mit komplexen Typsystemen und stellt gleichzeitig die Sicherheit von Rust auf ein solides Fundament.

Stacked Borrows ist der erste konkrete Vorschlag für ein System von Aliasing-Regeln in Rust und dient auch dank meiner engen Zusammenarbeit mit dem Rust-Team aktuell als de-facto Standard für die korrekte Verwendung ungeschützter Zeiger. Die von mir entwickelte Erweiterung für Miri wird vielfach eingesetzt, um die Konformität von `unsafe` Code mit diesen Regeln zu prüfen. Damit ist die Grundlage geschaffen, um Stacked Borrows letztendlich zu einem offiziellen Teil der Rust-Spezifikation weiterzuentwickeln.

Literaturverzeichnis

[Ah04] Ahmed, Amal: Semantics of Types for Mutable State. Dissertation, Princeton University, 2004.

[Ch20] Chromium project, The: Chromium Security: Memory safety. Blog post, https://www.chromium.org/Home/chromium-security/memory-safety, 2020.

[DAB11] Dreyer, Derek; Ahmed, Amal; Birkedal, Lars: Logical Step-Indexed Logical Relations. LMCS, 7(2:16):1–37, Juni 2011.

[Ju15] Jung, Ralf; Swasey, David; Sieczkowski, Filip; Svendsen, Kasper; Turon, Aaron; Birkedal, Lars; Dreyer, Derek: Iris: Monoids and Invariants as an Orthogonal Basis for Concurrent Reasoning. In: POPL. S. 637–650, 2015.

[Ju16] Jung, Ralf; Krebbers, Robbert; Birkedal, Lars; Dreyer, Derek: Higher-order ghost state. In: ICFP. S. 256–269, 2016.

[8] Siehe https://iris-project.org/ für weitere Informationen zu Iris.

[Ju18a] Jung, Ralf; Jourdan, Jacques-Henri; Krebbers, Robbert; Dreyer, Derek: RustBelt: Securing the Foundations of the Rust Programming Language. PACMPL, 2(POPL):66:1–66:34, 2018.

[Ju18b] Jung, Ralf; Krebbers, Robbert; Jourdan, Jacques-Henri; Bizjak, Aleš; Birkedal, Lars; Dreyer, Derek: Iris from the Ground Up: A Modular Foundation for Higher-Order Concurrent Separation Logic. JFP, 28:1–73, November 2018.

[Ju20a] Jung, Ralf: Understanding and Evolving the Rust Programming Language. Dissertation, Universität des Saarlandes, 2020.

[Ju20b] Jung, Ralf; Dang, Hoang-Hai; Kang, Jeehoon; Dreyer, Derek: Stacked Borrows: An Aliasing Model for Rust. PACMPL, 4(POPL), 2020.

[Kr17] Krebbers, Robbert; Jung, Ralf; Bizjak, Aleš; Jourdan, Jacques-Henri; Dreyer, Derek; Birkedal, Lars: The Essence of Higher-Order Concurrent Separation Logic. In: ESOP. Jgg. 10201 in LNCS, S. 696–723, 2017.

[Kr18] Krebbers, Robbert; Jourdan, Jacques-Henri; Jung, Ralf; Tassarotti, Joseph; Kaiser, Jan-Oliver; Timany, Amin; Charguéraud, Arthur; Dreyer, Derek: MoSeL: A General, Extensible Modal Framework for Interactive Proofs in Separation Logic. PACMPL, 2(ICFP):77:1–77:30, 2018.

[Mi78] Milner, Robin: A theory of type polymorphism in programming. Journal of Computer and System Sciences, 17(3):348–375, 1978.

[Re02] Reynolds, John C.: Separation logic: A logic for shared mutable data structures. In: LICS. S. 55–74, 2002.

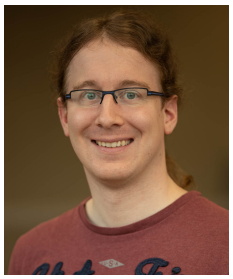
[SB14] Svendsen, Kasper; Birkedal, Lars: Impredicative Concurrent Abstract Predicates. In: ESOP. Jgg. 8410 in LNCS, S. 149–168, 2014.

[St94] Stroustrup, Bjarne: The Design and Evolution of C++. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1994.

[Ta67] Tait, W. W.: Intensional interpretations of functionals of finite type I. Journal of Symbolic Logic, 32(2), 1967.

[Th19] Thomas, Gavin: A proactive approach to more secure code. Blog post, https://msrc-blog.microsoft.com/2019/07/16/a-proactive-approach-to-more-secure-code/, 2019.

[WF94] Wright, Andrew K; Felleisen, Matthias: A syntactic approach to type soundness. Information and computation, 115(1), 1994.



Ralf Jung wurde am am 25. April 1990 in Wiesbaden, Deutschland, geboren. Er studierte Informatik an der Universität des Saarlandes und promovierte 2020 am MPI-SWS mit "summa cum laude". Seine Bachelorarbeit wurde mit dem FdSI-Bachelor-Preis für hervorragende Studienleistungen ausgezeichnet. In der Rust-Gemeinschaft ist er ein weithin anerkannter Experte für `unsafe` Code und leitet dort die Arbeitsgruppe zu "`unsafe` code guidelines". Er bloggt regelmäßig zu diesem Thema[9] und wirkt aktiv bei der Entwicklung der Sprache mit, unter anderem durch zwei Praktika bei Mozilla Research während seiner Promotionszeit. Derzeit arbeitet er als Post-Doc am MPI-SWS und ist zudem als externer Mitarbeiter eng in die PDOS-Forschungsgruppe am MIT eingebunden.

[9] Siehe https://www.ralfj.de/blog/categories/rust.html

Synthese im Kontext Parametrischer Markow-Modelle[1]

Sebastian Junges[2]

Abstract: Markow-Modelle sind ein prominenter Formalismus, um Systeme mit unsicherem Verhalten zu modellieren und zu analysieren. Ein Markov-Modell umfasst (System)-Zustände mit wahrscheinlichkeitsbehafteten Transitionen. Eine typische Fragestellung für ein gegebenes Modell lautet: *Beträgt die maximale Wahrscheinlichkeit, dass ein bestimmter Zustand erreicht wird, weniger als 0,01%?* Um diese Frage zu beantworten, ist es wichtig, dass die Wahrscheinlichkeiten im Markow-Modell exakt bekannt sind. Dies ist leider oft unrealistisch. Um den potentiellen Ungenauigkeiten in diesen Wahrscheinlichkeiten gerecht zu werden, betrachten wir parametrische Modelle, in denen Wahrscheinlichkeiten durch symbolische (genauer: parametrische) Ausdrücke statt durch konkrete Werte dargestellt werden. Es ergeben sich einige natürliche Fragestellungen, zum Beispiel: *Ist die maximale Wahrscheinlichkeit, dass ein bestimmter Zustand erreicht wird, weniger als 0,01% für jede Belegung der Parameter?* In diesem Exposé betrachten wir diese und verwandte Fragestellungen. Die geschilderten Ergebnisse liefern neue Erkenntnisse zur theoretischen Komplexität sowie neue und effektive Methoden. Diese Methoden wurden implementiert und sie verbessern den aktuellen Stand der Technik beträchtlich. Die Implementierungen sind nun in der Lage, Markow-Modelle mit tausenden Parametern und Millionen Zustände zu analysieren.

1 Einführung

Markow-Modelle sind mathematisches Modelle zur Darstellung wahrscheinlichkeitsbehafteter Prozesse und Phänomene. Die Analyse dieser Markow-Modelle zur Bewertung von komplexen Systemen mit stochastischer Dynamik ist allgegenwärtig. Sie ist Forschungsgegenstand, u.a. in der Künstlichen Intelligenz, der Zuverlässigkeitstechnik, der Systembiologie und in den Formalen Methoden. Der Kontext dieser Arbeit liegt im Bereich der Formalen Methoden, insbesondere in der automatisierten Analyse von sicherheitskritischen Systemen. Ein Markov-Modell kann solche Systeme passend erfassen, indem Systemzustände durch wahrscheinlichkeitsbehaftete Transitionen miteinander verbunden werden. Ein Beispiel ist die Wahrscheinlichkeit eines Ausfalls von der Fahrzeugelektronik zu modellieren, wobei die Zustände dann den Zustand verschiedener Fahrzeugkomponenten darstellen. Eine typische Fragestellung wäre nun etwa, wie hoch die Wahrscheinlichkeit ist, dass die Elektronik vor dem nächsten Werkstattbesuch ausfällt. In das Markow-Modell übertragen lautet die Fragestellung dann: *Wie hoch ist die maximale Wahrscheinlichkeit, dass ein bestimmter (Fehler-)Zustand erreicht wird?* Die Analyse ist naturgemäß stark abhängig von den Transitionswahrscheinlichkeiten in dem Modell. Diese sind allerdings mit grosser Unsicherheit verbunden, da sie oft auf (daten- und expertengestützten) Schätzungen basieren. Es stellt sich die Frage, wie die berechneten Wahrscheinlichkeiten im Rahmen dieser Unsicherheiten zu interpretieren sind. Im weiteren Sinne ist die grundlegende Annahme in dieser

[1] Englischer Titel der Dissertation: 'Parameter Synthesis in Markov Models'

[2] sjunges@berkeley.edu, University of California at Berkeley, USA

(a) Lösungsfunktion

(b) Parameterraumpartitionierung

Abb. 1: Parameterbelegungen und Erreichbarkeitswahrscheinlichkeit

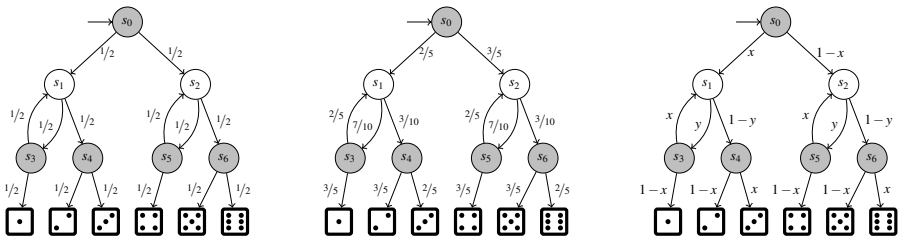
Arbeit, dass die algorithmische Analyse solcher Modelle diese Unsicherheiten in Betracht ziehen muss. Wir betrachten deshalb Markow-Modelle, deren Wahrscheinlichkeiten auch symbolisch statt konkret dargestellt werden können. Eine vergleichbare Situation zu den unsicheren Wahrscheinlichkeiten ergibt sich, wenn die Wahrscheinlichkeiten wählbar sind, was zum Beispiel in randomisierten Algorithmen der Fall ist. Die Frage ist dann, wie diese Wahrscheinlichkeiten geeignet gewählt werden sollten. Im weiteren Verlauf dieses Exposés werden wir sehen, dass eine solche Sichtweise eine natürliche Verbindung zu Planungsproblemen und dem Bestärkenden Lernen liefert.

Die Dissertation [Ju20] betrachtet parametrische Markow-Entscheidungsprozesse sowie den Sonderfall der *parametrischen Markow-Ketten* (*parametric Markov Chains*, pMCs). Das Ersetzen der Parameter durch eine konkrete Belegung induziert die klassischen parameterfreien Markow-Entscheidungsprozesse (MDPs) und Markow-Ketten (MCs). Für jeden dieser induzierten MCs (oder MDPs) können wir die (maximale) Wahrscheinlichkeit bestimmen. Dies ergibt dann einen Graphen wie in Abb. 1a, in dem wir für zwei Parameter die Erreichbarkeitswahrscheinlichkeit darstellen. Die Kurve ist rot gefärbt, wenn diese Wahrscheinlichkeit unterhalb einer bestimmten Grenze liegt. Im weiteren Verlauf dieses Exposés betrachten wir nur MCs mit Erreichbarkeitswahrscheinlichkeiten und verzichten auf MDPs. Die dargestellten Ergebnisse basieren in Teilen auf [De15, Wi19, Ce19a, Cu18, Qu16, Ce19b, Ju18].

2 Problemformulierung

Eine *Markow-Kette* ist ein Tupel aus einer (endlichen) Menge von Zuständen, einem Initialzustand sowie einer Transitionswahrscheinlichkeitsfunktion P, die jeden Zustand s auf eine Verteilung $P(s)$ über Nachfolgezustände abbildet. In dieser Arbeit konzentrieren wir uns auf *Erreichbarkeitswahrscheinlichkeiten*. Bei vorgegebenem Zielzustand t wird diese Wahrscheinlichkeit, t zu erreichen, bestimmt durch die Summe der Pfadwahrscheinlichkeiten aller Pfade, die im Initialzustand starten und in t enden. Die Pfadwahrscheinlichkeit ist dabei durch die Multiplikation der Transitionswahrscheinlichkeiten entlang des jeweiligen Pfades gegeben.

Beispiel 1. *Abb. 2a zeigt eine Markow-Kette mit 13 Zuständen. Transitionen werden durch Kanten dargestellt, welche mit der Transitionswahrscheinlichkeit $P(s)(s')$ beschriftet sind. Die Kette modelliert ein Protokoll, mit dem ein Würfelergebnis durch den wiederholten Wurf einer fairen Münze simuliert wird (der sogenannte Knuth-Yao Würfel). Tatsächlich*



(a) faire Münzen (b) gebiaste Münzen (c) parametrische MC

Abb. 2: Parametrischer Knuth-Yao Würfel

ist die Wahrscheinlichkeit um beispielsweise einen Zielzustand ⚀ zu erreichen genau $\sum_{i>0} 1/2 \cdot (1/2 \cdot 1/2)^i = 1/6$.

In einer parametrischen Markow-Kette (pMC) wird zusätzlich eine endliche *Parametermenge* fixiert. Die Transitionswahrscheinlichkeitsfunktion wird dann so abgewandelt, dass sie jedes Paar von Ursprungszustand s und Nachfolgezustand s' auf ein Polynom abbildet, das die Wahrscheinlichkeit beschreibt, von Zustand s zu Zustand s' zu wechseln.

Beispiel 2. *Was würde passieren, wenn die Münze nicht fair wäre und wir stattdessen zwei unfaire Münzen alternierend werfen würden? Insbesondere werfen wir die eine Münze in allen grauen Zuständen, die andere in allen weißen Zuständen. Abb. 2b zeigt die MC für einen spezifischen Fall. Zustand ⚀ wird nun nicht mal mehr mit einer Wahrscheinlichkeit größer* $3/20$ *erreicht. Wir interessieren uns jedoch für verschiedene unfaire Münzen und wir parameterisieren die Münzen demnach in ihrem Bias. Abb. 2c zeigt eine parameterisierte Variante des Knuth-Yao Würfels mit Parametern* $X = \{x, y\}$*: Hier stellen* x *und* y *die Wahrscheinlichkeiten für 'Kopf' in den grauen beziehungsweise weißen Zuständen dar.*

Eine *Belegung* weist Parametern konkrete Werte zu. In einer pMC ergibt das geeignete Ersetzen von Parameterwerten durch eine Belegung eine Markow-Kette, welche wir als *Instanziierung*. Jede Belegung ist so zu einer Markow-Kette *assoziiert*. Die Menge aller geeigneten Belegungen von Parametern nennen wir den *Parameterraum*. Teilmengen des Parameterraums bezeichnen wir durchgehend als *Region*.

Beispiel 3. *Eine mögliche Belegung der pMC in Abb. 2c ist gegeben durch* $\mathsf{val} := \{x \mapsto 2/5, y \mapsto 7/10\}$*. Die assoziierte Instanziierung ist in Abb. 2b dargestellt. Die pMC in Abb. 2a ist das Ergebnis des Ersetzens mit* $\mathsf{val}' := \{x, y \mapsto 1/2\}$*. Der Parameterraum ist* $\{\mathsf{val} \mid 0 \leq \mathsf{val}(x), \mathsf{val}(y) \leq 1\}$*. Wir interessieren uns zum Beispiel besonders fur die Region der fast-fairen Münzen:* $\{\mathsf{val} \mid 9/20 < \mathsf{val}(x), \mathsf{val}(y) < 11/20\}$.

Eine *Eigenschaft* ist eine Kombination aus einem oder mehreren Zielzuständen und einer Schranke die eine Erreichbarkeitswahrscheinlichkeit beschränken soll. Eine MC *erfüllt* die Eigenschaft, wenn die Erreichbarkeitswahrscheinlichkeit zum Zielzustand nicht unter dieser Schranke liegt. Für eine gegebene Region interessieren wir uns nun zum Beispiel für die Teilmenge der Belegungen, sodass die assoziierte MCs die Eigenschaft erfüllen.

Belegungen, deren assoziierte MC eine Eigenschaft erfüllen, nennen wir *akzeptierend*, alle anderen *ablehnend*. Es ergeben sich zwei zentrale Fragestellungen:

Für gegebene pMC, Eigenschaft, und Region:

- *Zulässigkeitssynthese*: Existiert eine akzeptierende Belegung in einer Region?
- *Regionsverifikation*: Ist eine Region akzeptierend, beziehungsweise, sind alle Belegungen in einer Region akzeptierend?

Beispiel 4. *Wir betrachten weiterhin die pMC aus Abb. 2c, zusammen mit einer Eigenschaft φ, die bedingt, dass $\boxdot$ mit einer Wahrscheinlichkeit nicht kleiner als $3/20$ erreicht wird. Es existiert eine akzeptierende Belegung, zum Beispiel die Belegung für faire Münzen. Sei $R = \{\mathsf{val}\colon X \to \mathbb{R} \mid 1/10 \leq \mathsf{val}(x) \leq 9/10$ und $3/4 \leq \mathsf{val}(y) \leq 5/6\}$ eine Region und sei Eigenschaft $\varphi' := \neg\varphi$ die Eigenschaft, die bedingt, dass $\boxdot$ mit einer Wahrscheinlichkeit mehr als $3/20$ erreicht wird. Das Verifikationsproblem besteht nun darin, zu bestimmen, ob alle Belegungen in R φ' akzeptieren. Die Frage ist äquivalent zu der Frage, ob alle Belegungen in R die Eigenschaft φ ablehnen. Da alle Belegungen in R eine induzierte Erreichbarkeitswahrscheinlichkeit zu Zustand $\boxdot$ kleiner $3/20$ haben, muss das Verifikationsproblem mit wahr beantwortet werden. Demnach akzeptiert R die Eigenschaft φ' and lehnt R φ ab.*

Weitere Fragestellungen, die im Rahmen dieser Arbeit untersucht wurden sind: (1) Das *Berechnen einer geschlossenen Darstellung*: Wie sieht die *Lösungsfunktion* aus, die Belegungen auf die Erreichbarkeitswahrscheinlichkeit in den assoziierten MCs abbildet? (2) Die *Parameterraumpartitionierung*: Wie finden wir eine Darstellung (oder Approximation) aller akzeptierenden Belegungen?

Beispiel 5. *Wir betrachten weiterhin Bsp. 4. Die Lösungsfunktion $x\cdot(1-y)\cdot(1-x)/(1-x\cdot y)$ beschreibt die Wahrscheinlichkeit $\boxdot$ zu erreichen (unter der Annahme, dass $\mathsf{val}(x)$ und $\mathsf{val}(y)$ strikt zwischen 0 und 1 liegen). In Abb. 1b approximieren wir akzeptierende und ablehnende Belegungen. Die grüne (gepunktete) Fläche, als Vereinigung von kleineren rechteckigen akzeptierenden Regionen beschreibt eben jene Belegungen, bei denen die Eigenschaft φ erfüllt ist (durch die assoziierte MC), und die rote (gestrichene) Fläche stellt die ablehnenden Belegungen dar. Die weiße Fläche stellt den Teil des Parameterraums dar, für den ein (geeigneter) Algorithmus noch nicht entschieden hat, ob die Belegungen akzeptierend oder ablehnend sind.*

Diese Problemstellungen wurden in Vorarbeiten, zum Beispiel in [HHZ11, HBK17, Ba11, Ch13], betrachtet. Intuitiv sind diese Fragen komplex durch die Abhängigkeiten zwischen den verschiedenen Parametern und dem Auftreten von den Parametern an verschiedenen Transitionen in der pMC.

3 Parametersynthese ist ETR-vollständig

Die Komplexitätsklasse ETR (von *Existential Theory of the Reals*) enthält alle Probleme die in Polynomialzeit auf das Entscheidungsproblem ETR reduziert werden können.

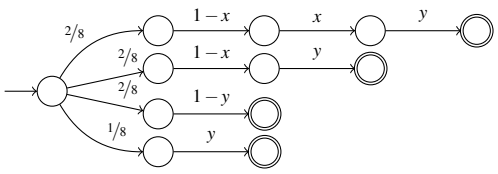


Abb. 3: Konstruktion zur ETR-Härte der Zulässigkeitssynthese (Beispiel)

Die Klasse umfasst die Klasse NP und ist in der Klasse PSPACE enthalten. Das Problem ETR besteht darin, zu entscheiden, ob ein gegebener existentiell-quantifizierter Satz $\exists x_1 \dots \exists x_n F(x_1, \dots, x_n)$ wahr ist, wobei F eine beliebige Boolesche Verknüpfung von polynomiellen Ungleichungen beschreibt, und $x_1, \dots x_n$ reellwertige Variablen sind.

Beispiel 6. *Betrachten wir die pMC in Abb. 4a, mit Region* $R = \{\mathsf{val} \mid 0 < \mathsf{val}(x) < 1, 0 < \mathsf{val}(y) < 1\}$*, sowie Zielzustand* s_3*. Es existiert eine Belegung sodass wir* s_3 *mit einer Wahrscheinlichkeit mindestens* $3/4$ *erreichen, genau dann, wenn der folgende Satz wahr ist:*

$$\exists p_0, \dots p_4, x, y : p_0 \geq 3/4 \wedge 0 < x < 1 \wedge 0 < y < 1 \wedge p_3 = 1 \wedge p_4 = 0 \tag{1}$$

$$s_1 = y \cdot s_2 + (1-y) \cdot s_3 \wedge s_2 = y \cdot s_1 + (1-y) \cdot s_4 \tag{2}$$

Für jeden Zustand s_i *führen wir eine Variable* p_i *ein, mit der intuitiven Bedeutung, dass sie die Erreichbarkeitswahrscheinlichkeit von Zustand* s_i *zum Zielzustand darstellt. Die Variablen* x, y *entsprechen den Parametern mit gleichem Namen. In* (1) *kodieren wir zunächst, dass die Wahrscheinlichkeit mindestens* $3/4$ *sein soll, und dass die Parameterwerte wie in* R *beschrieben sein sollen. Weiterhin wissen wir, dass vom Zielzustand* s_3 *sicherlich* s_3 *erreicht wird. Mit einem einfachen Graphalgorithmus bestimmen wir dazu, dass es keinen Pfad von* s_4 *zu* s_3 *gibt und die Wahrscheinlichkeit demnach* 0 *sein muss*[3]*. Die restlichen Gleichungen, in* (2)*, entsprechen einer rekursiven Beschreibung der Erreichbarkeitswahrscheinlichkeiten (oft Bellman-Gleichungen genannt). Die Länge des Satzes ist linear in der Größe des pMCs. Es ergibt sich somit eine polynomielle Reduktion.*

Bemerkenswerterweise funktioniert die Reduktion unter sehr milden Annahmen auch in die andere Richtung: Wir zeigen beispielhaft, wie wir aus einer beliebigen polynomiellen Ungleichung eine pMC erstellen, sodass deren Zielzustände mit einer gesammelten Wahrscheinlichkeit von mindestens λ erreicht werden, genau dann, wenn die Belegung die polynomielle Ungleichung erfüllt. Wichtige Vorarbeiten basieren auf [Ch17].

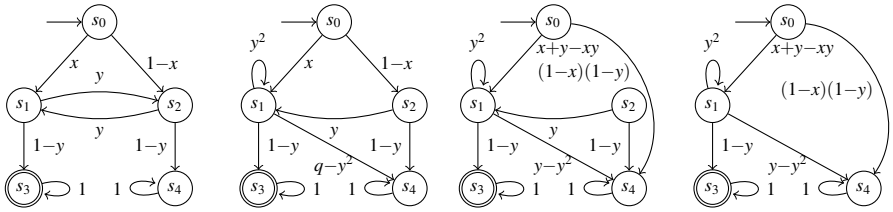
Beispiel 7. *Betrachten wir den Satz* $G = \exists x, y : -2x^2y + y \geq 5$*. Wir stellen die Ungleichung um, um den negativen Koeffizienten zu eliminieren und nach Ausmultiplizieren erhalten wir*

$$2 \cdot (1-x)xy + 2 \cdot (1-x)y + 2 \cdot (1-y) + y - 2 \geq 5.$$

In einem letzten Schritt addieren wir 2 *und teilen durch* 8*, sodass die Koeffizienten gemeinsam einer Verteilung entsprechen:*

$$2/8 \cdot (1-x)xy + 2/8 \cdot (1-x)y + 2/8 \cdot (1-y) + 1/8 \cdot y \geq 7/8$$

[3] Wenn y auch 1 werden darf, muss die Kodierung explizit um diese Graphensuche erweitert werden.



(a) Beispiel: pMC (b) $s_1 \to s_2$ eliminieren (c) $s_0 \to s_2$ eliminieren (d) s_2 unerreichbars

Abb. 4: Transitionseliminierung (Beispiel)

Aus dieser Ungleichung konstruieren wir nun Abb. 3, wobei die fehlende Wahrscheinlichkeitsmasse in eine Senke geleitet wird. Die Zielzustände werden genau dann mit einer Wahrscheinlichkeit mindestens $7/8$ erreicht, wenn der Satz G wahr ist.

Theorem 1. *Die Zulässigkeitssynthese ist ETR-vollständig.*

Das Theorem zeigt, dass dieses Problem im Allgemeinen schwer zu lösen ist und motiviert die weitere Suche nach einer guten Heuristik. Des Weiteren ist zu beachten, dass die Reduktion bereits für strukturell einfache pMCs funktioniert. Die Härte des Problems gilt demnach bereits fur viele praktisch relevante Subklassen.

4 Algorithmen zur Parametersynthese

Wir skizzieren Algorithmen, die den aktuellen Stand der Technik darstellen. Alle skizzierten Verfahren sind offen und frei verfügbar und wurden in der Dissertation empirisch evaluiert.

Berechnen der Lösungsfunktion

Der erste Ansatz, pMCs zu analysieren, basiert auf der Berechnung einer Lösungsfunktion. Eine Lösungsfunktion bildet Belegungen auf Erreichbarkeitswahrscheinlichkeiten ab, siehe Bsp. 5. Algorithmisch kann diese Abbildung aus den Gleichungen wie in Bsp. 6 durch Eliminierung der Zustandsvariablen berechnet werden. Tatsächlich eignet sich dazu ein Gauss-Algorithmus auf der Transitionsmatrix – allerdings zeigen andere Ansätze eine höhere Praxistauglichkeit. Die Herausforderung dabei besteht darin, dass wir über einem Polynomialring eliminieren und die Matrixeinträge demnach stark wachsen, was sich leider auch nicht ganz verhindern lässt. Praktisch hat sich die Zustandseliminierung (aus der Automatentheorie) durchgesetzt.

Beispiel 8. *Betrachten wir wieder die pMC in Abb. 4a. Wir eliminieren nun Zustand s_2. Dazu eliminieren wir erst die Transition $s_1 \to s_2$, siehe Abb. 4b, und dann die Transition $s_0 \to s_2$, siehe Abb. 4c. Damit ist Zustand s_2 nun unerreichbar und wir können den Zustand sowie die ausgehenden Kanten entfernen, siehe Abb. 4d. Nach dem Eliminieren von s_1 gibt es eine einzelne Kante zwischen Initial- und Zielzustand, die mit der Erreichbarkeitswahrscheinlichkeit beschriftet ist.*

Gleichungssystembasierte Ansätze zum Finden von Belegungen

Da die Zulässigkeitssynthese ein ETR-hartes Problem darstellt, gibt es nur wenig Hoffnung auf einen vollständigen und skalierbaren Algorithmus. Die Kodierung, wie in Beispiel 6 dargestellt, kann zwar von sogenannten SMT-Solvern gelöst werden. Allerdings skalieren diese Methoden bisher kaum zu interessanten Probleminstanzen. Da es zum Lösen der Zulässigkeitssynthese aber reicht, eine akzeptierende Belegung (nichtdeterministisch) zu raten und dann zu verifizieren, und diese Verifikation der Analyse einer (parameter-freien) MC entspricht, fokussieren sich einige Ansätze auf eine Suche im Parameterraum. Es hat sich herausgestellt, dass es wichtig ist, dabei einen Teil der Problemformulierung einfließen zu lassen, da ein blindes Suchen in einem hochdimensionalen Raum natürlich weiterhin ein schweres Problem ist.

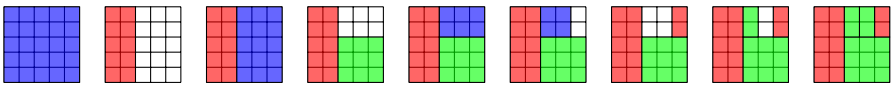
Wir vereinigen die beiden Sichtweisen wie folgt: Wir raten eine (beliebige) Belegung und linearisieren dann[4] um diese Belegung herum, um eine Belegung mit einer höheren Erreichbarkeitswahrscheinlichkeit zu finden. Dieses Verfahren iterieren wir, bis wir eine akzeptierende Belegung finden. Technisch erstellen wir ein Quadratisch-Beschränktes Quadratisches Optimierungsproblem, das wir durch eine Konvex-Konkavprozedur lösen. Das numerische Lösungsverfahren, das dabei zum Einsatz kommt, ist jedoch oft numerisch instabil. Wir binden deshalb die dedizierte Analyse von MCs so ein, dass sich ein effektives und effizientes Verfahren ergibt.

Gegenbeispielgelenkte Synthese

Die gegenbeispielgelenkte Synthese stellt die elementare Frage, ob wir von der Analyse einer ablehnenden Belegung schließen können, dass andere Belegungen auch ablehnend sein müssen, ohne eben jene andere Belegungen explizit zu betrachten. Wir betrachten hier einen Algorithmus, der für eine endliche Menge von Belegungen und zur Beantwortung der Zulässigkeitssynthese konzipiert ist. Weiterhin nehmen wir hier an, dass die betrachtete Eigenschaft erfordert, dass die Erreichbkarkeitswahrscheinlichkeit in einer induzierten MC kleiner als λ sein muss.

Wir bemerken erneut, dass das Finden einer akzeptierenden Belegung unser Problem lösen würde. Wir betrachten also den Fall, bei dem die geratene Belegung ablehnend ist. In diesem Fall ist die Erreichbarkeitswahrscheinlichkeit in der induzierten MC zu groß. Ein *Gegenbeispiel* besteht nun aus einer Teilmenge der Zustände der pMC, sodass für die gegebene Belegung val auf der assoziierten Teil-MC die Erreichbarkeitswahrscheinlichkeit auch λ überschreitet. Wenn in diesem Teilgraphen ein Parameter x nicht auftaucht, so können wir uns sicher sein, dass ein alleiniges Variieren von $\mathsf{val}(x)$ nicht ausreichen wird, um eine akzeptierende Belegung zu finden. Daraus folgt, dass wir nun eine Menge von Belegungen ausschließen können.

[4] Tatsächlich konvexifizieren wir in der Methode in der Dissertation.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Abb. 5: Verfeinerung in einer abstraktionsgelenkten Synthese zur Parameterraumpartitionierung

Abstraktionsgelenkte Synthese

Die wichtige Voraussetzung für die abstraktionsgelenkte Synthese ist die Wahl einer Abstraktion des Problems, die einfacher zu lösen ist: Die Komplexität der Zulässigkeitssynthese nimmt drastisch ab, wenn jeder Parameter nur an ausgehenden Transitionen von einem einzigen Zustand auftritt. Insbesondere kann der optimale Wert dann lokal entschieden werden, zum Beispiel durch eine Reduktion auf einen MDP und durch die effizienten Analyse dieses MDPs. Insbesondere gilt: Wenn es für die abstrakte Quotienten-pMC keine akzeptierende Belegung gibt, kann es auch keine akzeptierende Belegung in der Original-pMC geben. Dies rechtfertigt die Darstellung als Abstraktion. Die zweite Voraussetzung für die Abstraktionsgelenkte Synthese ist die Möglichkeit, zu verfeinern. In diesem Fall verfeinern wir den Parameterraum. Solche Verfeinerungen für den Knuth-Yao Würfel ergeben Abb. 1b.

Beispiel 9. *Wir skizzieren den Ansatz in Abb. 5. Das Gitter zeigt skizzenhaft 2 Parameter mit jeweils 5 Zuweisungen, d.h., jede Zelle korrespondiert zu einer Belegung. Blaue Regionen entstehen, wenn die Region sowohl akzeptierende und ablehnende Belegungen umfasst (in der Abstraktion), grün/rot zeigen akzeptierende und ablehnende Regionen. Wir betrachten die vollständige Region (blau), und bilden die Quotienten-pMC $\mathscr{D}'$. Die Analyse auf dieser pMC ergibt, dass es sowohl akzeptierende als auch ablehnende Belegungen für $\mathscr{D}'$ gibt. Wir teilen den Parameterraum auf. Eine Betrachtung der linken Region ergibt, dass alle Belegungen in $\mathscr{D}'$ ablehnend sind. Demnach sind sie auch ablehnend für $\mathscr{D}$. Wir iterieren auf diese Art weiter. Zu beachten ist, dass in Schritt (f) zwar alle Belegungen akzeptierend sind, wir das jedoch aus der Analyse der Quotienten-pMC nicht folgern können.*

5 Verbindung zu partiell beobachtbaren MDPs

MDPs sind weitverbreitet zur Analyse von Handlungsplanungsproblemen, bei denen sich eine Planungsentscheidung auf die Dynamik des Systems auswirkt. In MDPs kann der Planer sich jedem Zustand lokal zum Ausführen einer bestimmten Aktion entscheiden, d.h, ein Plan bildet Zustände auf Aktionen ab. Um einen solchen Plan umzusetzen, muss man allerdings den Zustand beobachten können. Diese Vorraussetzung ist in vielen Systemen nicht gegeben. Stattdessen erlauben diese Systeme nur eine partielle Beobachtbarkeit des Zustands. Um diesem Fall gerecht zu werden, wurden MDPs zu *partially observable MDPs* (POMDPs) [RN10] erweitert, indem jeder Zustand mit einer Beobachtung markiert wird.

Ein (optimaler) Plan in einem POMDP ist eine Abbiluding von einer Folge von Beobachtungen auf eine Aktion. Da solche Pläne unendlich groß werden können, sind Pläne mit beschränktem Speicher interessant [Me99]. Wir betrachten hier nur deren einfachste

(a) POMDP (b) pMC

Abb. 6: Von POMDPs zu pMCs.

Variante, sogenannte speicherlose Pläne. Ein speicherloser Plan für einen POMDP bildet Beobachtungen auf Verteilungen über Aktionen ab, d.h., man wählt bei jeder aktuellen Beobachtung durch einen passenden, gebiasten Münzwurf die nächste Aktion. Solche Pläne werden oft auch durch Bestärkendes Lernen konstruiert. Dieses Problem ist eng mit der Zulässigkeitssynthese in pMCs verwandt.

Beispiel 10. *Abb. 6a zeigt einen POMDP, wobei die Farben an den Zuständen die Beobachtung darstellen und die Buchstaben den Aktionen entsprechen. Beim Beobachten von einer grünen Beobachtung in Zustand s_1 oder s_3, wählt ein Plan mit einer gewissen Wahrscheinlichkeit q_1 Aktion α_1 und mit einer Wahrscheinlichkeit q_2 Aktion α_2. Die Wahrscheinlichkeiten p_1, p_2, p_3 bestimmen die Verteilung bei der Beobachtung in s_0. Tatsächlich liefert diese Sichtweise direkt eine pMC, wie gezeigt in Abb. 6b.*

6 Schluss

Im Rahmen der Dissertation wurde gezeigt, dass die Parametersynthese für Markov-Modelle ein vielversprechender Ansatz ist, um eine Reihe praktischer Probleme anzugehen. Obwohl das Problem theoretisch schwer ist, gibt es bereits verschiedene Ansätze, um dennoch gute Ergebnisse zu erzielen. Wir sehen natürlich weiterhin Möglichkeiten, den Stand der Technik in Zukunft zu verbessern, insbesondere durch interdisziplinäre Ansätze, die die neuesten Erkenntnisse aus der Probabilistischen Inferenz und der Robotik nutzen.

Danksagung. Ich möchte mich bei meinem Doktorvater Joost-Pieter Katoen, sowie bei meinen zahlreichen Ko-Autoren bedanken, mit denen ich jahrelang an der Parametersynthese geforscht habe. Zusätzlich bedanke ich mich bei Nils Jansen sowie Angelika und Christian Junges für die Unterstützung beim Erstellen dieser Zusammenfassung.

Literaturverzeichnis

[Ba11] Bartocci, Ezio; Grosu, Radu; Katsaros, Panagiotis; Ramakrishnan, C. R.; Smolka, Scott A.: Model Repair for Probabilistic Systems. In: TACAS. Jgg. 6605 in LNCS. Springer, S. 326–340, 2011.

[Ce19a] Ceska, Milan; Hensel, Christian; Junges, Sebastian; Katoen, Joost-Pieter: Counterexample-Driven Synthesis for Probabilistic Program Sketches. In: FM. Jgg. 11800 in LNCS. Springer, S. 101–120, 2019.

[Ce19b] Ceska, Milan; Jansen, Nils; Junges, Sebastian; Katoen, Joost-Pieter: Shepherding Hordes of Markov Chains. In: TACAS. Jgg. 11428 in LNCS. Springer, S. 172–190, 2019.

[Ch13] Chen, Taolue; Hahn, Ernst Moritz; Han, Tingting; Kwiatkowska, Marta Z.; Qu, Hongyang; Zhang, Lijun: Model Repair for Markov Decision Processes. In: TASE. IEEE CS, S. 85–92, 2013.

[Ch17] Chonev, Ventsislav: Reachability in Augmented Interval Markov Chains. CoRR, abs/1701.02996, 2017.

[Cu18] Cubuktepe, Murat; Jansen, Nils; Junges, Sebastian; Katoen, Joost-Pieter; Topcu, Ufuk: Synthesis in pMDPs: A Tale of 1001 Parameters. In: ATVA. Jgg. 11138 in LNCS. Springer, S. 160–176, 2018.

[De15] Dehnert, Christian; Junges, Sebastian; Jansen, Nils; Corzilius, Florian; Volk, Matthias; Bruintjes, Harold; Katoen, Joost-Pieter; Ábrahám, Erika: PROPhESY: A PRObabilistic ParamEter SYnthesis Tool. In: CAV. Jgg. 9206 in LNCS. Springer, S. 214–231, 2015.

[HBK17] Hutschenreiter, Lisa; Baier, Christel; Klein, Joachim: Parametric Markov Chains: PCTL Complexity and Fraction-free Gaussian Elimination. In: GandALF. Jgg. 256 in EPTCS, S. 16–30, 2017.

[HHZ11] Hahn, Ernst Moritz; Hermanns, Holger; Zhang, Lijun: Probabilistic reachability for parametric Markov models. STTT, 13(1):3–19, 2011.

[Ju18] Junges, Sebastian; Jansen, Nils; Wimmer, Ralf; Quatmann, Tim; Winterer, Leonore; Katoen, Joost-Pieter; Becker, Bernd: Finite-State Controllers of POMDPs using Parameter Synthesis. In: UAI. AUAI Press, S. 519–529, 2018.

[Ju20] Junges, Sebastian: Parameter synthesis in Markov models. Dissertation, RWTH Aachen University, Germany, 2020.

[Me99] Meuleau, Nicolas; Kim, Kee-Eung; Kaelbling, Leslie Pack; Cassandra, Anthony R.: Solving POMDPs by Searching the Space of Finite Policies. In: UAI. Morgan Kaufmann, S. 417–426, 1999.

[Qu16] Quatmann, Tim; Dehnert, Christian; Jansen, Nils; Junges, Sebastian; Katoen, Joost-Pieter: Parameter Synthesis for Markov Models: Faster Than Ever. In: ATVA. Jgg. 9938 in LNCS. Springer, S. 50–67, 2016.

[RN10] Russell, Stuart J.; Norvig, Peter: Artificial Intelligence – A Modern Approach (3. ed.). Pearson Education, 2010.

[Wi19] Winkler, Tobias; Junges, Sebastian; Pérez, Guillermo A.; Katoen, Joost-Pieter: On the Complexity of Reachability in Parametric Markov Decision Processes. In: CONCUR. Jgg. 140 in LIPIcs. Schloss Dagstuhl - LZI, S. 14:1–14:17, 2019.

Sebastian Junges wurde 1991 in den Niederlanden geboren und machte dort sein Abitur. Sein Studium absolvierte er an der RWTH Aachen, und schrieb dort seine Masterarbeit zur Fehlzustandsbaumanalyse. Nach dem Studium promovierte er an der RWTH Aachen im Bereich der Verifikation bei Prof. Dr. Joost-Pieter Katoen. Er ist einer der Entwickler des Model-Checkers Storm. Nach Abschluss der Promotion forscht er nun als Postdoctoral Researcher an der University of California im kalifornischen Berkeley zu Autonomie und Sicherheit von komplexen Systemen.

Formale Verifikation von Multiplizierern mit Computeralgebra [1]

Daniela Kaufmann[2]

Abstract: Arithmetische Schaltungen werden in Prozessoren zur Implementierung von Boolescher Algebra genutzt. Aufgrund des weitreichenden Einsatzes von Prozessoren ist es äußerst wichtig, die Korrektheit dieser Schaltungen garantieren zu können, um Fehler wie den berühmten Pentium FDIV-Bug zu vermeiden. Mithilfe formaler Verifikation kann festgestellt werden, ob eine Schaltung ihrer gewünschten Spezifikation entspricht. Allerdings stellen arithmetische Schaltungen, insbesondere Integer-Multiplizierer auf Gatterebene, eine Herausforderung für bestehende Verifikationstechniken dar. In dieser Dissertation [Ka20] werden aktuelle Verifikationsmethoden basierend auf Computeralgebra verbessert. Wir zeigen eine rigorose präzise mathematische Formulierung, welche auch die Anwendung der Mathematik in diesem Gebiet erweitert. Außerdem haben wir neue Methoden zur vollautomatischen Verifikation von Integer-Multiplizierern entworfen und implementiert, sowie ein kompaktes Beweisformat entwickelt, um das Ergebnis der Verifikation zertifizieren zu können.

1 Einleitung

Digitale Schaltungen sind ein wesentlicher Bestandteil von Computern und digitalen Systemen. Damit werden mittels binären digitalen Signalen logische Operationen ausgeführt und sie können zahlreiche digitale Komponenten und arithmetische Operationen modellieren. Die Grundfunktion einer digitalen Schaltung ist es aus gegebenen binären Eingangssignalen ein binäres Signal an den Ausgängen für die implementierte logische Funktion zu generieren. Die Hauptkomponenten digitaler Schaltungen sind Logikgatter, welche einfache Boolesche Funktionen repräsentieren. Beispiele für diese Gatter sind Negation (NOT), Konjunktion (UND) und Disjunktion (ODER). Diese logischen Gatter können kombiniert werden, um komplexere Funktionen darzustellen.

Eine Unterkategorie von digitalen Schaltungen sind Schaltnetze. In diesen hängt das Ausgangssignal nur von den gegebenen Eingängen ab, das heißt es gibt keine Rückkopplung der Ausgänge auf die Eingänge. Schaltnetze werden in Computern eingesetzt um Boolesche Algebra zu implementieren. Zum Beispiel ist die arithmetisch-logische Einheit (ALU) in einem Prozessor, die zur Berechnung von mathematischen Operationen eingesetzt wird, aus Schaltnetzen konstruiert. Arithmetische Schaltungen sind spezielle Schaltnetze, die Rechenoperationen wie zum Beispiel Multiplikation durchführen.

Aufgrund des weitreichenden, mitunter sicherheitskritischen, Einsatzes von digitalen Schaltungen ist es äußerst wichtig ihre Korrektheit sicherzustellen. Mittels formaler Verifikation lässt sich die Korrektheit von Software oder Hardware in Bezug auf eine zuvor definierte

[1] Englischer Titel der Dissertation: “Formal Verification of Multiplier Circuits using Computer Algebra”

[2] Johannes Kepler Universität Linz, daniela.kaufmann@jku.at

Spezifikation beweisen. Dazu wird das gegebene System in ein mathematisches Modell übersetzt und automatisierte Beweistechniken werden eingesetzt, um die gewünschten Korrektheitsbeweise zu erzielen. Die verschiedenen Verifikationstechniken unterscheiden sich durch die zugrundeliegenden mathematischen Modellierungen des Systems.

Formale Verifikation von arithmetischen Schaltungen kann helfen Fehler, wie den berühmten FDIV-Bug in frühen Intel Pentium Prozessoren, zu finden und zu vermeiden. Die Gleitkommaeinheit (FPU) dieser Prozessoren enthält einen Hardwarefehler, der bei Division von bestimmten Zahlen zu falschen Ergebnissen führt. Der Fehler wurde erst 1994, rund eineinhalb Jahre nach Markteinführung des Prozessors bekannt [SB94] und kostete Intel ca. 500 Millionen US-Dollar. Eine gleichartige Panne würde heutzutage ernsthafte finanzielle Schwierigkeiten, auch für große Prozessorhersteller, bedeuten. Daher wollen diese Firmen die Korrektheit ihrer Hardware zu hundert Prozent garantieren können.

Seit dem Bekanntwerden des FDIV-Bugs werden formale Verifikationstechniken entwickelt, um die Korrektheit von arithmetischen Schaltungen zu beweisen. Mehr als 25 Jahre später ist dieses Verifikationsproblem jedoch noch immer nicht vollautomatisiert lösbar. Besonders Integer-Multiplizierer, das sind arithmetische Schaltungen die Multiplikationen von ganzen Zahlen ausführen, stellen aufgrund ihres internen Aufbaus der Logikgatter eine Herausforderung für bestehende Verifikationsmethoden dar.

In der Praxis heißt das, dass industrielle Entwickler von arithmetischen Schaltungen momentan entweder aufwändige manuelle Verifikation mittels Theorembeweisern betreiben, oder sich komplett auf Simulationen verlassen. Immer bessere Optimierungen in der Entwicklung erhöhen einerseits die Effizienz einer Schaltung, steigern andererseits aber deren Komplexität. Dadurch sinkt der Prozentsatz der Werte, die simuliert werden können. Daher gelten Simulationen in der Praxis nicht mehr als vertrauenswürdig. Das Fehlen von vollautomatisierten Verifikationsmethoden für arithmetische Schaltungen ist aktuell immer noch ein großer Makel. An diesem Punkt verbessert diese Dissertation den Stand der Technik.

In dieser Dissertation [Ka20] werden formale Verifikationstechniken, welche auf Computeralgebra basieren, untersucht und verbessert. Das Ziel ist es, eine Verifikationsmethode zu erhalten, die für einen gegebenen Integer-Multiplizierer auf Gatterebene vollautomatisiert über dessen Korrektheit entscheidet, ohne dass der Entwickler manuell in den Verifikationsprozess eingreifen muss. Wir zeigen eine präzise mathematische Formalisierung dieses Problems und beweisen die Korrektheit und Vollständigkeit dieser Methode. Neue, in der Dissertation entwickelte, Algorithmen ermöglichen die Verifikation von komplexen Multiplizierer-Architekturen mit Bitbreiten von bis zu 2 048 Bits. Alle entstandenen implementierten Tools und Benchmarks sind als Open-Source verfügbar. Um das Verifikationsresultat des Verifikations-Tools validieren zu können, haben wir ein kompaktes Beweisformat entwickelt, welches erlaubt während der Verifikation einfache Beweiszertifikate zu generieren, die in einem eigenständigen Beweis-Checker validiert werden können. Während meiner Dissertation habe ich als Erstautor zu zehn Publikationen beigetragen und für meine erste Publikation [RBK17] den Best Paper Award bei FMCAD 2017, der wichtigsten Konferenz in Hardwareverifikation, erhalten. Sechs dieser Publikationen sind vollumfänglich in meiner kumulativen Dissertation enthalten.

2 Hintergrund

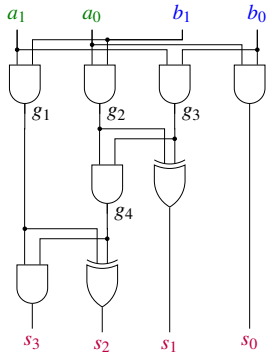
Seit der Entdeckung des FDIV-Bugs werden verschiedenste Verifikationstechniken zur Validierung der Korrektheit von Multiplizierern eingesetzt. Eine gängige Methode ist das Verifikationsproblem als ein Erfüllbarkeitsproblem der Aussagenlogik (SAT) zu kodieren. Bei dieser Methode wird die Schaltung zu einer Formel in konjunktiver Normalform (KNF) übersetzt und ein SAT-Solver eingesetzt, um die Erfüllbarkeit dieser Formel zu überprüfen. Im Jahr 2016 wurde eine größere Menge von solchen Kodierungen für arithmetische Schaltungen als Benchmarks zur jährlichen SAT-Competition eingereicht [Bi16]. Bei diesem Bewerb werden die aktuell besten SAT-Solver anhand einer Menge von Benchmarks ermittelt. Die Ergebnisse dieser Evaluierung zeigen allerdings, dass KNF-Kodierungen von Multiplizierern zu exponentiell großen Beweisen führen, dies deutet auf eine exponentielle Laufzeit von SAT-Solvern hin. Der Grund dafür ist, dass Multiplizierer aus Sequenzen von XOR-Gattern aufgebaut sind und diese mit aktuellen Lösungsmethoden in SAT-Solvern nicht effizient behandelt werden können. In der Theorie konnte die Notwendigkeit der exponentiellen Beweisgröße für simple Multiplizierer-Architekturen bereits widerlegt werden [BL17]. Dieses theoretische Resultat setzt auf strukturelle Kenntnisse über den Multiplizierer und konnte noch nicht praktisch in einem SAT-Solver implementiert werden.

Die erste Technik, mit der es möglich war, den FDIV-Bug zu vermeiden, basiert auf binären Entscheidungsdiagrammen. Genauer gesagt auf Binary Moment Diagrams (BMD) [CB95], da die Knotenanzahl in diesen Diagrammen linear in der Bitbreite bleibt. Jedoch benötigt diese Technik strukturelles Wissen über die Schaltung, da BMDs in einer vorbestimmten Ordnung aufgebaut werden müssen, um die lineare Größe zu garantieren. Daher kann diese Methode nicht vollautomatisiert in komplexen industriellen Designs eingesetzt werden.

Theorembeweiser, wie zum Beispiel ACL2 [Hu17], können in Kombination mit SAT-Solvern industrielle Multiplizierer beweisen. Jedoch ist diese Methode auch nicht vollautomatisiert einsetzbar. Das zu Grunde liegende Beweissystem baut auf einer Menge von problemspezifischen Axiomen und Inferenzregeln auf, und benötigt daher Hintergrundwissen über die Domäne des Problems. Termersetzungssysteme [Va07] benötigen gleichermaßen Hintergrundwissen über den Definitionsbereich und deren Anwendung auf dieses Problem ist auch nicht vollautomatisiert.

Methoden basierend auf Reverse-Engineering [SK04] nutzen arithmetische Darstellungen auf der Bitebene, welche von den gegebenen Multiplizierern auf Gatterebene extrahiert werden. Diese Technik erlaubt es einfache strukturierte Multiplizierer vollautomatisiert zu beweisen, scheitert jedoch an komplexeren industriellen Multiplizierer-Architekturen.

Seit 2016 werden Verifikationstechniken basierend auf Computeralgebra [Yu16, Sa16] als ein vielversprechender Zugang zur automatisierten Verifikation von arithmetischen Schaltungen und insbesondere Multiplizierern gesehen. In dieser Methode werden alle Logikgatter der Schaltung, sowie die Spezifikation als ein Polynom kodiert. Mittels algebraischer Reduktion basierend auf multivariater Polynomdivision kann nun das Problem, ob eine gegebene Schaltung einen korrekten Multiplizierer implementiert, gelöst werden, indem die Normalform des Spezifikationspolynoms bezüglich der Gatterpolynome berechnet wird. Der Multiplizierer ist fehlerfrei genau dann, wenn diese Normalform null ist.



Gatterpolynome	$s_3 = g_1 \wedge g_4$	$-s_3 + g_4 g_1,$
	$s_2 = g_1 \oplus g_4$	$-s_2 - 2g_4 g_1 + g_4 + g_1,$
	$g_4 = g_2 \wedge g_3$	$-g_4 + g_2 g_3,$
	$s_1 = g_2 \oplus g_3$	$-s_1 - 2g_2 g_3 + g_2 + g_3,$
	$g_1 = a_1 \wedge b_1$	$-g_1 + a_1 b_1,$
	$g_2 = a_0 \wedge b_1$	$-g_2 + a_0 b_1,$
	$g_3 = a_1 \wedge b_0$	$-g_3 + a_1 b_0,$
	$s_0 = a_0 \wedge b_0$	$-s_0 + a_0 b_0,$
Binäre Eingänge	$a_1, a_0 \in \mathbb{B}$	$-a_1^2 + a_1, -a_0^2 + a_0,$
	$b_1, b_0 \in \mathbb{B}$	$-b_1^2 + b_1, -b_0^2 + b_0$
Spezifikation	$8s_3 + 4s_2 + 2s_1 + s_0 - 4a_1 b_1 - 2a_1 b_0 - 2a_0 b_1 - a_0 b_0 = 0$	

Abb. 1: 2-Bit Multiplizierer (links) und dessen algebraische Modellierung (rechts).

3 Beitrag der Dissertation

In dieser Dissertation werden formale Verifikation von Multiplizierern auf Gatterebene mittels Computeralgebra untersucht und verbessert, sodass diese Technik auch für komplexe Multiplizierer eingesetzt werden kann. Wir betrachten n-Bit Multiplizierer mit Eingangssignalen $a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1} \in \{0,1\}$ und $2n$ Ausgängen $s_0, \ldots, s_{2n-1} \in \{0,1\}$. Eine Schaltung ist ein korrekter Multiplizierer genau dann, wenn für alle möglichen Eingänge $a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1} \in \{0,1\}$ und den von der Schaltung berechneten Ausgangssignalen $s_0, \ldots, s_{2n-1} \in \{0,1\}$ die Spezifikation $\sum_{i=0}^{2n-1} 2^i s_i - \left(\sum_{i=0}^{n-1} 2^i a_i\right)\left(\sum_{i=0}^{n-1} 2^i b_i\right) = 0$ gilt. In Worten gefasst bedeutet dies, dass der Ausgangsbitvektor gleich dem Produkt der beiden Eingangsbitvektoren ist. Man beachte, dass die Signale in der Schaltung binär sind, die Spezifikation aber als ein Polynom über den ganzen Zahlen dargestellt wird.

Die Komplexität eines Multiplizierers hängt von seiner Architektur, das heißt, der Anordnung und dem Aufbau der internen Logikgatter, ab. Im Allgemeinen werden Multiplizierer in drei Komponenten unterteilt. In der ersten Komponente werden partielle Produkte $a_i b_j$, für $0 \leq i, j < n$ aus den Eingangssignalen generiert. Dies kann entweder mittels (einer quadratischen Anzahl von) UND-Gattern geschehen oder auch durch eine komplexere Booth-Kodierung. In der zweiten Komponente werden diese partiellen Produkte mit Addierwerken, bestehend aus Voll- und Halbaddierern, akkumuliert und zu zwei Ebenen reduziert. Diese beiden Ebenen werden in einem *abschließenden Addierwerk* aufsummiert, welches das Ausgangssignal des Multiplizierers berechnet. In der Praxis werden die einzelnen Komponenten optimiert, um Multiplizierer platzsparender aufbauen zu können und die Berechnungsdauer der Ausgangssignale zu verringern.

Zu Beginn dieser Dissertation haben wir eine einfache und präzise mathematische Formalisierung für die Verifikation von arithmetischen Schaltungen definiert und umfassende Beweise für die *Korrektheit und Vollständigkeit* dieser Methode gezeigt. Obwohl die allgemeine Idee, arithmetische Schaltungen mit Hilfe von Polynomreduktion zu verifizieren bereits existierte [Yu16, Sa16], wurden diese entscheidenden Eigenschaften der Methode noch nicht belegt.

Die allgemeine Idee algebraischer Verifikation ist es für jedes Logikgatter ein Polynom zu definieren, welches die möglichen Signalkombinationen des Gatters modelliert. Für die Polynomkodierung muss ein zugrundeliegender Polynomring $R[X]$ festlegt werden. Hierfür haben wir zu Beginn dieser Dissertation die Menge der rationalen Zahlen $\mathbb{Q}$ als den Koeffizientenring R gewählt, da diese einen Körper bilden. Abbildung 1 zeigt die vollständige algebraische Kodierung eines 2-Bit Multiplizierers auf Gatterebene. Zum Beispiel impliziert das UND-Gatter $s_3 = g_1 \wedge g_4$ in Abb. 1 die Gleichung $-s_3 + g_4 g_1 = 0$ für die binären Signale s_3, g_1 und g_4. Zusätzlich wird für jedes Eingangssignal a_i (b_j) eine binäre Eingangsbedingung $-a_i^2 + a_i = 0$ ($-b_j^2 + b_j = 0$) generiert, welche definiert, dass diese Signale nur die Werte 0 und 1 annehmen können.

Es soll nun gezeigt werden, dass die Spezifikation aus den Gatterpolynomen und den binären Eingangsbedingungen folgt. Diese Implikation bedeutet algebraisch, dass die Spezifikation im Ideal, welches durch die Gatterpolynome und den binären Eingangsbedingungen erzeugt wird, enthalten ist. Die Idealzugehörigkeit der Spezifikation lässt sich eindeutig mit Hilfe der Theorie von Gröbnerbasen zeigen. Es gelten folgende wichtige Eigenschaften: Für jedes Ideal kann eine Gröbnerbasis ermittelt werden und danach steht einen Reduktionsalgorithmus zur Berechnung der Normalform zur Verfügung, der es uns erlaubt mittels wiederholter Polynomdivision die Frage der Idealzugehörigkeit eindeutig beantworten zu können.

Da wir zunächst als Koeffizientenring den Körper $\mathbb{Q}$ gewählt haben, können wir die Standardtheorie der Gröbnerbasen nutzen. Die Berechnung einer Gröbnerbasis für ein Ideal ist jedoch EXPSPACE-vollständig, und daher für größere Probleme praktisch nicht anwendbar. Wir können allerdings zeigen, dass die Berechnung einer Gröbnerbasis für unsere Anwendung nicht notwendig ist. Werden die Terme in den Polynomen anhand einer lexikographischen Variablenordnung sortiert, sodass die Ausgangsvariable eines Logikgatters immer größer als die Variablen der Eingänge ist, bilden die geordneten Gatterpolynome und binären Eingangsbedingungen automatisch eine Gröbnerbasis.

Auch nach dem Berechnen der Gröbnerbasis bleibt das Problem der Idealzugehörigkeit co-NP-hart [Ka20]. Experimente zeigen, dass bereits für einfache 8-Bit Multiplizierer die Normalform der Spezifikation nicht mehr in gängigen Computeralgebrasystemen ermittelt werden kann. Der Grund dafür ist, dass die Zwischenresultate der einzelnen Polynomdivisionen exponentiell anwachsen, bevor sie zu null reduzieren. Wir untersuchten mögliche Auslöser für diesen Wachstum und erlangten folgende Erkenntnisse. Zum einen wächst das Spezifikationspolynom quadratisch mit der Bitbreite des Multiplizierers. Zum anderen sind Voll- und Halbaddierer wiederkehrende Bausteine in den Multiplizierern, welche genutzt werden um drei bzw. zwei Bits aufzusummieren. Die Funktion dieser Addierer lässt sich mit einem einzelnen linearen Polynom kodieren. In der von der Schaltung induzierten Gröbnerbasis ist diese Polynom jedoch nicht enthalten. Die Gröbnerbasis enthält nur die internen nichtlinearen Gatterpolynome der Addierer, was zu einem exponentiellen Wachstum der Zwischenresultate im Reduktionsalgorithmus führt.

Um das Wachstum der Zwischenresultate zu umgehen, präsentieren wir einen neuen *inkrementellen Verifikationsalgorithmus* [RBK17]. Dieser Algorithmus beruht auf der Erkenntnis, dass die partiellen Produkte in einem Multiplizierer eindeutig in $2n$ Spalten teilbar

sind und jede Spalte maximal eine lineare Anzahl von partiellen Produkten enthält. Wir können dadurch für jede Spalte eine eigene Spezifikation festlegen. Wir erhalten eine Spaltenzerlegung automatisch aus den Einflussbereichen der Ausgangssignalen (und weit einfacher als eine Zeilenzerlegung, die zum Beispiel in BMDs notwendig ist und nicht notwendigerweise existiert). Die Korrektheit des Multiplizierers wird bewiesen, indem jede Spalte inkrementell auf ihre Korrektheit überprüft wird. Der Vorteil dieser Technik ist, dass das Gesamtproblem in mehrere kleinere Teilprobleme aufgeteilt wird und nicht mehr die gesamte quadratische Spezifikation im Reduktionsalgorithmus benötigt wird.

Des Weiteren betreiben wir *Preprocessing*, um eine reduzierte und kompaktere Darstellung [KBK20b] des Multiplizierers zu erhalten. Unsere Preprocessing-Technik fußt auf der Eliminationstheorie von Gröbnerbasen und erlaubt es, Variablen aus der Gröbnerbasis zu entfernen. In unserer ersten Version [KBK20b] suchen wir gezielt nach Voll- und Halbaddierern im gegebenen Multiplizierer und eliminieren deren interne Variablen. Dadurch enthält die Gröbnerbasis nun die gewünschten linearen Polynome, welche die Funktion dieser Addierer beschreiben. Für diese Optimierung führen wir die nötigen theoretischen Grundlagen ein und zeigen, dass in unserer Anwendung lokale Teile der Gröbnerbasis umgeschrieben werden können und die verbliebenen Polynome immer noch eine Gröbnerbasis bilden. Ohne dieses Ergebnis wäre es an dieser Stelle notwendig eine Gröbnerbasis für eine neue Variablenordnung zu berechnen, was uns wieder zu einem EXPSPACE-vollständigen Problem führen würde.

Eine gängige Darstellung von Schaltungen auf Gatterebene sind UND-Invertierer-Graphen (AIG). Wir haben ein Tool implementiert [RBK17, KBK20b], welches Multiplizierer im AIG-Format einliest, die Polynomkodierung übernimmt und die nötigen Preprocessing-Schritte durchführt. Für die eigentliche Berechnung der Normalform benutzen wir die Computeralgebrasysteme Mathematica und Singular. Unsere Experimente zeigen, dass diese Techniken einfache Multiplizierer, die komplett aus Voll- und Halbaddierern aufgebaut sind, verifizieren können, aber bei komplexeren Multiplizierern fehlschlagen.

Wie bereits beschrieben, wird in komplexen Architekturen beispielsweise eine Booth-Kodierung eingesetzt, um die Anzahl der Logikgatter für die Berechnung der partiellen Produkte zu reduzieren. Die Polynomrepräsentation dieser Kodierung hat im algebraischen Verifikationssystem zur Folge, dass in den Zwischenresultaten Terme mit Koeffizienten entstehen, die größer als 2^{2n} sind, welche durch den Ausgangsbitvektor nicht mehr abgedeckt sind. Wir können aufgrund des gewählten Körpers $\mathbb{Q}$ als Koeffiezientenrings diese Terme jedoch nicht kürzen, sondern müssen abwarten, bis sich diese Terme im Reduktionsalgorithmus durch Polynomdivision reduzieren. Darauf basierend erweitern wir unsere bisherigen theoretischen Ergebnisse und Beweise für Polynomringe über allgemeinere Koeffiezientenringe, sodass auch *modulares Kürzen* möglich ist [KBK19]. Modulares Beweisen ermöglicht es nicht nur, dass wir jetzt auch komplexe Integer-Multiplizierer erfolgreich verifizieren können, sondern auch abgeschnittene Multiplizierer behandeln können, bei denen die höchstwertigen Ausgangsbits vernachlässigt werden. Diese Multiplizierer finden zum Beispiel in der Satisfiability Modulo Theory Library (SMT-Lib) Verwendung und entsprechen der Integer-Multiplikation in gängigen Programmiersprachen wie C und Java.

Zusätzlich verallgemeinern wir unsere bestehenden Preprocessing-Techniken, sodass diese unabhängig von syntaktischen Mustern in der Schaltung sind. Einfache Veränderungen in der Gatteranordnung führen zu Fehlschlägen unseres bisherigen Preprocessings. Unsere neue Preprocessing-Methode eliminiert wiederholt Variablen, die nur in exakt einem anderen Polynom vorkommen. Wir zeigen durch einen einfachen Ansatz, dass dadurch unsere erste Preprocessing-Version subsumiert wird.

Nichtsdestotrotz sind die *abschließenden Addierwerke* von komplexen Multiplizierern, im Speziellen Paralleladdierer mit Übertragvorausberechnung schwer mit Computeralgebra zu verifizieren. Diese Art von Paralleladdierern beinhaltet Sequenzen von ODER-Gattern, welche zu einem exponentiellen Wachstum der Zwischenresultate, sowohl im Preprocessing, als auch im Reduktionsalgorithmus führen. Demgegenüber lassen sich diese Addierwerke aber einfach mit Hilfe eines SAT-Solvers verifizieren. Basierend auf dieser Beobachtung entwickeln wir eine *Verbindung von SAT-Solving und Computeralgebra* [KBK19]. Ist das abschließende Addierwerk ein Paralleladdierer, ersetzen wir dieses durch einen Ripple-Carry Addierer, der vollständig aus Volladdierern aufgebaut ist. Die Korrektheit dieser Substitution, im genaueren die Äquivalenz des ursprünglichen Addierwerkes und des Ripple-Carry Addierers wird unter Einsatz eines SAT-Solvers bewiesen. Der vereinfachte Multiplizierer wird mit unseren entwickelten Computeralgebra-Techniken verifiziert.

Als Teil dieser Dissertation haben wir das vollautomatisierte Tool AMULET implementiert, welches einen Multiplizierer im AIG-Format verifiziert. Im ersten Schritt wird in AMULET die Substitution des abschließenden Addierwerkes vorgenommen. Heuristiken erlauben es dieses Addierwerk zu identifizieren und durch einen Ripple-Carry Addierer zu ersetzen. AMULET erstellt eine KNF-Formel um die Korrektheit der Substitution zu beweisen. Der vereinfachte Multiplizierer kann mittels der implementierten Polynombibliothek in AMULET verifiziert werden. Die polynomialen Algorithmen in AMULET sind optimal an unseren Anwendungsfall angepasst und können so die syntaktischen Besonderheiten der Polynome gezielt nutzen, was ein wesentlicher Vorteil gegenüber Computeralgebrasystemen ist, die für allgemeine Anwendungsbereiche konzipiert sind. AMULET ist als Open-Source verfügbar und ermöglicht effiziente Verifikation nun auch von komplexen Multiplizierern. Evaluierungen dazu sind in Kapitel 4 zu finden.

An dieser Stelle müssen wir uns nun Fragen stellen, wie der Verifizierer verifiziert wird. Es kann sein, dass die Verifikationstools Fehler enthalten, welche zu einem falschen Ergebnis führen. Eine mögliche Methode um die Korrektheit der Tools zu garantieren, ist es, die Tools selbst zu verifizieren. Dies ist aber ein aufwändiger manueller Prozess. In der Praxis ist es daher gebräuchlicher, Beweiszertifikate während der Verifikation zu generieren. Diese Zertifikate können dann von einem eigenständigen Tool auf Richtigkeit überprüft werden. Bei der jährlichen SAT-Competition werden beispielsweise schon seit 2013 Beweiszertifikate gefordert. Für algebraisches Beweisen benötigen wir ein Beweissystem, welches über Polynomgleichungen urteilen kann. In dieser Dissertation haben wir das Beweiskalkül *Practical Algebraic Calculus* (PAC) [RBK18] entworfen, der ermöglicht systemnahe algebraische Beweise zu generieren. PAC basiert auf dem Polynomial Calculus, welcher vor allem in der Proof Complexity Community seine Anwendung findet, um über Komplexitätsergebnisse zu argumentieren. Praktische Anwendungen des Polynomial

Calculus kamen bisher nicht vor, da aufgrund seiner Formalisierung generierte Zertifikate nicht effizient überprüft werden konnten.

Wir haben daher PAC als eine Instantiierung des Polynomial Calculus entwickelt. Ähnlich zu einer Rechenprobe mit Rest für Division von Zahlen, modellieren wir in PAC ein Beweiszertifikat für die wiederholte multivariate Polynomdivision aus einer Reihe von Additions- und Multiplikationsschritten. Wir haben den Beweis-Checker PACHECK implementiert, welcher diese Beweiszertifikate auf ihre Richtigkeit überprüft. In der ursprünglichen Fassung von PAC haben wir explizit gefordert, dass alle Beweisschritte aufgelistet werden, was zu großen Beweisdateien führt. In einer erweiterten Version haben wir ein *kompakteres Beweisformat* [KFB20] definiert, welches ermöglicht Polynome mit Hilfe von Indizes anzusprechen. Des Weiteren nutzen wir aus, dass in unserem Anwendungsfall alle Variablen nur binäre Werte annehmen. PACHECK wurde dahingehend adaptiert. Wir generieren diese Beweiszertifikate als ein Nebenprodukt der Verifikation in AMULET und heben uns damit von verwandten Arbeiten auf diesem Gebiet ab, da diese keine Beweiszertifikate erzeugen.

Die Kombination von SAT-Solving und Computeralgebra [KBK19] hat zur Folge, dass Beweiszertifikate in unterschiedlichen Beweissystemen generiert werden. Für SAT wird vom SAT-Solver ein Beweis im sogenannten DRUP Beweissystem generiert. Für Computeralgebra generieren wir Zertifikate in PAC. Diese Beweise werden von unterschiedlichen Tools validiert, was eine Lücke im Beweiszertifikat hinterlässt. Kompositionelles Beweisen mittels eines Theorembeweisers könnte diese Lücke schließen, benötigt jedoch manuelle Unterstützung. Wir zeigen, dass sich *DRUP Beweise vollständig in PAC simulieren* lassen und wir entwickeln einen zugehörigen Übersetzungs-Ablauf [KBK20a]. Daher können wir ein vollständiges PAC Beweiszertifikat für die Verifikation von Multiplizierern generieren.

4 Resultate

Wir evaluieren die algebraische Verifikation von Multiplizierern und betrachten die Entwicklung dieser Technik innerhalb dieser Dissertation [RBK17, KBK19, KBK20b] und vergleichen die Ergebnisse mit aktuellen Tools von verwandten Arbeiten [Ci20, MGD19].

In ersten Experimenten evaluieren wir unsere Technik auf einem breiten Benchmarkset, welches aus 384 unterschiedlichen 64-Bit Multiplizierer-Architekturen besteht. Dieses Experiment ist in der linken Grafik von Abb. 2 zu sehen und zeigt die Anzahl der korrekt verifizierten Multiplizierern (sortiert nach Verifikationszeit) innerhalb eines Prozessorzeitlimits von 300 Sekunden. AMULET 1.0 ist in dieser Dissertation publiziert, AMULET 1.5 ist eine aktualisierte Version mit verbesserten Heuristiken zur Identifikation von Paralleladdierern mit Übertragvorausberechnung. Im zweiten Experiment, zu sehen auf der rechten Seite von Abb. 2, wählen wir einen simplen Multiplizierer, der sich vollständig in Voll- und Halbaddierer zerteilen lässt, mit unterschiedlichen Eingangsbitbreiten n. Das Zeitlimit für dieses Experiment wird auf 86 400 Sekunden (24 Stunden) gesetzt.

Es zeigt sich, dass AMULET und insbesondere dessen aktualisierte Version eine Größenordnung schneller ist als vergleichbare Tools [MGD19] und im Gegensatz zu [Ci20] auch komplexe Multiplizierer effizient verifizieren und zertifizieren kann.

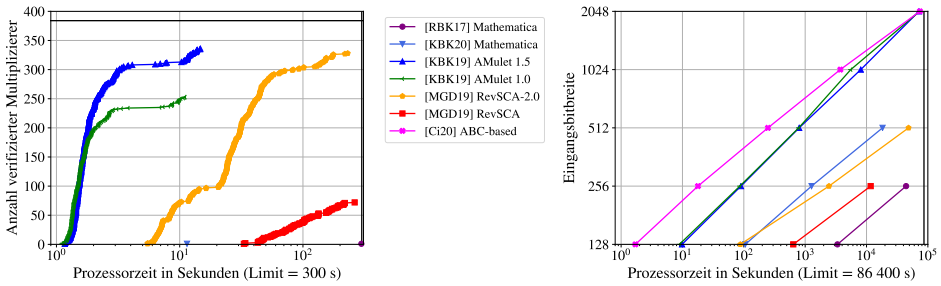


Abb. 2: Verifikation von 384 verschiedenen 64-Bit Multiplizierer-Architekturen (links) und von einfachen Multiplizierern mit großen Bitbreiten (rechts).

5 Schlussfolgerung

Automatisierte formale Verifikation von arithmetischen Schaltungen, insbesondere von Integer-Multiplizierern, wird in der Praxis immer noch als große Herausforderung angesehen. In dieser Dissertation betrachten und verbessern wir Verifikationsmethoden basierend auf Computeralgebra, um effiziente Verifikation von komplexen Multiplizierern zu ermöglichen. Wir zeigen eine umfassende mathematische Formalisierung, welche auf der Theorie der Gröbnerbasen basiert und daher auch die Verwendung von Mathematik in diesem Gebiet erweitert. Eine neue Kombination von logischem und algebraischem Rechnen führt zu innovativen und wirkungsvollen Methoden für die Verifikation von Multiplizierern. Wir präsentieren erstmalig ein algebraisches Beweissystem, welches erlaubt kompakte Zertifikate für die Verifikation zu generieren. Es zeigt sich, dass ein Zusammenspiel der von uns entwickelten Techniken ein effizientes Lösen einer Vielfalt von komplexen Multiplizierer-Architekturen ermöglicht.

Literaturverzeichnis

[Bi16] Biere, Armin: Collection of Combinational Arithmetic Miters Submitted to the SAT Competition 2016. In: SAT Competition 2016. Jgg. B-2016-1 in Dep. of Computer Science - Series of Publications B. University of Helsinki, S. 65–66, 2016.

[BL17] Beame, Paul; Liew, Vincent: Towards Verifying Nonlinear Integer Arithmetic. In: CAV 2017. Jgg. 10427 in LNCS. Springer, S. 238–258, 2017.

[CB95] Chen, Yirng-An; Bryant, Randal E.: Verification of Arithmetic Circuits with Binary Moment Diagrams. In: DAC 1995. ACM, S. 535–541, 1995.

[Ci20] Ciesielski, Maciej J.; Su, Tiankai; Yasin, Atif; Yu, Cunxi: Understanding Algebraic Rewriting for Arithmetic Circuit Verification: a Bit-Flow Model. IEEE TCAD, 39(6):1346–1357, 2020.

[Hu17] Hunt, Jr., Warren A.; Kaufmann, Matt; Strother Moore, J; Slobodova, Anna: Industrial Hardware and Software Verification with ACL2. Philos. Trans. Royal Soc. A, 375(2104):20150399, 2017.

[Ka20] Kaufmann, Daniela: Formal Verification of Multiplier Circuits using Computer Algebra. Dissertation, Informatik, Johannes Kepler University Linz, 2020.

[KBK19] Kaufmann, Daniela; Biere, Armin; Kauers, Manuel: Verifying Large Multipliers by Combining SAT and Computer Algebra. In: FMCAD 2019. IEEE, S. 28–36, 2019.

[KBK20a] Kaufmann, Daniela; Biere, Armin; Kauers, Manuel: From DRUP to PAC and Back. In: DATE 2020. IEEE, S. 654–657, 2020.

[KBK20b] Kaufmann, Daniela; Biere, Armin; Kauers, Manuel: Incremental Column-wise Verification of Arithmetic Circuits using Computer Algebra. FMSD, 56(1):22–54, 2020.

[KFB20] Kaufmann, Daniela; Fleury, Mathias; Biere, Armin: Pacheck and Pastèque, Checking Practical Algebraic Calculus Proofs. In: FMCAD 2020. TU Vienna Academic Press, S. 264–269, 2020.

[MGD19] Mahzoon, Alireza; Große, Daniel; Drechsler, Rolf: RevSCA: Using Reverse Engineering to Bring Light into Backward Rewriting for Big and Dirty Multipliers. In: DAC 2019. ACM, S. 185:1–185:6, 2019.

[RBK17] Ritirc, Daniela; Biere, Armin; Kauers, Manuel: Column-Wise Verification of Multipliers Using Computer Algebra. In: FMCAD 2017. IEEE, S. 23–30, 2017.

[RBK18] Ritirc, Daniela; Biere, Armin; Kauers, Manuel: A Practical Polynomial Calculus for Arithmetic Circuit Verification. In: SC2 Workshop 2018. CEUR-WS, S. 61–76, 2018.

[Sa16] Sayed-Ahmed, Amr; Große, Daniel; Kühne, Ulrich; Soeken, Mathias; Drechsler, Rolf: Formal Verification of Integer Multipliers byCombining Gröbner Basis with Logic Reduction. In: DATE 2016. IEEE, S. 1048–1053, 2016.

[SB94] Sharangpani, H.; Barton, M. L.: Statistical Analysis of Floating Point Flaw in the Pentium Processor. 1994.

[SK04] Stoffel, Dominik; Kunz, Wolfgang: Equivalence Checking of Arithmetic Circuits on the Arithmetic Bit Level. IEEE TCAD, 23(5):586–597, 2004.

[Va07] Vasudevan, Shobha; Viswanath, Vinod; Sumners, Robert W.; Abraham, Jacob A.: Automatic Verification of Arithmetic Circuits in RTL Using Stepwise Refinement of Term Rewriting Systems. IEEE Trans. Comput., 56(10):1401–1414, 2007.

[Yu16] Yu, Cunxi; Brown, Walter; Liu, Duo; Rossi, André; Ciesielski, Maciej J.: Formal Verification of Arithmetic Circuits by Function Extraction. IEEE TCAD, 35(12):2131–2142, 2016.

Daniela Kaufmann, geboren 1991 in Linz/Österreich, studierte von 2011 bis 2016 an der Johannes Kepler Universität Linz das Bachelorstudium technische Mathematik, sowie das Masterstudium Computermathematik. Daniela Kaufmann begann 2016 ihr Doktorratsstudium in Informatik unter der Betreuung von Prof. Armin Biere am Institut für Formale Modelle und Verifikation an der Johannes Kepler Universität Linz. Ihre erste Publikation über formale Verifikation von Multiplizierern wurde mit dem Best Paper Award der Intl. Conference for Formal Models in Computer Aided Design (FMCAD), einer der Top-Tier Konferenzen für Hardwareverifikation, ausgezeichnet. In 2020 wurde ihre Forschung mit dem JKU Young Researcher Award prämiert. Ihre Forschungsinteressen umfassen angewandte formale Methoden, Hardwareverifikation, Beweissysteme, sowie SAT Solving und Computeralgebra.

Steffen Hölldobler et al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 179

Scharfe Worst-Case-Garantien und Approximationsalgorithmen für verschiedene Klassen geometrischer Optimierungsprobleme[1]

Phillip Keldenich[2]

Abstract: In diesem Beitrag betrachten wir verschiedene NP-schwere geometrische Optimierungsprobleme aus den Bereichen der *konfliktfreien Färbung* von Graphen, der *kollisionsfreien Bewegungsplanung* und der geometrischen *Packung* und *Überdeckung*, und fassen die Ergebnisse unserer Arbeit zusammen, die in [Ke20] ausführlich beschrieben werden. Neben anderen Ergebnissen präsentieren wir zu verschiedenen Problemvarianten aus diesen Problemfeldern Garantien, die den Faktor zwischen einer offensichtlichen Schranke an die optimale Lösung einer Instanz und dem tatsächlichen Wert einer optimalen Lösung beschränken. In vielen Fällen sind diese Garantien bestmöglich, was bedeutet dass es Familien von Instanzen gibt, für die der garantierte Faktor angenommen wird. Die konstruktiven Beweise für diese Garantien basieren auf Algorithmen, die sich in jedem Fall auch als effiziente Approximationsalgorithmen mit konstantem Approximationsfaktor interpretieren lassen.

1 Einführung

Viele praxisrelevante geometrische Optimierungsprobleme sind NP-schwer. Das bedeutet unter anderem, dass es für sie vermutlich keinen effizienten Algorithmus gibt, der zu jeder gegebenen Instanz eine optimale Lösung berechnet. Um mit dieser Situation umzugehen, kann man von vornherein auf eine optimale Lösung verzichten und stattdessen versuchen, eine möglichst gute Lösung zu berechnen. Hier ist zwischen heuristischen Lösungen und Approximationsalgorithmen zu unterscheiden. Bei heuristischen Lösungen hat man in der Regel keine Garantien an den Faktor zwischen einer optimalen Lösung und einer heuristisch gefundenen Lösung. Approximationsalgorithmen hingegen besitzen einen sogenannten Approximationsfaktor c und man weiß, dass die gefundene Lösung höchstens um einen Faktor c schlechter sein kann als eine optimale Lösung.

Viele Approximationsalgorithmen aus der Literatur sind allerdings rein theoretische Ergebnisse. Einige Approximationsalgorithmen haben eine polynomielle Laufzeit, die aber zu hoch für einen praktischen Einsatz ist. Einige andere Approximationsalgorithmen haben Approximationsfaktoren, die für einen praktischen Einsatz inakzeptabel groß sind, und werden oft in der Praxis von heuristischen Ansätzen geschlagen. Bei vielen Approximationsalgorithmen gibt es aber uch eine Lücke zwischen dem Approximationsfaktor und der Abweichung zwischen optimaler Lösung und gefundener Lösung auf praxisrelevanten Instanzen. Dies liegt daran, dass der Approximationsfaktor auch im schlimmsten Fall noch gelten muss.

[1] Englischer Titel der Dissertation: „Tight Worst-Case Guarantees and Approximation Algorithms for Several Classes of Geometric Optimization Problems"

[2] TU Braunschweig, Abteilung Algorithmik, keldenich@ibr.cs.tu-bs.de

Außerdem gibt es bei manchen Approximationsalgorithmen auch eine Lücke zwischen dem bewiesenen Approximationsfaktor und dem tatsächlichen Approximationsfaktor. Dies liegt oft daran, dass der Wert einer optimalen Lösung in der mathematischen Analyse des Approximationsalgorithmus' schwer fassbar ist und der Algorithmus daher mit einer einfacheren Schranke an die optimale Lösung verglichen wird. Beispielsweise wird bei der Analyse von Überdeckungsalgorithmen statt der optimalen Lösung oft die Fläche bzw. das Volumen des zu überdeckenden Containers für den Vergleich herangezogen. Beim Vergleich mit dieser sogenannten *container volume bound* nimmt man implizit an, dass die optimale Lösung den Container immer überlappungsfrei ohne Verluste überdecken kann.

Beim Vergleich mit solchen Schranken nimmt man in Kauf, dass der bewiesene Approximationsfaktor den schlimmstmöglichen Faktor zwischen der optimalen Lösung und der Schranke enthalten muss. Neben der problemspezifischen Motivation hat diese Überlegung insbesondere unsere Arbeit zu Pack- und Überdeckungsproblemen motiviert, in der wir diesen schlimmstmöglichen Faktor für bestimmte Problemvarianten bestimmen.

Dieser Beitrag ist wie folgt strukturiert: In Abschnitt 2 fassen wir unsere Ergebnisse zur konfliktfreien Färbung von Graphen zusammen, in Abschnitt 3 die Ergebnisse zur kollisionsfreien Bewegungsplanung und in Abschnitt 4 die Ergebnisse zu geometrischen Pack- und Überdeckungsproblemen. Jeder der Abschnitte 2–4 beschreibt dabei auch die untersuchten Probleme und die Motivation unserer Arbeit. Abschließend ziehen wir in Abschnitt 5 ein Fazit.

2 Konfliktfreie Färbung

Die Klasse der Färbungsprobleme auf Graphen ist eine der ältesten Problemklassen, die man auf Graphen untersucht hat. Das klassische k-Färbungsproblem verlangt für einen gegebenen Graphen $G = (V,E)$ eine Färbung $c : V \to \{1,\dots,k\}$, die jedem Knoten eine von k Farben zuweist, sodass keine benachbarten Knoten dieselbe Farbe besitzen. Die kleinste Zahl an Farben, für die es eine solche Färbung für einen Graphen G gibt, nennt man die chromatische Zahl $\chi(G)$ von G. Das Färbungsproblem lässt sich auch als Überdeckungsproblem von V durch möglichst wenige unabhängige $S \subseteq V$ betrachten; viele Probleme, bei denen eine gewisse Menge von Objekten möglichst günstig durch Teilmengen überdeckt werden soll, deren Elemente in Konflikt miteinander stehen können, lassen sich als Färbungsproblem darstellen.

Als klassische Motivation für das Färbungsproblem wird oft die Vergabe von Frequenzen zur Datenübertragung angeführt; hier sollen durch verschiedene Frequenzen bei benachbarten Knoten Interferenzen verhindert werden. Allerdings ist die klassische Forderung, dass alle benachbarten Knoten unterschiedliche Farben haben, hier in vielen Fällen stärker als eigentlich notwendig. Soll beispielsweise ein Gebiet mit Funktürmen abgedeckt werden, so müssen nicht an jedem Punkt in diesem Gebiet alle dort erreichbaren Funktürme unterschiedliche Frequenzen haben, sondern es reicht für gewöhnlich aus, wenn wenigstens eine der Frequenzen nur einmal auftritt. Diese Überlegung ist eine Möglichkeit, das sogenannte *konfliktfreie Färbungsproblem* zu motivieren.

Das konfliktfreie Färbungsproblem in der von uns untersuchten Form verlangt für einen gegebenen Graphen $G = (V, E)$ eine Färbung einer Teilmenge $S \subseteq V$ der Knoten, sodass jeder Knoten v einen gefärbten Nachbarn hat, dessen Farbe in der Nachbarschaft von v nur einmal auftritt. Dieses Problem wurde vor unserer Arbeit beispielsweise von Pach und Tardos [PT09] studiert, die unter anderem gezeigt haben, dass ein Graph mit n Knoten höchstens $O(\log^2 n)$ Farben für eine konfliktfreie Färbung benötigt. Für einen vollständigeren Überblick über die Literatur verweisen wir aus Platzgründen auf [Ke20]. Analog zur chromatischen Zahl $\chi(G)$ eines Graphen definieren wir die konfliktfreie chromatische Zahl $\chi_{CF}(G)$ als die kleinste Zahl k von Farben, für die G eine konfliktfreie Färbung mit k Farben zulässt. Abbildung 1 zeigt ein Beispiel für eine konfliktfreie 2-Färbung mit abgeschlossenen Nachbarschaften.

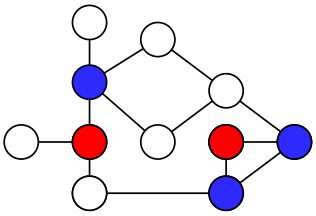


Abb. 1: Eine konfliktfreie 2-Färbung eines Graphen; weiße Knoten sind ungefärbt.

Neben der oben definierten Variante untersuchen wir verschiedene andere Varianten des Problems. Einerseits betrachten wir auch den Fall, dass alle Knoten des Graphen gefärbt werden müssen; andererseits betrachten wir das Problem sowohl für abgeschlossene Nachbarschaften, in denen ein Knoten zu seiner eigenen Nachbarschaft gehört, als auch für offene Nachbarschaften, wo dies nicht der Fall ist.

Außerdem lässt sich das Problem als Erweiterung des Problems DOMINATING SET auffassen, bei dem für jeden Knoten ein dominierender Nachbar mit eindeutiger Farbe existieren muss; in diesem Fall kann es erstrebenswert sein, bei einer gegebenen Anzahl von Farben die Anzahl gefärbter Knoten zu minimieren, was wir ebenfalls untersuchen.

Eines der bekanntesten Ergebnisse zur Färbung von Graphen ist der 4-Farben-Satz, der lange vor seinem Beweis durch Appel und Haken [AH77] als 4-Farben-Vermutung existierte und besagt, dass jeder planare Graph mit vier Farben gefärbt werden kann. Eine Verallgemeinerung des 4-Farben-Satzes stellt Hadwigers Vermutung [Ha43] dar. Sie besagt, dass jeder Graph mit k Farben gefärbt werden kann, der keinen vollständigen Graphen mit $k+1$ Knoten als Minor besitzt. Für das oben definierte konfliktfreie Färbungsproblem mit abgeschlossenen Nachbarschaften beweisen wir in unserer Arbeit [Ab18] die folgenden Sätze.

Satz 1. *Jeder Graph, der weder einen K_{k+2} noch einen K_{k+3}^{-3} als Minor hat, lässt sich mit k Farben konfliktfrei färben.*

Dabei steht K_{k+2} für einen vollständigen Graphen auf $k+2$ Knoten und K_{k+3}^{-3} für einen vollständigen Graphen auf $k+3$ Knoten, aus dem die drei Kanten eines Dreiecks entfernt wurden. Die Färbung kann mit einem einfachen Algorithmus effizient bestimmt werden.

Da der K_{k+1} in beiden verbotenen Minoren enthalten ist, folgt daraus insbesondere, dass Hadwigers Vermutung für konfliktfreie Färbungen für beliebiges k gilt.

Korollar 2. *Jeder Graph, der keinen K_{k+1} als Minor hat, ist konfliktfrei k-färbbar.*

Des Weiteren erhalten wir folgendes Korollar, da der vollständige bipartite Graph $K_{3,3}$ in K_6^{-3} enthalten ist.

Korollar 3. *Planare Graphen sind konfliktfrei 3-färbbar. Außenplanare Graphen sind konfliktfrei 2-färbbar.*

Diese Schranken sind scharf; insbesondere ist es NP-schwer zu entscheiden, ob ein planarer Graph konfliktfrei 1- oder 2-färbbar ist.

Satz 4. *Konfliktfreie k-Färbbarkeit ist auf allgemeinen Graphen für jedes $k \geq 1$ NP-schwer zu entscheiden. Für $k \in \{1,2\}$ gilt dies auch für planare Graphen.*

Falls alle Knoten gefärbt werden müssen, so lässt sich dies unter anderem erreichen, indem man allen ungefärbten Knoten eine neue Farbe zuweist. In diesem Fall benötigen planare Graphen vier Farben und außenplanare Graphen drei; auch diese Schranken sind scharf. Auch das Minimieren der Anzahl gefärbter Knoten ist ein NP-schweres Problem.

Satz 5. *Falls $\mathsf{P} \neq \mathsf{NP}$, so gibt es für kein $k \geq 3$ einen Approximationsalgorithmus für die Minimierung der Anzahl gefärbter Knoten in einer konfliktfreien k-Färbung, selbst wenn der gegebene Graph garantiert konfliktfrei k-färbbar ist. Für $k = 3$ Farben gilt dies auch auf planaren Graphen.*

Auf der positiven Seite gilt für $k \geq 4$ Farben, dass sich aus jedem DOMINATING SET D eines planaren Graphen in polynomieller Zeit eine konfliktfreie 4-Färbung mit $|D|$ gefärbten Knoten berechnen lässt. Dies erlaubt unter anderem, FPT-Algorithmen und polynomielle Approximationsschemata für planares DOMINATING SET zu benutzen, um konfliktfreie 4-Färbungen planarer Graphen mit möglichst geringer gefärbter Knotenzahl zu erhalten.

Für offene Nachbarschaften beweisen wir die folgenden Aussagen.

Satz 6. *Alle bipartiten planaren Graphen können mit vier Farben konfliktfrei gefärbt werden. Es gibt bipartite planare Graphen, die vier Farben benötigen. Für $k < 4$ Farben ist es NP-schwer zu entscheiden, ob k Farben für einen bipartiten planaren Graphen ausreichen. Jeder planare Graph kann mit acht Farben konfliktfrei gefärbt werden.*

Neben planaren Graphen untersuchen wir das konfliktfreie Färbungsproblem mit abgeschlossenen Nachbarschaften auch auf Schnittgraphen geometrischer Objekte [FK18]. Für frei skalierbare Objekte, die beliebig dünn werden können, wie beispielsweise Ellipsen und Rechtecke, aber auch allgemeinere Klassen wie (konvexe) Polygone, kann der Schnittgraph von n Objekten $\Omega(\log n / \log\log n)$ Farben für eine konfliktfreie Färbung benötigen. Für Objekte, die zwar beliebig skaliert werden können aber nicht beliebig dünn werden können,

kann der Schnittgraph von n Objekten $\Omega(\sqrt{\log n})$ Farben benötigen. Außerdem ist das Problem bereits auf Schnittgraphen von Kreisen oder Quadraten NP-schwer. Weiterhin zeigen wir, dass jeder Schnittgraph von Einheitskreisen mit sechs und jeder Schnittgraph von Einheitsquadraten mit vier Farben konfliktfrei gefärbt werden kann; wenigstens zwei Farben sind für diese Graphen manchmal notwendig.

3 Kollisionsfreie Bewegungsplanung

Die gleichzeitige Bewegung von Robotern in einem Schwarm (oder anderen mobilen Objekten, wie beispielsweise Flugzeugen, Fahrzeugen oder Drohnen) führt zu Herausforderungen, die bei der Betrachtung eines einzelnen beweglichen Objekts keine Rolle spielen. Insbesondere muss bei der Planung der Bewegungen dafür gesorgt werden, dass die Objekte nicht miteinander kollidieren. Gleichzeitig haben die Objekte individuelle Start- und Zielpositionen, die möglicherweise dazu führen, dass ein Objekt einen Umweg in Kauf nehmen oder warten muss, um Kollisionen mit anderen Objekten zu vermeiden. Solange die Objekte von einer einzelnen Kontrollinstanz gesteuert werden — wie beispielsweise Roboter in einem Lagerhaus — ist es möglich und oft sinnvoll zu versuchen, die Zeit zu minimieren, die vergeht, bis das letzte Objekt an seiner Zielposition angekommen ist.

Aufgrund der hohen praktischen Relevanz der Probleme, die bei der Koordination gleichzeitiger Bewegungen auftreten, ist es nicht verwunderlich, dass es sowohl von praktischer als auch theoretischer Seite bereits eine Menge wissenschaftlicher Arbeiten zu diesen Themen gibt. Wir verweisen aus Platzgründen für einen Überblick auf [Ke20].

Eine *Konfiguration* weist jedem Roboter eine Position zu. Unser Ziel ist es, die Roboter von einer gegebenen Startkonfiguration in eine Zielkonfiguration zu überführen, dabei Kollisionen zu vermeiden und die insgesamt benötigte Zeit (den sogenannten *Makespan*) zu minimieren. Dies setzt in der Regel voraus, dass sich viele der Objekte gleichzeitig bewegen.

Wir untersuchen in unserer Arbeit [De19] zwei Varianten dieses Problems. Einerseits untersuchen wir das Problem auf zweidimensionalen Gittern, die aus Pixeln bestehen, die zu jeder Zeit nur von höchstens einem Roboter besetzt sein dürfen. Andererseits untersuchen wir das Problem in der euklidischen Ebene, wobei Roboter als zusammenhängende zweidimensionale Objekte betrachtet werden, die sich zu keiner Zeit überlappen dürfen.

In der von uns untersuchten Problemvariante auf dem Gitter bewegen sich die Roboter synchron in parallelen Bewegungsschritten, wobei jeder Roboter entweder stehenbleibt oder sich in eines der horizontal oder vertikal benachbarten Pixel bewegt. Eine Bewegung von einem Roboter auf Pixel p auf ein benachbartes Pixel q ist dabei in einem Schritt S zulässig, wenn q entweder leer ist, oder der Roboter auf q in Schritt S von q auf ein anderes Pixel $r \neq p$ bewegt wird. Außerdem darf kein Pixel in S Ziel von mehr als einem Roboter sein. Dieses Modell erlaubt es, Roboter entlang eines vollständig besetzten Kreises von Pixeln zu bewegen, aber es erlaubt keine Vertauschung benachbarter Roboter innerhalb eines Schritts. Abbildung 2 zeigt Beispiele erlaubter und verbotener Schritte.

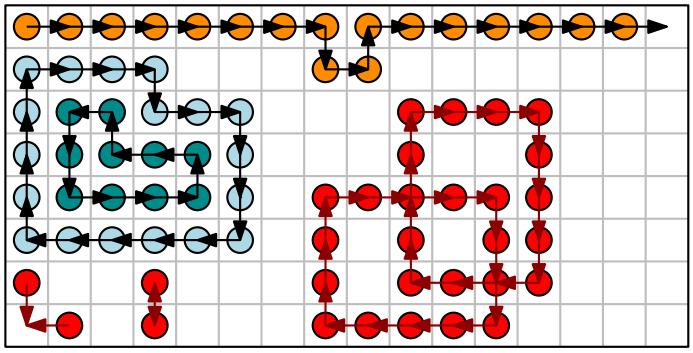


Abb. 2: Beispiele für in einem parallelen Bewegungsschritt verbotene Bewegungen (rot) und erlaubte Bewegungen (andere Farben).

Das Minimieren des Makespans ist bereits auf einem rechteckigen Gitter ohne Hindernisse NP-schwer.

Satz 7 ([De19])**.** *Es ist* NP*-schwer zu entscheiden, ob zu gegebenen Start- und Zielkonfigurationen auf dem Gitter ein gegebener Makespan M ohne Kollisionen realisiert werden kann.*

Jeder Roboter kann sich in einem Schritt höchstens einen Pixel weit bewegen. Eine einfache untere Schranke für den Makespan einer optimalen Lösung erhält man daher, indem man die maximale Distanz d zwischen der Start- und Zielposition eines Roboters berechnet. Das Hauptergebnis unserer Arbeit [De19] ist der folgende Satz.

Satz 8. *Es gibt eine Konstante c, sodass für alle Start- und Zielkonfigurationen mit maximaler Distanz d auf rechteckigen Gittern mit Seitenlängen mindestens 2×3 ein Bewegungsplan aus höchstens $c \cdot d$ Schritten existiert. Es gibt einen Algorithmus, der einen solchen Bewegungsplan in polynomieller Zeit bestimmt.*

Der Algorithmus aus Satz 8 ist insbesondere ein c-Approximationsalgorithmus für die Minimierung des Makespans. Wir können diesen Satz auch auf zweidimensionale Roboter in der euklidischen Ebene anwenden, um sie mit konstantem Overhead aus einer Start- in eine Zielkonfiguration zu überführen. Das funktioniert, solange alle Start- und Zielpositionen weit genug voneinander entfernt sind. Dazu skalieren wir die Konfigurationen und die Roboter, sodass jeder Roboter in einer Kreisscheibe mit Radius 1 enthalten ist. Außerdem nehmen wir an, dass sich alle Roboter höchstens mit Geschwindigkeit 1 bewegen können. Wir beweisen folgenden Satz, indem wir ein geeignetes Gitter über die Roboterkonfigurationen legen und den Algorithmus aus Satz 8 darauf anwenden.

Satz 9. *Für jedes Paar von Start- und Zielkonfiguration, bei dem die paarweise Distanz zwischen allen Start- und Zielpositionen mindestens 4 ist, lässt sich mit einem Polynomialzeitalgorithmus ein Makespan von $O(d)$ realisieren, wobei d die maximale Distanz zwischen der Start- und Zielposition eines Roboters ist.*

Falls die Distanz zwischen Start- und Zielpositionen zu klein ist, gibt es Instanzen in denen kein konstanter Faktor zwischen Makespan und der maximalen Distanz d möglich ist; siehe Abbildung 3 für eine Skizze einer solchen Instanz.

Satz 10. *Es gibt eine Familie $\mathscr{F}$ von Instanzen mit N Robotern mit maximaler Distanz $d \in \Theta(1)$, für die der minimale kollisionsfrei erreichbare Makespan $M \in \Omega(N^{1/4})$ ist.*

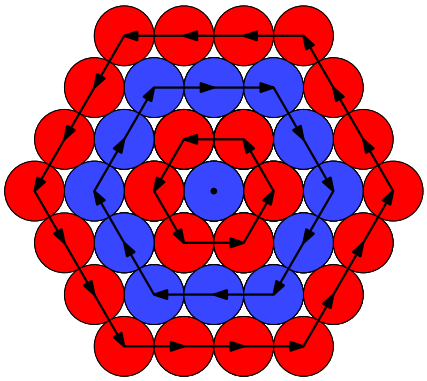


Abb. 3: Eine Instanz aus der Instanzfamilie $\mathscr{F}$ aus Satz 10, bestehend aus Schichten von roten und blauen Robotern. Es können beliebig viele weitere Schichten von Robotern hinzugefügt werden. Um die miteinander verschränkten Schichten rotieren zu können, muss man die Roboter weit genug auseinander bewegen; wir zeigen, dass dies dazu führt, dass ein konstanter Faktor zwischen Makespan und Distanz d unmöglich wird.

Auf der positiven Seite präsentieren wir einen Algorithmus, der auf beliebigen Instanzen in der euklidischen Ebene funktioniert.

Satz 11. *Es gibt einen Polynomialzeitalgorithmus, der zu einem gegebenen Paar von Start- und Zielkonfigurationen einen kollisionsfreien Bewegungsplan mit Makespan $O(d+\sqrt{N})$ berechnet.*

4 Geometrische Pack- und Überdeckungsprobleme

Bei geometrischen Packproblemen geht es in der Regel darum, eine Menge von Objekten in einen Container C zu packen, sodass sich die Objekte nicht überlappen und vollständig im Container enthalten sind; analog geht es bei Überdeckungsproblemen darum, einen gegebenen Container vollständig mit einer gegebenen Menge von Objekten zu überdecken; hierbei dürfen die Objekte sich und den Rand des Containers überlappen. Geometrische Pack- und Überdeckungsprobleme treten in unterschiedlicher Form in vielen praktischen Anwendungsbereichen auf, unter anderem beim Packen von Paketen, Zuschneiden von Stoff aus Bahnen oder Baumaterial aus vorgefertigten Platten oder bei der Abdeckung eines Gebiets mit Funktürmen oder Sicherheitskameras. Viele dieser Probleme sind NP-schwer; oft ist nicht einmal klar, ob sie in NP enthalten sind. Neben den praktischen Anwendungen sind viele grundlegende Pack- und Überdeckungsprobleme auch wegen ihrer theoretischen Bedeutsamkeit studiert worden; für einen Überblick über die Literatur verweisen wir aus Platzgründen auf [Ke20].

Wir untersuchen in unseren Arbeiten das Packen von Quadraten [Fe21] und Kreisen [FKS19] in Kreise sowie das Überdecken von Rechtecken mit Kreisen [Fe20]. Zu jedem dieser Probleme präsentieren wir einen *worst-case-optimalen* Algorithmus. Dies ist für Packungsprobleme ein Algorithmus, der garantiert, dass jede beliebige Instanz mit einer zu packenden Gesamtfläche von höchstens $w(C)$ erfolgreich in einen Container C gepackt wird, wobei $w(C)$ nicht vergrößert werden kann, weil es für jedes $\varepsilon > 0$ eine Instanz mit zu packender Gesamtfläche $w(C) + \varepsilon$ gibt, die sich nicht in C passt. Analog gilt ein Algorithmus für Überdeckungsprobleme als worst-case-optimal, wenn er für jede Instanz mit einer Gesamtfläche von mindestens $w(C)$ erfolgreich den Container C überdeckt, wobei $w(C)$ die Fläche der überdeckenden Objekte ist, die für eine erfolgreiche Überdeckung immer ausreichend und manchmal erforderlich ist.

Bei der Überdeckung von Rechtecken durch Kreise bekommen wir als Eingabe ein Rechteck als Container und wollen dieses mit einer Menge von Kreisen überdecken, die als endliche Folge $r_1, \ldots, r_n$ von Radien gegeben ist. Wir nehmen ohne Einschränkung an, dass das Rechteck die Kantenlängen $\lambda \times 1$ mit $\lambda \geq 1$ hat; alle anderen Fälle gehen durch Skalierung und Rotation aus diesem Fall hervor. Ferner nehmen wir an, dass $r_1 \geq \cdots \geq r_n$, das heißt, dass die Kreise nach Größe absteigend geordnet sind. Wir beweisen dazu folgenden Satz.

Satz 12 ([Fe20]). *Sei $\lambda \geq 1$ und sei $\mathscr{R}$ ein Rechteck der Größe $\lambda \times 1$. Seien ferner*

$$\lambda_2 := \sqrt{\frac{\sqrt{7}}{2} - \frac{1}{4}} \text{ und } w(\lambda) = \begin{cases} 3\pi\left(\frac{\lambda^2}{16} + \frac{5}{32} + \frac{9}{256\lambda^2}\right), & \text{falls } \lambda < \lambda_2, \\ \pi\frac{\lambda^2+2}{4}, & \text{sonst.} \end{cases}$$

(1) Für jedes $a < w(\lambda)$ gibt es eine Menge D^- von Kreisen mit Gesamtfläche a, die nicht ausreicht, um $\mathscr{R}$ abzudecken. (2) Sei $r_1 \geq \cdots \geq r_n > 0$ eine Folge von Radien mit einer Kreisfläche von insgesamt $\pi\sum_{i=1}^{n} r_i^2 \geq w(\lambda)$. Dann kann man in polynomieller Zeit eine Platzierung von n Kreisen mit Radien $r_1, \ldots, r_n$ finden, die $\mathscr{R}$ vollständig abdeckt.

Teil (1) von Satz 12 basiert auf zwei Instanzklassen, die in Abbildung 4(a) dargestellt sind. Teil (2) basiert auf einem rekursiven Überdeckungsalgorithmus, der verschiedene relativ einfache Überdeckungsroutinen verfolgt, die jeweils eine konstante Zahl von Kreisen explizit platzieren, den ggf. unabgedeckten Teil von $\mathscr{R}$ in Teilrechtecke aufteilen, die verbleibenden Kreise entsprechend aufteilen und dann den Algorithmus rekursiv auf jedes der Teilrechtecke anwenden. Jede Routine hat ein hinreichendes Erfolgskriterium, das lediglich von λ und r_1 bis r_7 abhängt und die Anwendbarkeit der Routine und den Erfolg der Rekursion garantiert.

Die Garantie, dass der Algorithmus mit mindestens $w(\lambda)$ Kreisfläche immer erfolgreich ist, wird durch Induktion über die Anzahl der Kreise bewiesen. Dabei wird in Induktionsanfang und Induktionsschritt gezeigt, dass immer wenigstens eines der Erfolgskriterien der Überdeckungsroutinen erfüllt ist. Aufgrund der großen Anzahl von Routinen wird dieser Teil des Beweises mit Computerunterstützung bewiesen. Der computerunterstützte Teil des Beweises basiert auf der Unterteilung des Raums, der von λ und $r_1, \ldots, r_7$ aufgespannt wird, in eine große endliche Anzahl von Hyperrechtecken. Für jedes einzelne Hyperrechteck kann dann mithilfe von Intervallarithmetik automatisiert bewiesen werden, dass wenigstens eines der Erfolgskriterien für alle Punkte innerhalb des Hyperrechtecks erfüllt ist.

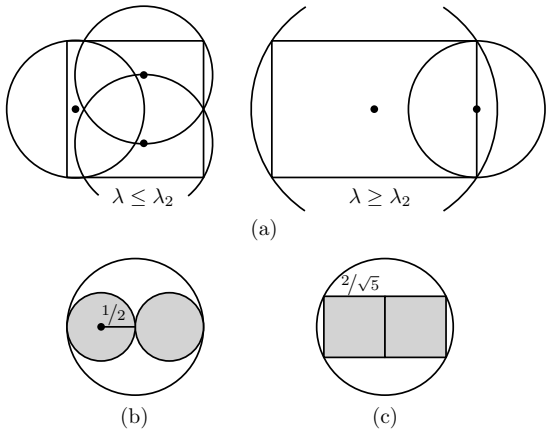


Abb. 4: (a) Die zwei schlimmstmöglichen Fälle für das Überdecken von Rechtecken mit Kreisen. Verringert man den Radius des größten Kreises um ein $\varepsilon > 0$, kann das Rechteck nicht mehr überdeckt werden. Unten: Der schlimmstmögliche Fall für die Packung von Kreisen (b) oder Quadraten (c) in Kreise.

In zwei anderen Arbeiten beweisen wir, ebenfalls mit Computerunterstützung, ähnliche Sätze für die Packung von Kreisen [FKS19] bzw. Quadraten [Fe21] in Kreisscheiben. Die Garantien, die diese Sätze geben, sind ebenfalls bestmöglich; Abbildung 4 zeigt Instanzen, für die die Objekte nicht vergrößert werden können, ohne Überlappungen zu erzwingen.

Satz 13 ([FKS19])**.** *Für jede Folge $r_1 \geq \cdots \geq r_n > 0$ mit einer Kreisfläche von insgesamt $\pi \sum_{i=1}^{n} r_i^2 \leq \pi/2$ existiert eine Packung von n Kreisen mit Radien $r_1, \ldots, r_n$ in die Einheitskreisscheibe. Eine solche Packung kann in polynomieller Zeit gefunden werden.*

Satz 14 ([Fe21])**.** *Für jede Folge $s_1 \geq \cdots \geq s_n > 0$ von Seitenlängen von Quadraten mit Gesamtfläche $\sum_{i=1}^{n} s_i^2 \leq 8/5$ existiert eine Packung von n Quadraten mit Seitenlängen $s_1, \ldots, s_n$ in die Einheitskreisscheibe. Eine solche Packung kann n polynomieller Zeit gefunden werden.*

5 Fazit

Neben anderen Resultaten haben wir in unserer Arbeit für verschiedene Problemklassen Worst-Case-Garantien und worst-case-optimale Algorithmen vorgestellt. Es ist wenig überraschend, dass die verschiedenen Problemfelder grundsätzlich verschiedene Ansätze zum Beweis solcher Resultate brauchen. Insbesondere im Fall der Resultate für Pack- und Überdeckungsprobleme sind wir aber dennoch optimistisch, dass sich die von uns eingesetzten Techniken und Werkzeuge auch für andere Problemvarianten und möglicherweise auch für andere Problemfelder als nützlich erweisen könnten.

Literaturverzeichnis

[Ab18] Abel, Zachary; Alvarez, Victor; Demaine, Erik D.; Fekete, Sándor P.; Gour, Aman; Hesterberg, Adam; Keldenich, Phillip; Scheffer, Christian: Conflict-Free Coloring of Graphs. SIAM Journal on Discrete Mathematics, 32:2675–2702, 2018.

[AH77] Appel, K.; Haken, W.: Every planar map is four colorable. Part I. Discharging. Illinois Journal of Mathematics, 21:429–490, 1977.

[De19] Demaine, Erik D.; Fekete, Sándor P.; Keldenich, Phillip; Scheffer, Christian; Meijer, Henk: Coordinated Motion Planning: Reconfiguring a Swarm of Labeled Robots with Bounded Stretch. SIAM Journal on Computing, 48:1727–1762, 2019.

[Fe20] Fekete, Sándor P.; Gupta, Utkarsh; Keldenich, Phillip; Scheffer, Christian; Shah, Sahil: Worst-Case Optimal Covering of Rectangles by Disks. In: Proceedings of the 36th International Symposium on Computational Geometry (SoCG 2020). S. 42:1–42:23, 2020.

[Fe21] Fekete, Sándor P.; Gurunathan, Vijaykrishna; Juneja, Kushagra; Keldenich, Phillip; Kleist, Linda; Scheffer, Christian: Packing Squares into a Disk with Optimal Worst-Case Density. In: Proceedings of the 37th International Symposium on Computational Geometry (SoCG 2021). 2021. Zur Veröffentlichung angenommen.

[FK18] Fekete, Sándor P.; Keldenich, Phillip: Conflict-Free Coloring of Intersection Graphs. International Journal of Computational Geometry & Applications, 28:289–307, 2018.

[FKS19] Fekete, Sándor P; Keldenich, Phillip; Scheffer, Christian: Packing Disks into Disks with Optimal Worst-Case Density. In: Proceedings of the 35th International Symposium on Computational Geometry (SoCG 2019). S. 35:1–35:19, 2019.

[Ha43] Hadwiger, Hugo: Über eine Klassifikation der Streckenkomplexe. Vierteljahresschrift der Naturforschenden Gesellschaft in Zürich, 88:133–143, 1943.

[Ke20] Keldenich, Phillip: Tight Worst-Case Guarantees and Approximation Algorithms for Several Classes of Geometric Optimization Problems. Dissertation, TU Braunschweig, 2020.

[PT09] Pach, J.; Tardos, G.: Conflict-Free Colourings of Graphs and Hypergraphs. Combinatorics, Probability and Computing, 18(05):819–834, 2009.

Phillip Keldenich wurde am 8. November 1989 in Langenfeld (Rheinland) geboren. Von 1996 bis 2000 besuchte er die Gemeinschaftsgrundschule Birkenhöhe in Wuppertal. Danach besuchte er ebenda von 2000 bis 2009 das Gymnasium Bayreuther Straße, das er 2009 nach dem Abitur verließ. Von 2009 bis 2010 war er Zivildienstleistender in der Bergischen Diakonie in Aprath. Ab dem Wintersemester 2010/11 folgte ein Bachelorstudium der Informatik an der RWTH Aachen. Dieses beendete er im Sommersemester 2013 mit Auszeichnung mit seiner Bachelorarbeit zu Lernalgorithmen für reguläre Automaten mit schwacher Akzeptanzbedingung über unendlichen Wörtern. Für seinen Abschluss wurde er mit dem Schöneborn-Preis geehrt. Auf das Bachelorstudium folgte ab dem Wintersemester 2013/14 ein Masterstudium der Informatik, ebenfalls an der RWTH Aachen, das er im Oktober 2015 ebenfalls mit Auszeichnung abschloss. Im November 2015 trat er eine Stelle als wissenschaftlicher Mitarbeiter in der Abteilung Algorithmik an der TU Braunschweig an und begann sein Promotionsstudium unter Prof. Dr. Sándor Fekete, das er im November 2020 mit der Note summa cum laude abschloss.

Verifikation Nebenläufiger Programme: Verfeinerung, Synchronisation, Sequenziellisierung[1]

Bernhard Kragl[2]

Abstract: Unsere Gesellschaft vertraut maßgeblich und stetig zunehmend auf verteilte IT Systeme. Allerdings sind Entwurf und Verifikation nebenläufiger Programme berüchtigt komplizierte, zeitintensive und fehleranfällige Aufgaben, selbst für Fachexperten. Der Grund dafür ist die enorme (unendliche) Menge an verzahnten Ausführungen nebenläufiger Programme. All diese Ausführungen müssen in einem formalen Korrektheitsbeweis erfasst und berücksichtigt werden. Solche Beweise basieren bekanntermaßen auf induktiven Invarianten über den globalen Programmzustand. Die Erarbeitung induktiver Invarianten für konkrete Implementierungen ist zwar theoretisch möglich, aber praktisch undenkbar. In dieser Dissertation präsentieren wir eine Verifikationsmethodik basierend auf dem Konzept der schrittweisen Verfeinerung, welche die Konstruktion formaler Korrektheitsbeweise grundlegend vereinfacht. Wir präsentieren eine Programmiersprache zur kompakten Beschreibung von Beweisen über mehrere Abstraktionsebenen. Eine mächtige Beweisregel unterteilt das Verifikationsproblem in zahlreiche automatisierbare Unterprobleme. Neue Beweistaktiken für asynchrone Programme ermöglichen die Reduktion von komplexen nebenläufigen Ausführungen zu simplen sequenziellen Ausführungen. Unsere Methodik ist in unserem statischen Analyseprogramm CIVL implementiert und dessen Effektivität wird in zahlreichen Fallstudien demonstriert. CIVL wurde bereits in mehreren Publikationen anderer Forscher verwendet.

1 Einführung

Nebenläufigkeit ist in heutigen Computersystemen allgegenwärtig. Geografisch verteilte Datenzentren benötigen fehlertolerante verteilte Algorithmen zur Sicherstellung eines konsistenten Systemzustandes; massiv-parallele Supercomputer ermöglichen gigantische wissenschaftliche Berechnungen; Mobil- und Webanwendungen verwenden ereignisgetriebene asynchrone Programmierung um ein schnelles und flüssiges Benutzererlebnis zu bieten; die Kontrollsoftware von eingebetteten Systemen muss auf asynchrone Ereignisse der realen Welt reagieren; nur um einige Beispiele zu nennen. In nebenläufigen Programm erfolgen mehrere Berechnungsschritte gleichzeitig. Zum Beispiel können auf einem Mehrkernprozessor mehrere Threads (je einer pro Prozessorkern) gleichzeitig laufen, welche Daten über ein geteiltes Speichersystem austauschen. In einem verteilten System kommunizieren Prozesse auf unabhängigen Computern durch den austausch von Nachrichten welche über ein Netzwerk geschickt werden. Formell verstehen wir "gleichzeitig" als eine Verzahnung der atomaren (d.h. nicht weiter in Teilschritte unterteilbare) Aktionen der nebenläufigen Berechnungen. Sind zum Beispiel $A_1;A_2;A_3$ die atomaren Aktionen von Prozess A und $B_1;B_2;B_3$ die atomaren Aktionen von Prozess B, so ist $A_1;B_1;B_2;A_2;B_3;A_3$ eine der 20 möglichen verzahnten Ausführungen der beiden Prozesse.

[1] Englischer Titel der Dissertation: "Verifying Concurrent Programs: Refinement, Synchronization, Sequentialization

[2] IST Austria, bernhard.kragl@gmail.com

Die Anzahl an verzahnten Ausführungen steigt exponentiell mit der Anzahl nebenläufiger Prozesse und der Anzahl atomarer Aktionen pro Prozess. Diese explodierende Anzahl an Verzahnungen macht das Nachvollziehen aller möglichen Ausführungen nebenläufiger Programme eine beinahe unmögliche mentale Aufgabe, insbesondere wenn Systeme adaptiert und weiterentwickelt werden. Unvorhergesehene Ausnahmesituationen definitiv auszuschließen ist extrem schwierig. Das Paxos Protokoll [La98], welches im Kern beinahe jedes replizierten Systems verwendet wird, ist berüchtigt für seine Komplexität; schon auf Papier [La02a, OO14, vRA15] und erst recht in konkreten Implementierungen [CGR07]. Im Chord Algorithmus [St01] für verteilte Hashtabellen fanden sich über 10 Jahre in allen publizierten Varianten Korrektheitsfehler [Za12].

Bevor wir sagen können *ob* ein System korrekt ist, müssen wir definieren *was es bedeutet* für das System korrekt zu sein; wir brauchen eine *Spezifikation*. Spezifikationen variieren in Form und Gestalt. Zum Beispiel können wir uns interessieren für flache generische Eigenschaften (wie Speichersicherheit oder Deadlockfreiheit), tiefe funktionale Eigenschaften, temporale Eigenschaften, Hypereigenschaften (wie Datenschutz), etc. In dieser Dissertation zielen wir auf tiefe funktionale Safety-Eigenschaften [MP90] ab, welche in mathematischer Logik ausgedrückt werden. Beispielsweise drückt in einem System mit einer endlichen Menge an Prozessen P die Formel

$$\forall\, p_1, p_2 \in P.\ p_1.\text{hasDecided} \wedge p_2.\text{hasDecided} \implies p_1.\text{decidedValue} = p_2.\text{decidedValue}$$

aus, dass sich die Prozesse auf einen konsistenten Wert einigen müssen (ein Teil des berühmten Konsensusproblems). Wenn sich je zwei beliebige Prozesse p_1 und p_2 für einen Wert entschieden haben, so muss dieser Wert der gleiche sein.

Es gibt viele Ansätze um Vertrauen in die Korrektheit eines Computerprogramms zu bekommen. Etwa durch Testen, statische Analyse, Modellprüfung, oder deduktive Verifikation. Diese Ansätze spannen ein breites Spektrum zwischen Kosten (Zeit, Ressourcen, Expertise, etc.) und Nutzen (stärke der resultierenden Garantie). Der Fokus dieser Dissertation ist das Beweisen starker benutzerdefinierter Spezifikationen *aller* (unendlich vieler) nebenläufiger Ausführungen durch deduktive Verifikation. Dabei ist unser Ziel, einerseits den nötigen manuellen (kreativen) Aufwand so einfach und angenehm als möglich zu gestalten, und andererseits alle automatisierbaren Schritte an einen Computer zu übergeben.

2 Deduktive Verifikation nebenläufiger Programme

Wir wiederholen die fundamentalen Konzepte der *induktiven Invarianten*, der *Reduktion*, und der *Verfeinerung*, und beschreiben unsere Innovationen auf Basis dieser Konzepte.

Induktive Invarianten. Angenommen ein Programm ist modelliert als Transitionssystem $(\mathit{Var}, \mathit{Init}, \mathit{Next}, \mathit{Safe})$, wobei Var die Menge an Programmvariablen, Init das Initialzustandsprädikat über Var, Next das Transitionsprädikat über $\mathit{Var} \cup \mathit{Var}'$, und Safe ein Sicherheitsprädikat über Var ist. Wir wollen sicherstellen, dass keine Programmausführung einen unsicheren Zustand erreichen kann, d.h., für alle Zustandssequenzen $s_1, \ldots, s_n$ mit $s_1 \models \mathit{Init}$

und $s_i, s'_{i+1} \models \mathit{Next}$ für alle Paare von aufeinander folgenden Zuständen soll $s_n \models \mathit{Safe}$ gelten. Wir erreichen dies durch Finden einer *induktiven Invariante*—einem Prädikat *Inv* über *Var* so dass (1) $\mathit{Init} \implies \mathit{Inv}$, (2) $\mathit{Inv} \wedge \mathit{Next} \implies \mathit{Inv}'$, und (3) $\mathit{Inv} \implies \mathit{Safe}$ gelten. Die Invariante gilt im Initialzustand, bleibt durch Zustandsübergänge erhalten, und gilt nicht in unsicheren Zuständen. Daher überapproximiert *Inv* die tatsächlich erreichbaren Zustände und trennt diese von den unsicheren Zuständen.

Die modellierung von Programmen als Transitionssysteme ist adäquat für theoretische Betrachtungen, birgt allerdings Probleme für die Verifikation in der Praxis. Das *Next* Prädikat muss den Kontrollfluss des Programms kodieren, was zu einer Fallunterscheidung über alle möglichen Schritte in allen möglichen Zuständen führt. Diese Fallunterscheidung verstärkt sich dann in *Inv*, was das Auffinden und Formulieren von induktiven Invarianten zu einem äußerst komplizierten und mühsamen Unterfangen macht.

Für sequentielle Programme hat Floyd [Fl67] gezeigt wie Programmbeweise durch Annotation des Programmtextes mit *induktiven Zusicherungen* konstruiert werden können. Das resultierende Beweissystem wird heute als *Floyd-Hoare Logik* [Ho69] bezeichnet, und bildet die Basis moderner Programmbeweiser [Ba05]. Aussagen in Floyd-Hoare Logik werden üblicherweise als $\{\varphi\}\, c\, \{\psi\}$ geschrieben, mit der Bedeutung, dass wenn Kommando c in einem Zustand der die *Vorbedingung* φ erfüllt ausgeführt wird und terminiert, dann erfüllt der Endzustand nach der Ausführung von c die *Nachbedingung* ψ.

Owicki und Gries [OG76] erfanden folgende Beweisregel für nebenläufige Programme.

$$\frac{\Psi_1 : \{\varphi_1\}\, c_1\, \{\psi_1\} \quad \Psi_2 : \{\varphi_2\}\, c_2\, \{\psi_2\} \quad \Psi_1, \Psi_2 \text{ interferenzfrei}}{\{\varphi_1 \wedge \varphi_2\}\, c_1 \parallel c_2\, \{\psi_1 \wedge \psi_2\}}$$

Um eine Eigenschaft der parallelen Ausführung von c_1 and c_2 zu beweisen, können wir demnach über beide Kommandos unabhängig schließen. Allerdings müssen die resultierenden Beweise Ψ_1 und Ψ_2 *interferenzfrei* sein, was bedeutet dass keine Aussage in Ψ_1 duch einen Schritt in c_2 invalidiert werden kann, und gleichermaßen für Ψ_2 und c_1. Abbildung 1 zeigt einen Beispielbeweis für $\{x = 0\}\, x := x+1 \parallel x := x+2\, \{x = 3\}$, wo zwei Threads eine gemeinsame Variable x um 1 bzw. 2 erhöhen. Beide Zuweisungen an x werden als atomar angenommen. Im linken Thread wird die Vorbedingung $x = 0$ zu $x = 0 \vee x = 2$ geschwächt, wodurch wir nach $x := x+1$ die Nachbedingung $x = 1 \vee x = 3$ erhalten. Im rechten Thread wird $x = 0$ zu $x = 0 \vee x = 1$ geschwächt, wodurch wir nach $x := x+2$ die Nachbedingung $x = 2 \vee x = 3$ erhalten. Zusammen folgt aus ψ_1 und ψ_2, dass $x = 3$. Die *Interferenzbedingungen* sind rechts in Abbildung 1 gelistet. Zum Beispiel folgt aus $\varphi_1 \wedge \varphi_2$ dass $x = 0$, und daher gilt nach $x := x+2$ dass $x = 2$, was wiederum φ_1 erfüllt.

Es mag überraschend wirken, dass wir trotz Abschwächung von φ_1 und φ_2 die präzise Nachbedingung $x = 3$ aus ψ_1 und ψ_2 wiedergewinnen konnten. Können wir den Beweis aus Abbildung 1 auf folgendes Programm adaptieren, in dem x von beiden Threads um 1 erhöht wird: $\{x = 0\}\, x := x+1 \parallel x := x+1\, \{x = 2\}$. Die Antwort ist nein. Jede Annotation die nur x erwähnt ist entweder zu schwach um die Nachbedingung $x = 2$ zu folgern, oder zu stark um interferenzfrei zu sein. Als Lösung dieses Problems können sogenannte *Hilfsvariablen* für Beweiszwecke in das Programm eingeführt werden [Ow75]. Allerdings müssen diese Hilfsvariablen zumeist die Kontrollpunkte aller Threads kodieren, was

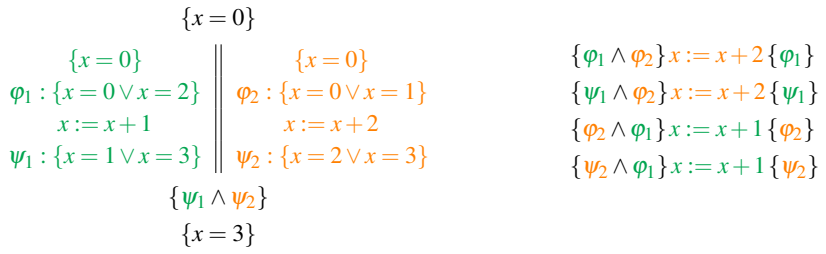


Abbildung 1: Owicki-Gries Beweis eines simplen nebenläufigen Programms.

wiederum (wie bei flachen Transitionssystemen) zu exzessiven Fallunterscheidungen in den Beweisannotationen führt. Dies ist ein massives praktisches Problem für Beweise von komplexen realistischen Programmen.

In dieser Dissertation behandeln wir das Hauptproblem der Verifikation von nebenläufigen Programmen—der Konstruktion von induktiven Invarianten—durch Bereitstellung neuer Techniken und Methodologien zur Unterstützung der interaktiven Beweiskonstruktion. Der Fokus liegt auf Möglichkeiten zur Dekomposition und Strukturierung von Beweisen, wodurch die mentale Aufgabe in überschaubare Teilprobleme zerlegt wird, und der Beweisprozess angenehmer und produktiver gestaltet wird.

Reduktion. Nebenläufige Berechnungen sind normalerweise nicht völlig unabhängig. Trotz vieler möglicher Verzahnungen ist zu erwarten, dass viele dieser Verzahnungen "äquivalent" sind, also zum gleichen Resultat führen. Zwei typische Anwendungen von Nebenläufigkeit sind (1) die Beschleunigung von Berechnungen, und (2) die Zustandsreplikation um Fehler zu tolerieren. In beiden Fällen wird Nebenläufigkeit nicht benötigt um das gewünschte Programmverhalten zu erhalten. Vielmehr muss die Nebenläufigkeit durch geeignete Synchronisation so kontrolliert werden, dass nur "sinnvolles" Verhalten erlaubt ist. Tatsächlich liegen den meisten Konsistenzbedingungen für nebenläufige Programme, wie etwa der *Linearisierbarkeit* [HW90], sequenzielle Spezifikationen zugrunde.

Wir erinnern an das Beispiel aus Abbildung 1. Dieses Beispiel hat nur zwei mögliche Ausführungen, $x := x + 1; x := x + 2$ and $x := x + 2; x := x + 1$. Müssen wir wirklich beide Betrachten, und—im Wesentlichen—die möglichen Zwischenzustände $x = 2$ and $x = 1$ in den Beweisannotationen auflisten? Da die Addition *kommutativ* ist, spielt die Reihenfolge in diesem Beispiel keine Rolle. Wir können eine Reihenfolge wählen und argumentieren, dass die andere Reihenfolge "äquivalent" ist. Um Kommutativität im Allgemeinen auszunutzen um Beweise auf eine Untermenge von Verzahnungen zu reduzieren müssen folgende Fragen beantwortet werden: (1) Wie bestimmen wir die Kommutativität einzelner Operationen? (2) Wie spezifizieren wir die zu betrachtende Untermenge an Verzahnungen? (3) Wie begründen wir, dass alle anderen Verzahnungen implizit abgedeckt sind?

In einer grundlegenden Arbeit stellte Lipton [Li75] das Konzept von *rechts Movern* und *links Movern* vor. In Lipton's ursprünglicher Definition ist eine Operation ein rechts Mover, wenn die Operation in jeder Ausführung nach rechts (d.h. später) über jede Operation

eines anderen Threads permutiert werden kann, ohne den Wert einer Programmvariable im Endzustand zu verändern. Analog ist eine Operation ein links Mover, wenn sie über Operationen in anderen Threads nach links (d.h., früher) permutiert werden kann. Beispielsweise ist das Erwerben eines Locks ein rechts Mover, und das Freigeben eines Locks ein links Mover. *Lipton's Reduktion* ersetzt eine Sequenz $c_1;\ldots;c_n$ mit dem atomaren Kommando $[c_1;\ldots;c_n]$, falls für ein i alle $c_1,\ldots,c_{i-1}$ rechts Mover und alle $c_{i+1},\ldots,c_n$ links Mover sind (c_i is uneingeschränkt). In [FQ03] wurde die theoretische Idee der Mover in ein Typsystem zum Beweis von Atomarität von Methoden in einer nebenläufigen objektorientierten Sprache übernommen. Dieses Typsystem basiert auf der fixen Klassifizierung von bestimmten Operationen als Mover. Die Arbeit am QED Verifizierer [EQT09] stellte das Konzept der *bedingten atomaren Aktionen* vor, welche eine atomare Aktion nicht nur als Menge von Zustandsübergängen ausdrücken, sondern zusätzlich mit einer *Schranke* versehen, welche eine Bedingung angibt die zur Ausführung der Aktion gelten muss. Schranken erfassen kontextbezogene Information, wodurch wir Kommutativitätseigenschaften von atomaren Aktionen in isolation erheben können. Anstatt einer a priori Klassifizierung spezifischer Operationen als Mover, können wir *Movertypen* an bedingte atomare Aktionen durch paarweise Kommutativitätsanalyse zuweisen. Durch die Verwendung von Schranken identifizierten Elmas et al. [EQT09] die *Abstraktion* als symbiotisches Pendant zur Reduktion, wodurch iterative Programmvereinfachung ermöglicht wird. Abstraktion einer atomaren Aktion (d.h., Stärkung der Schranke oder Schwächung der Übergangsrelation) kann deren Movertyp stärken und dadurch Reduktion ermöglichen. Reduktion formt neue grobkörnige atomare Aktionen, welche wiederum abstrahiert werden können um Reduktion erneut anwendbar zu machen.

Reduktion erleichtert die Konstruktion von induktiven Invarianten, da Invarianten für ein reduziertes Programm um ein Vielfaches einfacher sein können als für das ursprüngliche Programm, was den Aufwand für die Reduktion mehr als kompensiert. In dieser Dissertation präsentieren wir neue reduktionsbasierte Beweisregeln für asynchrone Programme und verteilte Systeme. Dadurch ermöglichen wir Reduktionsbeweise für eine völlig neu Klasse von Programmen.

Verfeinerung. Programmentwicklung durch *schrittweise Verfeinerung* ist die Idee zur Entwicklung eines Programms durch sukzessive Verfeinerung einer abstrakten Spezifikation zu einer konkreten Implementierung. Alternative kann eine konkrete Implementierung sukzessive abstrahiert werden, um eine abstrakte Spezifikation zu beweisen. Oder der top-down und bottom-up Ansatz wird kombiniert. Formale Verifikation durch schrittweise Verfeinerung wird, in der Theorie, seit langem zur Konstruktion verifizierter nebenläufiger Programme vorgeschlagen (z.B. [BvW98, Sc97, Ro01]).

Zurück zum Modell der Transitionssysteme, sei K ein konkretes und A ein abstraktes Transitionssystem. Ein Standardansatz zum Beweis dass K eine korrekte Implementierung von A ist bedient sich einer *Verfeinerungsfunktion* von konkreten Zuständen von K zu abstrakten Zuständen von A. In einer sogenannten *Vorwärtssimulation* wird gezeigt, dass die Verfeinerungsfunktion jedem konkreten Schritt in K einen abstrakten Schritt in A zweist. Dadurch wird jedes konkrete Verhalten durch ein abstraktes Verhalten gerechtfertigt.

Die Arbeit in dieser Dissertation ist im Kontext des Verifikationssystems CIVL, erstmals Beschrieben von Hawblitzel et al. [Ha15]. Als reduktionsbasierter Verifizierer verfechtet CIVL die Verfeinerung von nebenläufigen Programmen über mehrere *Abstraktionsschichten* hinweg. Ein charakteristisches Designmerkmal von CIVL ist, dass alle Schichten in einem Verfeinerungsbeweis *strukturierte Programme* bleiben, d.h., Programme mit Prozeduren, imperativem Kontrollfluss, strukturellem Parallelismus, und asynchroner Nebenläufigkeit. Dies ist im Gegensatz zu bisherigen verfeinerungsbasierten Verifikationssystemen, wie TLA+ [La02b] oder Event-B [Ab96], welche nebenläufige Systeme als flache Transitionssysteme repräsentieren. Zwei Vorteile der Repräsentation als strukturierte Programme sind (1) der natürliche Brückenschluss von abstrakten Modellen zu realen Implementierungen, und (2) der Erhalt der im Programmtext enthaltenen Strukturinformation. In CIVL entspricht jeder Beweisschritt einer kleinen, vereinfachenden Prgrammtransformation, welche die atomaren Aktionen im Programm immer grobkörniger und abstrakter, und das Programm insgesamt immer weniger nebenläufig machen. Die benötigten Invarianten zur Begründung einzelner Beweisschritte sind vergleichsweise simpel, und die ermöglichte Dekomposition macht Beweise einfacher zu konstruieren und wiederzuverwenden. Die Implementierung von CIVL übersetzt das Verifikationsproblem in eine Menge an modularen Verifikationsbedingunen, welche von einem automatischen Theorembeweiser überprüft werden.

3 Beiträge der Dissertation

In dieser Dissertation beschäftigen wir uns mit formalen Methoden, die Entwickler und Programmierer dabei unterstützen, zuverlässigere nebenläufige Systeme zu entwickeln. Wir stellen neue Techniken, Methodiken, und Werkzeuge zur rigorosen Analyse und Verifikation solcher Systeme vor. Grob gesagt fallen die Beiträge dieser Dissertation in zwei Kategorien. Als erstes präsentieren Kapitel 2 und Kapitel 3 die formalen Grundlagen zur Entwicklung eines verfeinerungsbasierten Verifizierers, welcher alle Abstraktionsschichten als strukturierte Programme repräsentiert. Als zweites präsentieren Kapitel 4 and Kapitel 5 neue reduktionsbasierte Beweisregeln, welche neuartige simple Beweise für asynchrone Programme ermöglichen. Jedes Kapitel entspricht einem Forschungspapier, welches bei einer top-tier Konferenz veröffentlicht wurde.[3]

Kapitel 2: Layered Concurrent Programs [KQ18]
Bernhard Kragl, Shaz Qadeer
CAV 2018 (30th International Conference on Computer Aided Verification)
Konferenzrang: A (ERA) / A1 (Qualis)

Kapitel 3: Refinement for Structured Concurrent Programs [KQH20]
Bernhard Kragl, Shaz Qadeer, Thomas A. Henzinger
CAV 2020 (32nd International Conference on Computer Aided Verification)
Konferenzrang: A (ERA) / A1 (Qualis)

[3] Die Quelle für Konferenzreihungen ist `http://www.conferenceranks.com`. ERA reiht von A (=am besten) bis C (=am schlechtesten). Qualis reiht nach A1 (=am besten), A2, B1, ..., B5 (=am schlechtesten).

Kapitel 4: Synchronizing the Asynchronous [KQH18]
Bernhard Kragl, Shaz Qadeer, Thomas A. Henzinger
CONCUR 2018 (29th International Conference on Concurrency Theory)
Konferenzrang: A (ERA) / A2 (Qualis)

Kapitel 5: Inductive Sequentialization of Asynchronous Programs [Kr20b]
Bernhard Kragl, Constantin Enea, Thomas A. Henzinger, Suha O. Mutluergil, Shaz Qadeer
PLDI 2020 (41st ACM SIGPLAN Conference on Programming Language Design and Implementation)
Konferenzrang: A (ERA) / A1 (Qualis)

Zusammengefasst leistet diese Dissertation [Kr20a] folgende technische Beiträge.

Layered Concurrent Programs. Wir präsentieren *Layered Concurrent Programs* (dt. *geschichtete nebenläufige Programme*) (Kapitel 2), einen kompakten Formalismus zur Repräsentation aller Programme in einem mehrschichtigen Verfeinerungsbeweis als eine einzige syntaktische Einheit. Dadurch wird exzessive Duplikation von jenen Programmteilen vermieden, welche sich in einzelnen Beweisschritten nicht ändern.

Yield Invariants. Wir präsentieren *Yield Invariants* (Kapitel 3), ein neues Spezifikationsidiom, welches induktive Invarianten benennt, parametrisiert, und als wiederverwendbare Einheiten zusammenfasst. So können Yield Invariants spezifisch für eine bestimmte Aufrufstelle instanziiert werden (ähnlich wie Prozeduren). Yield Invariants kombinieren die Präzision von Invarianten à la Owicki-Gries und die Kompaktheit von Rely-Guarantee Spezifikationen [Jo83]. Die Portierung existierende Beispiele zur Benutzung von Yield Invariants ergab signifikant Verbesserungen von Beweiskomplexität und Performance.

Verfeinerung über strukturierte Programme. Wir präsentieren eine mächtige Beweisregel zur schrittweisen Verfeinerung, welche das Verifikationsproblem über strukturierte Programme in modulare Verifikationsbedingungen zerlegt (Kapitel 3). Diese Beweisregel integriert Yield Invariants mit einem flexiblen System für *linear Berechtigungen* zur Verbesserung der Beweislokalität. Wir unterstützen sowohl die *Einführung* als auch die *Elimination* von lokalen und globalen Variablen, sowie die modulare Abstraktion von rekursiven Prozeduren.

Synchronization. Wir präsentieren ein Reduktionsprinzip namens *Synchronization* (dt. *Synchronisation*) (Kapitel 4), welches die Umwandlung von asynchronen Prozeduraufrufen zu synchronen Prozeduraufrufen ermöglicht. Dadurch wird das Verifikationsproblem erheblich erleichtert, da wir nicht mehr über asynchrone Berechnungsschritte zu einem beliebigen späteren Zeitpunkt in einer Berechnung schließen müssen, sonder (für Beweiszwecke) so tun können, als ob die Berechnung sofort ausgeführt wird.

Inductive Sequentialization. Wir präsentieren ein Reduktionsprinzip namens *Inductive Sequentialization* (dt. *Induktive Sequenziellisierung*) (Kapitel 5), welches das Schließen über ein verteiltes System auf eine einzige sequenzielle Ausführung des Systems reduziert. Wir zeigen, dass selbst komplizierte Protokolle wie Paxos simple sequenzielle Reduktionen erlauben. Unsere Beweise mittels Inductive Sequentialization sind um ein Vielfaches

einfacher als existierende Beweise mittels standard induktiver Invarianten, da Inductive Sequentialization das Problem umgeht, über beliebig viele und beliebig lange verzahnte Ausführungen zu Schließen.

Pending Asyncs. Wir erweitern bedingte atomare Aktionen mit der Idee von *Pending Asyncs* (dt. *ausstehende asynchrone Aktionen*). Mittels Pending Asyncs spezifizieren atomare Aktionen nicht nur Zustandsänderungen globaler Variablen, sondern auch die Erzeugung asynchroner Berechnungen. Pending Asyncs wurden erstmals für die Arbeit an Synchronization präsentiert, und bildeten anschließend die technische Grundlage für Inductive Sequentialization. Die Beweisregel in Kapitel 3 beschreibt die "Erzeugung" von Pending Asyncs, wohingegen Kapitel 4 und Kapitel 5 Techniken zur "Eliminierung" von Pending Asyncs beschreiben.

CIVL Verifizierer. Alle Techniken in dieser Dissertation wurden in unserem Verifizierungssystem CIVL[4] implementiert, welches als Teil von Boogie[5] frei verfügbar ist. Konkret bildete die Theorie aus Kapitel 2 und Kapitel 3 die Basis für ein neues Design und Implementierung von CIVL, und die Techniken aus Kapitel 4 and Kapitel 5 sind als neue Beweistaktiken verfügbar, welche symbiotisch mit den bereits existierenden Taktiken integriert wurden. Mittels unserer Implementierung demonstrierten wir in zahlreichen Fallstudien die Anwendbarkeit und Nützlichkeit unserer Verifikationsmethodik.

4 Ausblick

In dieser Dissertation präsentierten wir einen neuen Ansatz zur deduktiven Verifikation nebenläufiger Programme. Die vorgestellte Verfeinerungsmethodik über strukturierte nebenläufige Programme ermöglicht die schrittweise Abstraktion von feinkörniger Prozeduren zu grobkörnigen atomaren Aktionen. Die Konstruktion von formalen Beweisen durch Benutzer wird in überschaubare Teilprobleme untergliedert, und die Beweisprüfung durch einen Computer wird in modulare Verifikationsbedingungen übersetzt. Wir sind überzeugt, dass formal verifizierte Implementierungen nur dann gängige Praxis werden können, wenn Programmierung und Verifikation in eine vereinte Aktivität zusammengeführt wird. Unsere Arbeit fördert diese Zusammenführung durch die einheitliche Repräsentation aller Abstraktionsschichten eines mehrschichtigen Verfeinerungsbeweises (von konkreter Implementierung zu abstrakter Spezifikation) im selben Formalismus, und die kompakte Repräsentation aller Abstraktionsschichten und deren Zusammenhang in einem einzigen geschichteten nebenläufigen Programm.

Wir integrierten neuartige reduktionsbasierte Programmvereinfachungen in unsere Methodik, welche asynchrone Programmausführungen *synchronisieren* bzw. *sequenziellisieren*, und dadurch die Intuition von Programmierern über simple Verzahnungen ausnutzen. Wir haben unser Verifikationsverfahren in zahlreichen anspruchsvollen Fallstudien angewendet und dabei demonstriert, dass unser Verfahren viel einfachere Beweise erlaubt als bisher bekannte Beweise. Besonders entscheidend ist, dass diese Vereinfachung nicht nur die

[4] https://civl-verifier.github.io
[5] https://github.com/boogie-org/boogie

Komplexität des finalen Beweises betrifft, sondern die intellektuelle Herausforderung der eigentlichen Beweiskonstruktion!

Insgesamt stellt diese Dissertation einen bedeutenden Fortschritt auf dem Stand der Technik der Programmierung zuverlässiger nebenläufigen und verteilten Systemen dar. Einige unabhängige Forscher publizierten bereits Artikel unter der Verwendung unseres Verifizierers CIVL. Obgleich die formale Verifikation von anspruchsvollen nebenläufigen Algorithmen und realistischen Implementierungen durchaus eine Herausforderung bleiben wird, so ermöglicht es unser Dekompositions- und Strukturierungsmechanismus, dass sich Programmierer dieser Herausforderung bestens gewappnet stellen können.

Literaturverzeichnis

[Ab96] Abrial, Jean-Raymond: The B-book - assigning programs to meanings. 1996.

[Ba05] Barnett, Michael; Chang, Bor-Yuh Evan; DeLine, Robert; Jacobs, Bart; Leino, K. Rustan M.: Boogie: A Modular Reusable Verifier for Object-Oriented Programs. In: FMCO. 2005.

[BvW98] Back, Ralph-Johan; von Wright, Joakim: Refinement Calculus - A Systematic Introduction. Graduate Texts in Computer Science. 1998.

[CGR07] Chandra, Tushar Deepak; Griesemer, Robert; Redstone, Joshua: Paxos made live: an engineering perspective. In: PODC. 2007.

[EQT09] Elmas, Tayfun; Qadeer, Shaz; Tasiran, Serdar: A calculus of atomic actions. In: POPL. 2009.

[Fl67] Floyd, Robert W.: Assigning Meanings to Programs. Proceedings of Symposium on Applied Mathematics, 19, 1967.

[FQ03] Flanagan, Cormac; Qadeer, Shaz: A type and effect system for atomicity. In: PLDI. 2003.

[Ha15] Hawblitzel, Chris; Petrank, Erez; Qadeer, Shaz; Tasiran, Serdar: Automated and Modular Refinement Reasoning for Concurrent Programs. In: CAV. 2015.

[Ho69] Hoare, C. A. R.: An Axiomatic Basis for Computer Programming. Commun. ACM, 12(10), 1969.

[HW90] Herlihy, Maurice; Wing, Jeannette M.: Linearizability: A Correctness Condition for Concurrent Objects. ACM Trans. Program. Lang. Syst., 12(3), 1990.

[Jo83] Jones, Cliff B.: Specification and Design of (Parallel) Programs. In: IFIP Congress. 1983.

[KQ18] Kragl, Bernhard; Qadeer, Shaz: Layered Concurrent Programs. In: CAV. 2018.

[KQH18] Kragl, Bernhard; Qadeer, Shaz; Henzinger, Thomas A.: Synchronizing the Asynchronous. In: CONCUR. 2018.

[KQH20] Kragl, Bernhard; Qadeer, Shaz; Henzinger, Thomas A.: Refinement for Structured Concurrent Programs. In: CAV. 2020.

[Kr20a] Kragl, Bernhard: Verifying Concurrent Programs: Refinement, Synchronization, Sequentialization. Dissertation, IST Austria, 2020.

[Kr20b] Kragl, Bernhard; Enea, Constantin; Henzinger, Thomas A.; Mutluergil, Suha Orhun; Qadeer, Shaz: Inductive sequentialization of asynchronous programs. In: PLDI. 2020.

[La98] Lamport, Leslie: The Part-Time Parliament. ACM Trans. Comput. Syst., 16(2), 1998.

[La02a] Lamport, Leslie: Paxos Made Simple, Fast, and Byzantine. In: OPODIS. 2002.

[La02b] Lamport, Leslie: Specifying Systems, The TLA+ Language and Tools for Hardware and Software Engineers. 2002.

[Li75] Lipton, Richard J.: Reduction: A Method of Proving Properties of Parallel Programs. Commun. ACM, 18(12), 1975.

[MP90] Manna, Zohar; Pnueli, Amir: A Hierarchy of Temporal Properties. In: PODC. 1990.

[OG76] Owicki, Susan S.; Gries, David: Verifying Properties of Parallel Programs: An Axiomatic Approach. Commun. ACM, 19(5), 1976.

[OO14] Ongaro, Diego; Ousterhout, John K.: In Search of an Understandable Consensus Algorithm. In: USENIX Annual Technical Conference. USENIX Association, 2014.

[Ow75] Owicki, Susan S.: Axiomatic Proof Techniques for Parallel Programs. Dissertation, Cornell University, 1975.

[Ro01] de Roever, Willem P.; de Boer, Frank S.; Hannemann, Ulrich; Hooman, Jozef; Lakhnech, Yassine; Poel, Mannes; Zwiers, Job: Concurrency Verification: Introduction to Compositional and Noncompositional Methods. Cambridge Tracts in Theoretical Computer Science. 2001.

[Sc97] Schneider, Fred B.: On Concurrent Programming. Graduate Texts in Computer Science. 1997.

[St01] Stoica, Ion; Morris, Robert Tappan; Karger, David R.; Kaashoek, M. Frans; Balakrishnan, Hari: Chord: A scalable peer-to-peer lookup service for internet applications. In: SIGCOMM. 2001.

[vRA15] van Renesse, Robbert; Altinbuken, Deniz: Paxos Made Moderately Complex. ACM Comput. Surv., 47(3), 2015.

[Za12] Zave, Pamela: Using lightweight modeling to understand Chord. Comput. Commun. Rev., 42(2), 2012.

Bernhard Kragl ist ein Applied Scientist in der S3 Automated Reasoning Group bei Amazon Web Services (AWS). Seine Forschungsinteressen sind Programmiersprachen und formale Methoden für die Entwicklung zuverlässiger Computersysteme. Er promovierte am IST Austria unter der Betreuung von Thomas A. Henzinger. Seine Dissertation beschäftigt sich mit Beweistechniken für nebenläufige und verteilte Systeme. Zuvor schloss er mit Arbeiten zum automatischen Schließen und Theorembeweisen ein Bachelor- und Masterstudium an der Technischen Universität Wien ab.

CogniCrypt— Sichere Integration Kryptographischer Software[1]

Stefan Krüger[2]

Abstract: Empirische Studien zeigen, dass Fehlbenutzungen von Crypto APIs weit verbreitet sind. Die Literatur liefert mehrere Ansätze, diese zu beheben, aber keiner adressiert das Problem vollständig. Das Resultat ist eine lückenhafte Landschaft verschiedener Ansätze. In dieser Arbeit adressiere ich das Problem solcher Fehlbenutzungen systematisch durch CogniCrypt. CogniCrypt integriert verschiedene Arten von Tool Support in einen Ansatz, der Entwickler davon befreit, wissen zu müssen, wie die APIs benutzt werden. Zentral für meinen Ansatz ist CrySL, eine Beschreibungssprache, mit der spezifiziert wird, wie APIs benutzt werden. Ich habe einen Compiler für CrySL und zwei darauf aufsetzende Supporttools, die Code-Analyse CogniCrypt$_{SAST}$ und den Code-Generator CogniCrypt$_{GEN}$, entwickelt. Schlussendlich habe ich CogniCrypt prototypisch implementiert und in einer empirischen Evaluierung die Effektivität von CogniCrypt gezeigt.

1 Einleitung

Digitale Geräte werden vielfach verwendet um sensitive Daten zu speichern. Kryptographie ist das Hauptwerkzeug um solche Daten vor Fälschung zu schützen. Damit dieser Schutz effektiv ist, müssen Algorithmen nicht nur konzeptionell sicher und korrekt implementiert sein, sondern auch korrekt in den Anwendungscode integriert werden. Das scheint jedoch oft nicht zu funktionieren. Lazar et al. [La14] untersuchten 269 kryptographische Schwachstellen und stellten fest, dass 83% dieser von Entwicklern verursacht werden, die Algorithmen unsicher integrieren. Nachfolgende Forschung zu diesem Phänomen kam zu dem Schluss, dass das Problem von signifikanter Größe ist [Eg13, Gu19, Fe19, Ve17]. Die Gründe hinter jedem individuellen Fehler mögen mannigfaltig sein. Das Design der kryptographischen Bibliotheken und ihrer Anwendungsschnittstellen (API) sind aber wohl zentral für die Schwierigkeiten von Entwicklern. Das zeigt auch folgendes Beispiel.

1.1 Ein Motivierendes Beispiel

Die Java Cryptography Architecture (JCA) [Or20], Javas Haupt-Crypto-API, stellt die API `PBEKeySpec` zur passwort-basierten Generierung von Schlüsseln bereit. Abbildung 1 zeigt eine Nutzung von `PBEKeySpec` wie sie in unzähligen realen Programmen zu finden ist.

Das Codebeispiel erstellt zunächst ein `PBEKeySpec`-Objekt durch den Aufruf des Konstruktors. Dieser erwartet ein Passwort, einen Salt, eine Iterationszahl und die Schlüsselänge.

[1] Englischer Originaltitel der Dissertation: "CogniCrypt– The Secure Integration of Cryptographic Software"

[2] krueger@cqse.eu

```
1  String pwd = "password"; // Festverdrahtet zur Demonstration
2  byte[] salt = {15, -12, 94, 0, 12, 3, -65, 73, -1, -84, -35};
3  PBEKeySpec spec = new PBEKeySpec(pwd.toCharArray(), salt, 100000, 256);
4
5  SecretKeyFactory skf =
       SecretKeyFactory.getInstance("PBKDF2WithHmacSHA256");
6  byte[] keyMaterial = skf.generateSecret(spec).getEncoded();
7
8  SecretKeySpec cipherKey = new SecretKeySpec(keyMaterial, "AES");
```

Abb. 1: Beispiel einer *inkorrekten* passwortbasierten Schlüsselgenerierung in Java.

Das `PBEKeySpec`-Objekt wird an ein `SecretKeyFactory`-Objekt weitergereicht, das den Schlüsselableitungsalgorithmus `PBKDF2` verwendet um Schlüsselmaterial zu generieren (Zeilen 5–6). In der letzten Zeile erstellt das Code-Beispiel daraus dann den Schlüssel.

Das Beispiel enthält mehrere Fehler um den Konstruktor-Aufruf von `PBEKeySpec`. Der erste liegt beim Salt. Dieser muss zufällig und unvorhersehrbar sein [Bu17], im Codebeispiel ist er aber festverdrahtet. Der zweite Fehler betrifft das Passwort. Der Konstruktor erwartet das Passwort aus gutem Grund als Char-Array: Passwörter sollten nicht länger im Speicher bleiben als notwendig, das heißt, sie sollten aufgeräumt werden. Strings sind allerdings unveränderbar, sie aufzuräumen ist unmöglich. Zuletzt fehlt im Code ein Aufruf der Methode `clearPassword()`. Trotz der Fehler läuft der Code ohne Exceptions.

1.2 Beiträge Dieser Arbeit

Die Probleme im gezeigten Code-Beispiel mögen wie Einzelfälle aussehen, wie die obige Forschung aber zeigt, sind sie es aber nicht. Eine Vielzahl unterschiedlicher Ansätze hat versucht das Problem anzugehen. Neben den oben referenzierten Studien und deren Programmanalysen wurde unter Anderem versucht, fehlerhafte Programme zu reparieren, Hilfsmittel für APIs zu erweitern oder bessere Designs für APIs zu entwickeln. Diese Ansätze sind hilfreich: Sie fördern das Verständnis über gute API-Nutzbarkeit, sie verringern bestehende Probleme oder können den Weg ebnen für neue APIs. Am Ende liefern sie allerdings keine systematische Lösung. Meine Arbeit geht daher einen anderen Weg.

Man sieht, dass Anwendungsentwickler einerseits sofortige Hilfe benötigen. Daher unterstützt meine Lösung Entwickler, wenn sie die APIs nutzen. Es ist weiterhin klar, dass Unterstützung aus verschiedenen Richtungen kommen muss. Nur eine Programmanalyse oder nur bessere Hilfsmittel genügen nicht. Stattdessen präsentiere ich eine Lösung, die mehrere Formen von Unterstützung integriert. Meine Arbeit hat insgesamt folgendes Ziel:

Fehlbenutzungen von Crypto APIs können signifikant reduziert werden durch einen integrierten Ansatz, der Entwickler davon befreit, wissen zu müssen, wie die APIs genutzt werden müssen.

Zur Realisierung dieses Ziels präsentiere ich den integrierten Ansatz CogniCrypt. Ich habe CogniCrypt als Plugin für die IDE Eclipse entworfen und prototypisch implementiert, damit es sich in den Arbeitsablauf von Entwicklern integriert. CogniCrypt besteht

aus mehreren Komponenten. Dem Ansatz zugrunde liegt die textuelle Spezifikationssprache CRYSL (Abschnitt 3). Mit Hilfe von CRYSL können Domänen-Experten definieren, wie eine API benutzt werden soll. Ich habe die JCA in CRYSL modelliert. Auf Basis von CRYSL können verschiedene Formen von Tool-Unterstützung gebaut werden.

Eine Form von Tool Support liefert COGNICRYPT$_{SAST}$ (Abschnitt 4), eine statische Analyse, die ein Java- oder Android-Programm auf ihre Befolgung von CRYSL-Regeln hin überprüft. Ich habe weiterhin den Code-Generator COGNICRYPT$_{GEN}$ (Abschnitt 5) entwickelt. Dieser generiert anwendungsfallspezifischen sicheren Code für Crypto APIs.

Zuletzt habe ich die Effektivität von COGNICRYPT in einem kontrollierten Experiment evaluiert (Abschnitt 6). Die Ergebnisse zeigen, dass Entiwckler mit COGNICRYPT sowohl sichereren als auch funktionaleren Code produzieren. Ausführlich habe ich meine Arbeit und ihre Ergebnisse in meiner Dissertation [Kr20] festgehalten.

2 COGNICRYPT

Ich präsentiere COGNICRYPT hier zunächst aus einer Nutzerperspektive um zu demonstrieren, wie es Entwickler unterstützt. Meine prototypische Implementierung für Eclipse ist öffentlich verfügbar und open-source[3]. Aktuell unterstützt der Prototyp nur Java, die vorgestellten Konzepte sind allerdings nicht auf Java beschränkt. Das Eclipse-Plugin COGNICRYPT kombiniert die beiden Ansätze COGNICRYPT$_{GEN}$ und COGNICRYPT$_{SAST}$.

Der Code-Generator COGNICRYPT$_{GEN}$ kann *sicheren* Code für unter anderem das Beispiel in Abbildung 1 generieren. Dazu klickt der Nutzer den COGNICRYPT$_{GEN}$-Button in der Toolbar von Eclipse und wählt im erscheinenden Dialogfenster den passenden Anwendungsfall aus. Anschließend beantwortet er mehrere Fragen, die COGNICRYPT$_{GEN}$ zur Konfiguration verwendet. Zuletzt wählt er die Klasse aus, in die COGNICRYPT Code generiert. Im Anschluss generiert COGNICRYPT$_{GEN}$ zwei Artefakte. Einmal generiert es die eigentliche Implementierung für den ausgewählten Anwendungsfall in das Paket `de.cognicrypt.crypto`. Dazu generiert COGNICRYPT eine Methode `templateUsage()`, die zeigt, wie die Implementierung aufgerufen wird, in die vorher ausgewählte Klasse.

COGNICRYPT benachrichtigt Nutzer über Fehlbenutzungen von Crypto APIs, indem es kontinuierlich die Analyse COGNICRYPT$_{SAST}$ laufen lässt. COGNICRYPT$_{SAST}$ stellt sicher, dass die API-Nutzungen sicher bleiben, wenn Entwickler sie ändern. Darüber hinaus hilft COGNICRYPT$_{SAST}$ auch, wenn Entwickler APIs direkt, also ohne COGNICRYPT$_{GEN}$, nutzen. Bei einer Fehlbenutzung generiert COGNICRYPT einen Eclipse Error Marker.

3 CrySL

Damit COGNICRYPT weiß, wie Crypto APIs verwendet werden, benutzt es Regeln in CRYSL. CRYSL ist eine Spezifikationssprache, die Experten in die Lage versetzt zu de-

[3] `www.cognicrypt.org`

finieren, wie eine API genutzt wird. Die Sprache folgt größtenteils einem White-Listing-Ansatz und die Syntax ist bewusst einfach und Java-nah gehalten.

3.1 Ausdruckskraft von CrySL

Eine Regel besteht aus bis zu vier Abschnitten: (1) Ein Methoden-Sequenz-Muster, (2) Nutzungsauflagen für Methodenparameter, (3) Verbotene Methoden und (4) Nutzungsauflagen für Interaktionen verschiedener Klassen.

```
9  SPEC javax.crypto.spec.PBEKeySpec
10 OBJECTS
11   char[] password;
12   byte[] salt;
13   int iterationCount;
14   int keylength;
15
16 ...
17 EVENTS
18   c1: PBEKeySpec(password, salt, iterationCount, keylength);
19   cP: clearPassword();
20
21 ORDER
22   c1,  cP
23
24 CONSTRAINTS
25   iterationCount >= 10000;
26   neverTypeOf(pwd, java.lang.String);
27
28 REQUIRES
29   randomized[salt];
30 ENSURES
31   speccedKey[this, keylength] after c1;
32 NEGATES
33   speccedKey[this, _];
```

Abb. 2: CrySL-Regel für JCA-Klasse `javax.crypto.spec.PBEKeySpec`.

Zur Illustration von CrySL zeige ich in Abbildung 2 die CrySL-Regel von `PBEKeySpec`. Die Regel definiert unter dem Schlüsselwort `SPEC`, welche Klasse sie spezifiziert. Unter `OBJECTS` listet sie dann die vier in der Regel verwendeten Objekte. Eines davon ist das `char[]`-Objekt `password`. Die Abschnitte `EVENTS` und `ORDER` spezifizieren das Methoden-Sequenz-Muster. Dazu werden unter `EVENTS` alle Methoden, die zur korrekten Verwendung von `PBEKeySpec` beitragen können, als Methoden-Event-Muster definiert (Zeilen 18–19). Der `ORDER`-Abschnitt definiert anschließend valide Aufrufsequenzen der Methoden-Events als regulären Ausdruck. Mit Hilfe von Labels (z.B. `cP`) vereinfacht CrySL die spätere Referenz auf die Methoden. Das Muster für `PBEKeySpec` ist relativ einfach: Der Konstruktor `c1` muss einmal aufgerufen werden, gefolgt von einem Aufruf von `cP`. Der Abschnitt `CONSTRAINTS` fügt die oben erwähnten Nutzungsauflagen für Parameter hinzu.

Zur Definition von Auflagen für Interaktionen von Klassen bietet CrySL drei weitere Schlüsselwörter an, die einen Rely/Guarantee-Mechanismus implementieren. Zuerst ist da `ENSURES`, mit dem eine Klasse eine Garantie ausgibt, wenn sie richtig verwendet wurde. `PBEKeySpec` garantiert `speccedKey`, also dass ein korrekt instantiiertes `PBEKeySpec`-

Objekt sicheres Schlüsselmaterial bereitstellt. Die Regel nutzt das Schlüsselwort `after` um anzuzeigen, dass die Garantie nach dem Aufruf des Konstruktors gegeben ist. Über das Schlüsselwort `NEGATES` zerstört `PBEKeySpec` seine `speccedKey`-Garantie nach dem Aufruf von `cP` wieder. Eine Garantie zu verlangen geschieht über das Schlüsselwort `REQUIRES`. `PBEKeySpec` selbst verlangt, dass sein Salt von einer anderen Klasse randomisiert wurde. Ähnlich können auch Regeln die Garantie `speccedKey` verlangen. Insgesamt deckt die CrySL-Regel von `PBEKeySpec` alle oben diskutierten Fehlbenutzungen ab.

3.2 Implementierung

Ich habe die gesamte JCA in insgesamt 23 CrySL-Regeln modelliert. Darüber hinaus wurden unabhängig von dieser Arbeit noch die APIs Google Tink, Bouncy Castle und Bouncy Castle als JCA-Provider in CrySL modelliert. Weiterhin habe ich einen auf dem Sprachentwicklungswerkzeug XText basierenden Compiler für CrySL implementiert. Xtext stellt auf Basis der Grammatik von CrySL grundlegenden Support wie einen Editor bereit. Ich habe auch einen Parser entwickelt, der CrySL-Regeln in ein Java-Objekt-Modell übersetzt, das von auf CrySL aufbauendem Tool-Support verwendet werden kann und von CogniCrypt$_{SAST}$ und CogniCrypt$_{GEN}$ verwendet wird.

4 CogniCrypt$_{SAST}$

CogniCrypt$_{SAST}$ ist eine auf CrySL aufsetzende statische fluss- und kontextsensitive Datenflussanalyse. Als Eingabe erhält CogniCrypt$_{SAST}$ neben dem zu analysierenden Java- oder Android-Programm eine Menge von CrySL-Regeln. Im Folgenden analysiert CogniCrypt$_{SAST}$ dann, ob das Programm die CrySL-Regeln einhält.

4.1 Ablauf & Prototypische Implementierung

Zunächst parst CogniCrypt$_{SAST}$ die Regeln mit dem CrySL-Compiler. Anschließend überführt CogniCrypt$_{SAST}$ das Zielprogramm mit Hilfe des Analyseframeworks Soot in die Zwischenrepräsentation Jimple und erstellt einen Call Graph vom Zielprogramm. Den Call Graph durchsucht CogniCrypt$_{SAST}$ im Anschluss nach verbotenen Methoden.

Im Anschluss ermittelt CogniCrypt$_{SAST}$ in einer Prä-Analyse die zu analysierenden Objekte, wie das `PBEKeySpec`-Objekt in Abbildung 1. Objekte werden hier über Allocation Sites approximiert. Dabei zeichnet es ebenfalls die Methodenaufrufe auf den Objekten und die potentiellen Laufzeitwerte deren Parameter auf. Ich habe einen Constraint Solver entwickelt, der diese Werte mit jenen in der entsprechenden CrySL-Regel vergleicht. Für `PBEKeySpec` in Abbildung 1 gilt das beispielsweise für die Nutzungsauflage für `iterationCount`. Um die Einhaltung der Methoden-Sequenz-Muster zu überprüfen, übersetzt CogniCrypt$_{SAST}$ diese in Zustandsautomaten und führt mit diesem und den

aufgezeichneten Methodenaufrufen über das Datenflussanalyseframework IDE^{al} eine Typestate-Analyse durch. Für `PBEKeySpec` in Abbildung 1 findet CogniCrypt$_{SAST}$ einen Typestate-Fehler, da es `clearPassword()` nicht aufruft, um das Passwort aufzuräumen.

Es gibt auch für die Interaktion von Klassen eine eigene Analyse in CogniCrypt$_{SAST}$, die dem in Abschnitt 3 beschriebenen Ablauf folgt. Für die richtige Auflösung der Garantien und Versicherungen, hält CogniCrypt$_{SAST}$ während der Analyse eine Liste von Garantien vor. Wenn CogniCrypt$_{SAST}$ beispielsweise das `PBEKeySpec`-Objekt in Abbildung 1 analysiert, überprüft es, ob die Garantie vorliegt, dass `salt` randomisiert ist. Wenn das der Fall ist und das `PBEKeySpec`-Objekt keine weiteren Teile der Regeln verletzen würde, generierte CogniCrypt$_{SAST}$ auch dessen Garantie. Da das Objekt jedoch, wie oben dargelegt, mehrfach falsch benutzt wird, stellt CogniCrypt$_{SAST}$ keine Garantie aus.

4.2 Evaluierung

Für die Evaluierung von CogniCrypt$_{SAST}$ habe ich diese Forschungsfragen betrachtet:

RQ1 Wie hoch sind Präzision und Recall von CogniCrypt$_{SAST}$?
RQ2 Welche Arten von Fehlbenutzungen findet CogniCrypt$_{SAST}$?
RQ3 Wie schnell läuft CogniCrypt$_{SAST}$?
RQ4 Wie gut schneidet CogniCrypt$_{SAST}$ ab im Vergleich mit dem Stand der Technik?

4.2.1 Setup

Zur Beantwortung habe ich CogniCrypt$_{SAST}$ mit dem JCA-Regelsatz auf 10.000 Android-Apps angewendet. Zur Beantwortung von **RQ1** habe ich 50 der Apps zufällig ausgewählt und die Typestate- und Parameterergebnisse von CogniCrypt$_{SAST}$ manuell verifiziert. Um **RQ3** zu beantworten, habe ich die Analysedauer für alle Apps gemessen. Für **RQ4** habe ich CryptoLint [Eg13] auch auf die 10,000 Apps angewendet. Ich hatte keinen Zugriff auf das Tool, sondern haben das Tool simuliert, indem ich CogniCrypt$_{SAST}$ mit CrySL-Regeln, die denen von CryptoLint entsprachen, laufen ließ.

4.2.2 Ergebnisse

RQ1: CogniCrypt$_{SAST}$ findet insgesamt 156 Fehlbenutzungen in den 50 Apps. Von den laut CogniCrypt$_{SAST}$ insgesamt 27 Typestate-Fehlern konnte ich 25 bestätigen. CogniCrypt$_{SAST}$ übersieht aufgrund der verwendeten Alias-Analyse aber vier Typestate-Fehler. CogniCrypt$_{SAST}$ findet 129 Fehler bei Nutzungsauflagen für Parameter, 19 davon sind falsch positiv. Ich konnte keine Falschnegativen finden. Insgesamt ergeben sich daher für die Typestate-Analyse 92.6% Präzision und 86.2% Recall, für die Analyse der Nutzungsauflagen fü Parameter dagegen 85.3% Präzision und ein Recall von 100%.

RQ2: CogniCrypt$_{SAST}$ findet in 4.439 Apps JCA-Nutzungen, in 95% davon mindestens einen Fehler. Mit 3.955 Apps treten Fehler bei den Nutzungsauflagen für Parameter am häufigsten auf, gefolgt von 2.896 Apps mit Typestate-Fehlern. Unsichere Interaktionen von Klassen kommen in 1.367 Apps vor, Aufrufe verbotener Methoden nur in 62 Apps.

RQ3: Im Durchschnitt dauert die Analyse einer App 101 Sekunden, die Zeiten schwanken zwischen 10 Sekunden und 28,6 Minuten. Die Ursache dafür liegt in den Größen der Apps, da CogniCrypt$_{SAST}$ 83% der Analysezeit auf die Erstellung des Call-Graphs verwendet.

RQ4: CogniCrypt$_{SAST}$ entdeckt 20.426 Fehlbenutzungen über 23 Klassen verteilt in den 10.000 Android-Apps. Im Kontrast dazu findet das nachgebaute CryptoLint nur 5.507 in sechs Klassen. Insgesamt findet CogniCrypt$_{SAST}$ also fast vier mal so viele Fehler.

5 CogniCrypt$_{GEN}$

CogniCrypt$_{GEN}$ ist ein Code-Generator für Crypto APIs, der auf CrySL aufsetzt. CogniCrypt$_{GEN}$ generiert anwendungsfallspezifischen Code, der sicher und korrekt in Abhängigkeit der CogniCrypt$_{GEN}$ bereitgestellten CrySL-Regeln ist. Als weitere Eingabe neben den CrySL-Regeln erhält CogniCrypt$_{GEN}$ ein Code-Template, das sicherheitsunkritischen Wrapper-Code für den jeweiligen Anwendungsfall bereitstellt.

5.1 Ablauf & Prototypische Implementierung

```
34  byte[] salt = new byte[32];
35  char[] pwd = {'p', 'w', 'd'}; //Festverdrahtet zur Demonstration
36  javax.crypto.SecretKey encryptionKey = null;
37
38  CrySLCodeGenerator.getInstance().
39   includeClass("java.security.SecureRandom").addParameter(salt,"salt").
40   includeClass("java.security.PBEKeySpec").addParameter(pwd,"password").
41   includeClass("javax.crypto.SecretKeyFactory").
42   includeClass("java.security.SecretKey").
43   includeClass("javax.crypto.SecretKeySpec").addReturnObject(encryptionKey).
44   generate();
```

Abb. 3: CogniCrypt$_{GEN}$-Template, das korrekte Variante von Abbildung 1 generiert.

Um zu illustrieren, wie CogniCrypt$_{GEN}$ funktioniert, beziehe ich mich auf das Template in Abbildung 3, mit dessen Hilfe CogniCrypt$_{GEN}$ eine *sichere* Variante des `PBEKeySpec`-Beispiels in Abschnitt 1 generiert. Zeilen 34 – 36 erstellen drei Objekte und bilden den Wrapper-Teil des Templates. Im Anschluss folgt der Aufruf des eigentlichen Code-Generators durch eine Fluent API. Zwischen Instantiierung und Abschluss (Zeilen 38 und 44) erfolgt dessen Konfiguration über drei Methoden. Durch den Aufruf von `includeClass()` in Zeile 40 generiert CogniCrypt$_{GEN}$ Code für die Klasse `PBEKeySpec`. Der Aufruf von `addParameter()` gibt dann die Variable `pwd` im Wrapper-Code, also das nutzerspezifizierte Passwort, über die Variable `password` in der CrySL-Regel von `PBEKeySpec` an den generierten Code weiter. Die Rückgabe von Objekten vom generierten an den Wrapper-Code erfolgt über die Methode `addReturn()`, wie in Zeile 43 zu sehen. Hier wird der ge-

nerierte Schlüssel in der Code-Template-Variable `encryptionKey` gespeichert. Nach der Verarbeitung des Templates, generiert CogniCrypt$_{GEN}$ die `templateUsage()`-Methode.

Die Implementierung von CogniCrypt$_{GEN}$ setzt auf existierender Infrastruktur auf. So nutzt CogniCrypt$_{GEN}$ den CrySL-Compiler um die Regeln einzulesen. Die Lösung zur Verarbeitung der Templates und Generierung des Codes setzt auf dem Eclipse JDT auf.

5.2 Evaluierung

In der Evaluation von CogniCrypt$_{GEN}$ habe ich folgende Forschungsfragen betrachtet:

RQ5 Unterstützt CogniCrypt$_{GEN}$ oft genutzte kryptographische Anwendungsfälle?
RQ6 Produziert CogniCrypt$_{GEN}$ schnell genug Code um in der alltäglichen Softwareentwicklung eingesetzt werden zu können?
RQ7 Nehmen Entwickler von CogniCrypt$_{GEN}$ das Tool als nutzbarer wahr als einen vergleichbaren Code-Generator?

5.2.1 Setup

Zur Beantwortung von **RQ5** habe ich aus verschiedenen Quellen Anwendungsfälle zusammengetragen und diese mit CogniCrypt$_{GEN}$ implementiert. Um **RQ6** zu beantworten, habe ich anschließend alle implementierten Anwendungsfälle mit CogniCrypt$_{GEN}$ generiert und die Laufzeit gemessen. Für die Beantwortung von **RQ7** habe ich 16 Doktoranden und Studierende an meiner lokalen Universität gebeten Implementierungen mehrerer Anwendungsfälle in CogniCrypt$_{GEN}$ und einem ähnlichen Code-Generator, der auf XSL basiert, vorzunehmen und anschließend ihre Eindrücke zu schildern.

5.2.2 Ergebnisse

RQ5: Insgesamt habe ich elf Anwendungsfälle aus drei Quellen zusammengetragen. Die Anwendungsfälle reichen von passwort-basierter Verschlüsselung von Dateien und hybrider Verschlüsselung von Byte-Arrays bis zum Signieren von Strings. Ich habe alle elf Anwendungsfälle erfolgreich in CogniCrypt$_{GEN}$ implementiert.

RQ6: Die Laufzeit von CogniCrypt$_{GEN}$ schwankte für die elf Anwendungsfälle zwischen 6,6 und 8,1 Sekunden. CogniCrypt$_{GEN}$ ist also ausreichend performant.

RQ7: Alle 16 Teilnehmer schlossen die an sie gestellten Aufgaben erfolgreich ab. Wenn gebeten die Nutzbarkeit der beiden Code-Generatoren zu vergleichen, bewerteten Teilnehmer CogniCrypt$_{GEN}$ als signifikant nutzbarer als die XSL-basierte Lösung.

6 Nutzerstudie

Ich evaluiere schließlich die Effektivität von CogniCrypt und somit ob es das oben aufgestellte Ziel erfüllt. Für genauere Beobachtungen betrachte ich diese Forschungsfragen:

RQ8 Welchen Einfluss hat CogniCrypt auf die Funktionalität der Anwendungen?
RQ9 Welchen Einfluss hat CogniCrypt auf die Sicherheit der Anwendungen?
RQ10 Welchen Einfluss hat CogniCrypt auf die Entwicklungszeit der Anwendungen?
RQ11 Nehmen Entwickler CogniCrypt als nutzbarer wahr im Vergleich zu Eclipse?

6.1 Setup

Zur Beantwortung der vier Forschungsfragen führe ich eine Nutzerstudie an der prototypischen Eclipse-Implementierung durch. Dazu habe ich 24 Masterstudierende rekrutiert und diese zwei kryptographische Programme implementieren lassen, eins mit CogniCrypt, eins mit einem regulären Eclipse. Für **RQ8** und **RQ9** habe ich die Lösungen der Teilnehmer mit vorher definierten Sicherheits- und Funktionalitätskriterien abgeglichen. Um **RQ10** zu beantworten, habe ich die Zeiten gemessen, die Teilnehmer für ihre Lösungen brauchten. Für **RQ11** habe ich schließlich alle Teilnehmer zu ihren Eindrücken interviewt.

6.2 Ergebnisse

RQ8: Insgesamt waren die mit CogniCrypt erstellten Lösungen signifikant funktionaler. Von den 22 mit CogniCrypt implementierten Programme wurden 18 komplett richtig implementiert. Von den ohne CogniCrypt entwickelten hingegen waren nur drei korrekt.

RQ9: Die Sicherheitsanalyse liefert nochmal eindeutigere Ergebnisse. Ohne CogniCrypt implementierte nur ein Teilnehmer eine sichere Lösung. Mit CogniCrypt waren dagegen 18 von 22 Programme sicher. Bei einer der Aufgaben waren alle Programme mit CogniCrypt sicher, bei der anderen erhielten waren sie zu 87.5% sicher.

RQ10: Für eine der Aufgaben existieren signifikant weniger funktionale Lösungen ohne CogniCrypt als mit CogniCrypt. Daher vergleiche ich die Entwicklungszeit nur für die andere Aufgabe. Ohne CogniCrypt haben Teilnehmer im Schnitt zwischen 16 und 27,5 Minuten benötigt, mit CogniCrypt waren es nur zwischen neun und 20 Minuten. Insgesamt waren Teilnehmer mit CogniCrypt also signifikant schneller.

RQ11: Teilnehmer bewerten CogniCrypt generell positiv und signifikant besser als das reguläre Eclipse. Eclipse wurde als schlecht nutzbar bewertet. Einige Teilnehmer hatten aber auch Kritik an CogniCrypt, insbesondere in Bezug auf die Integration in Eclipse.

7 Schlussfolgerungen

In dieser Arbeit habe ich den integrierten Ansatz CogniCrypt vorgestellt. Wie ich gezeigt habe, reduziert CogniCrypt signifikant Fehlbenutzungen kryptographischer APIs

durch eine Mischung aus Code-Analyse und Code-Generierung. Diese ermöglicht es Entwicklern Crypto APIs zu nutzen, ohne sich in sie einarbeiten zu müssen.

Durch seinen modularen Aufbau mit der Spezifikationssprache CrySL als Basis für verschiedene Formen von Toolsupport ist CogniCrypt vielfach erweiterbar. Abseits dieser Arbeit wurden bereits das Program-Repair-Tool CogniCrypt$_{\text{FIX}}$, der Testfallgenerator CogniCrypt$_{\text{TEST}}$, der Dokumentationsgenerator CogniCrypt$_{\text{DOC}}$, ein auf OpenJ9-basierendes Hot-Patching-System und eine JCA-Erweiterung, die mit AspectJ Fehlbenutzungen automatisch umgeht, entwickelt. Das zeigt die Wirkmacht con CrySL und der hier vorgestellten Lösung. Andere Erweiterungen auf Basis von CrySL sind denkbar.

Literaturverzeichnis

[Bu17] Bundesamt fuer Sicherheit in der Informationstechnik (BSI): Cryptographic Mechanisms: Recommendations and Key Lengths. Bericht BSI TR-02102-1, BSI, Januar 2017.

[Eg13] Egele, Manuel; Brumley, David; Fratantonio, Yanick; Kruegel, Christopher: An empirical study of cryptographic misuse in android applications. In: ACM Conference on Computer and Communications Security. S. 73–84, 2013.

[Fe19] Feichtner, Johannes: A Comparative Study of Misapplied Crypto in Android and iOS Applications. In: Proceedings of the 16th International Joint Conference on e-Business and Telecommunications, ICETE 2019 - Volume 2: SECRYPT, Prague, Czech Republic, July 26-28, 2019. S. 96–108, 2019.

[Gu19] Gu, Zuxing; Wu, Jiecheng; Liu, Jiaxiang; Zhou, Min; Gu, Ming: An Empirical Study on API-Misuse Bugs in Open-Source C Programs. In: 43rd IEEE Annual Computer Software and Applications Conference, COMPSAC 2019, Milwaukee, WI, USA, July 15-19, 2019, Volume 1. S. 11–20, 2019.

[Kr20] Krüger, Stefan: CogniCrypt - The Secure Integration of Cryptographic Software. Dissertation, Universität Paderborn, Heinz Nixdorf Institut, Softwaretechnik, Oktober 2020.

[La14] Lazar, David; Chen, Haogang; Wang, Xi; Zeldovich, Nickolai: Why does cryptographic software fail?: a case study and open problems. In: ACM Asia-Pacific Workshop on Systems (APSys). S. 7:1–7:7, 2014.

[Or20] Oracle Inc.: , Java Cryptography Architecture (JCA), 2020. `https://docs.oracle.com/en/java/javase/15/security/java-cryptography-architecture-jca-reference-guide.html`.

[Ve17] VeraCode: , State of Software Security 2017. `https://info.veracode.com/report-state-of-software-security.html`, 2017.

Stefan Krüger hat von 2009 bis 2014 an der Otto-von-Guericke Universität Magdeburg Informatik studiert und mit dem Master of Science abgeschlossen. Von 2015 an hat Krüger zunächst an der TU Darmstadt, dann ab April 2016 an der Universität Paderborn zur sicheren Integration kryptographischer Software promoviert. Krügers Arbeit fand im Rahmen des Sonderforschungsbereich CROSSING statt. Dessen Ziel ist es Lösungen für zukunftssichere Kryptographie und IT-Sicherheit zu entwickeln. Inzwischen arbeitet Krüger für die CQSE GmbH.

Semantik, Sprache und Geometrie: Szenenverständnis lernen[1]

Iro Laina[2]

Abstract: Die Dissertation [La20] befasst sich mit grundlegenden Problemen auf dem Gebiet des Szenenverständnisses und bietet eine abgerundete Sicht auf das Thema. Szenenverständnis ist der Prozess der Wahrnehmung einer komplexen Umgebung durch sensorische Eingaben, der das Verständnis der Struktur, der darin liegenden Objekte und deren Interaktion untereinander und mit der Umgebung beinhaltet (aber nicht darauf beschränkt ist). In der Praxis umfasst das Szenenverständnis eine Vielzahl von Aufgaben, die darauf abzielen, detaillierte Informationen über die Szene zu extrahieren, und wurde in den letzten Jahren durch Deep-Learning-Techniken revolutioniert. Dieser Bericht bietet eine Zusammenfassung der Dissertation und ihrer Beiträge, die in zwei größere Kategorien — Wahrnehmung und Sprache — gegliedert sind. Maschinelle Wahrnehmung befasst sich speziell mit dem Verständnis von Geometrie und Semantik. In dieser Hinsicht bringt der erste Teil der Dissertation den Stand der Technik bei Problemen wie Tiefenschätzung, Lokalisierung, semantische Segmentierung und Szenenrekonstruktion voran. Der zweite Teil definiert die wichtige Rolle der natürlichen Sprache beim Verstehen von Szenen, durch die intelligente Systeme in der Lage sind, ihr Verständnis zu kommunizieren oder mit menschlichen Benutzern zu interagieren.

1 Einleitung

Eines der langjährigen Ziele der Künstlichen Intelligenz (KI) ist es, Maschinen zu konstruieren, die in der Lage sind, ihre Umgebung wahrzunehmen und autonom und rational zu handeln. Obwohl echte Intelligenz noch nicht erreicht ist, haben Fortschritte in den Bereichen Bildverarbeitung und maschinelles Lernen im letzten Jahrzehnt zu beispiellosen Erkenntnissen in verschiedenen Teilproblemen der KI beigetragen.

Um Intelligenz zu erreichen, sollten Maschinen in erster Linie mit der grundlegenden Fähigkeit ausgestattet sein, ihre Umgebung *wahrzunehmen*, was die Voraussetzung für komplexere Aufgaben ist, wie z. B. Navigation oder Interaktion mit dem Benutzer. Der Schwerpunkt dieser Dissertation liegt auf der Wahrnehmung der Umgebung von visuellen Eingaben, wodurch das Spektrum der visuellen Informationen, die automatisch aus einem Bild einer realen Szene extrahiert werden können, erweitert wird. In der Bildverarbeitung ist dies als *Szenenverständnis* (Scene Understanding) bekannt und umfasst eine Vielzahl von Aufgaben, die mit visueller, *geometrischer* oder *semantischer* Wahrnehmung zusammenhängen. Geometrische Wahrnehmung bezieht sich auf das Ableiten der Struktur und Anordnung der realen 3D-Welt, während semantische Wahrnehmung sich mit dem Erkennen von Entitäten, deren Attributen und Interaktionen beschäftigt.

[1] Englischer Titel der Dissertation: “Semantics, Language and Geometry: Learning to Understand the Scene”

[2] University of Oxford (Visual Geometry Group), iro.laina@eng.ox.ax.uk. Nominierung bei der Technischen Universität München.

Der erste Teil der Dissertation befasst sich mit solchen visuellen Verständnisaufgaben, sowohl in komplexen Szenen als auch in ausgewählten objektzentrierten Ansichten. Szeneverständnissysteme werden jedoch oft nicht nur ihr Verständnis an den Endbenutzer kommunizieren müssen, sondern müssen auch Feedback-Informationen berücksichtigen und ihre Leistung an die spezifischen Ziele des jeweiligen Benutzers anpassen. Um dies zu erreichen, bildet die Fähigkeit zur Verarbeitung von *natürlicher Sprache* eine weitere Achse im Szenenverständnis und damit auch den zweiten Teil dieser Dissertation.

Tabelle 1 fasst einige der grundlegenden Fähigkeiten zusammen, die ein visuelles System benötigt, um sich autonom zu verhalten. Wir assoziieren diese mit repräsentativen geometrischen, semantischen und linguistischen Aufgaben, die es zu lösen gilt; unsere Beiträge haben den Stand der Technik bei jedem dieser Probleme vorangetrieben. In den folgenden Abschnitten diskutieren wir diese im Detail.

Erforderliche Fähigkeiten eines autonomen Systems	**Repräsentative Aufgaben**	**Relevante Veröffentlichungen**
Seine Umgebung verstehen, sich bewegen und navigieren können	Tiefenabschätzung Semantische Segmentierung SLAM	[La16, Ta17, Dh19]
Objekte lokalisieren und erkennen, mit ihnen interagieren können	Semantische Segmentierung Lokalisierung Griffpositionserkennung	[La17, Ru17, Gh18]
Die Szene in der Sprache des Benutzers beschreiben können	Bildbeschriftung	[LRN19]
Benutzerinteraktion, bei Rückmeldungen falsches Verständnis korrigieren	Mensch-Maschine Interaktion	[Ru18]

Tab. 1: Übersicht über erforderliche Fähigkeiten für autonome visuelle Systeme und deren Aufschlüsselung in repräsentative Aufgaben, zu denen diese Dissertation Beiträge geleistet hat.

2 Geometrie und Semantik

Die Dissertation leistet einen Beitrag zum Stand der Technik bei verschiedenen Problemen des visuellen Verständnisses und entwickelt diesen weiter. Der erste Teil konzentriert sich speziell auf die maschinelle Wahrnehmung, die sich mit dem geometrischen und semantischen Verständnis von komplexen und objektzentrischen Bildern beschäftigt. Die behandelten Probleme, sind Tiefenschätzung, semantische Segmentierung und Lokalisierung aus Farbbildern; diese haben einen hohe praktischen Nutzen in realen Anwendungen und sind grundlegend für eine Höhere-Intelligenz. Aus technischer Sicht liegen die folgenden Beiträge in der Schnittmenge von Computer Vision und Deep Learning.

2.1 Komplexe Szenen

Bilder sind zweidimensionale Projektionen der realen, dreidimensionalen Welt. Die Entwicklung eines Computersystems, das eine reale 3D-Szene aus nur einer einzigen Ansicht wahrnehmen kann, ist ein bekanntes Problem in der Bildverarbeitung. Es gibt mehrere

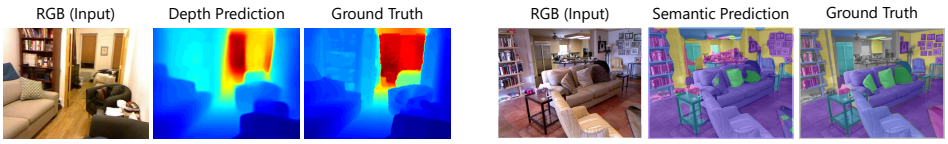
Aspekte, die zur visuellen Wahrnehmung beitragen, d. h. zum Verständnis dessen, was wir in einer Szene sehen. In Analogie zur visuellen Informationsverarbeitung im biologischen Gehirn (two-streams hypothesis [Go92]) beziehen sich diese Aspekte darauf, sowohl die räumliche Position (wo) als auch die semantische Bedeutung und die Eigenschaften (was) von Objekten in unserer Umgebung zu erfassen. Genauso sind beim maschinellen Sehen zwei der grundlegendsten Aufgaben Tiefenvorhersage und semantische Segmentierung.

Das Ziel der Tiefenvorhersage ist es, die Entfernungen von der Kamera zu Objekten in der Szene sowie deren Form zu schätzen. Dies ist eine wichtige Aufgabe in Szenarien, in denen Sensoren für direkte Tiefenmessungen nicht anwendbar oder nicht verfügbar sind, und als solches ist es ein gut erforschtes Thema in der Bildverarbeitung.

Seit 2012 hat der weit verbreitete Erfolg von Convolutional Neural Networks (CNNs), die Art und Weise, wie die Tiefenschätzung angegangen wird, verändert. Dies führte zu zahlreichen lernbasierten Methoden, die entweder vollständig überwacht oder selbstüberwacht sind und eine noch nie dagewesene Qualität bei der Schätzung von 3D-Strukturen aus einem einzigen Bild aufweisen. Der erste Beitrag dieser Dissertation, der in Kapitel 4 vorgestellt und in 3DV 2016 [La16] veröffentlicht wurde, gehört zu den ersten Deep-Learning-Methoden, die für die monokulare Tiefenschätzung vorgeschlagen wurden. Die allererste Deep-Learning-Lösung für dieses Problem stammt aus [EPF14], das im Wesentlichen die Vektorausgabe des Netzwerks in eine 2D-Vorhersage mit anschließender Verfeinerung umformt. In [La16] stellen wir jedoch fest, dass vollverknüpfte Schichten kein optimales Design für Aufgaben mit einer hochdimensionalen Ausgabe darstellen, wie z. B. Bildvorhersageaufgaben und folglich Tiefenschätzung. Wir schlagen daher eine Encoder-Decoder-Architektur vor, indem wir die Idee des Residual Learning [He16] auf "Up-Convolutions" erweitern, was zu einem "Fully Convolutional Residual Network" (FCRN) führt. Wir schlagen außerdem vor, das Netzwerk mit einer umgekehrten Huber Fehlerfunktion (Berhu) zu trainieren, der für diese Aufgabe besser geeignet ist als die gängigste L_2-Zielfunktion. Sowohl die Architektur als auch die Zielfunktion hatten einen erheblichen Einfluss auf das Gebiet der monokularen Tiefenschätzung und die Open-Source-Codebasis wurde von zahlreichen Gruppen verwendet.

In [Dh19] zerlegen wir anstelle der konventionellen Tiefenschätzung die Szene in Tiefenschichten; dies ermöglicht es uns, die Geometrie von Szenenregionen zu schätzen, die von Vordergrundobjekten verdeckt werden, was in Augmented-Reality-Anwendungen eine große praktische Bedeutung hat. Indem wir uns thematisch von geometrischen zu semantischen Aspekten des Szenenverständnisses bewegen, befassen wir uns auch mit der Aufgabe der semantischen Bildsegmentierung mit dem Fokus auf komplexe Innenraumumgebungen, wo unsere Methode Ergebnisse erzielt, die mit dem damaligen Stand der Technik konkurrieren können (Abbildung 1).

Unsere ersten Experimente führten zu der Beobachtung, dass die Qualität der geschätzten Tiefenkarten ausreicht, um aus einem einzigen Bild eine 3D-Szenenpunktwolke zu erstellen. Dies war ein ermutigendes erstes Zeichen für den Einsatz von gelernten Modellen zur Tiefenvorhersage bei der simultanen Positionsbestimmung und Kartierung (SLAM). SLAM ist ein grundlegendes Problem in der Computer Vision, das darauf abzielt, die Karte einer Umgebung zu schätzen, um so ein 3D-Modell der Szene zu erstellen und gleichzeitig



(a) Depth estimation **(b)** Semantic segmentation

Abb. 1: Vorhersagen des Fully Convolutional Residual Network (FCRN) bei den Aufgaben **(a)** Tiefenschätzung und **(b)** semantische Segmentierung.

die relative Position und Orientierung des sich bewegenden Agenten (oder der Kamera) zu bestimmen. Echtzeit-SLAM-Systeme finden sich in einer Vielzahl von Anwendungen, insbesondere für die Szenenrekonstruktion und Navigation von autonomen Fahrzeugen (z. B. selbstfahrende Autos), Haushaltsrobotern (z. B. Staubsauger) und Augmented Reality.

Unser Fokus liegt dabei insbesondere auf *visual* SLAM aus monokularen Videos. Traditionelle monokulare SLAM-Systeme basieren auf Punktkorrespondenzen oder Bildähnlichkeiten zwischen Einzelbildern. Ein repräsentatives Beispiel ist LSD-SLAM [ESC14], das direkt mit Bildintensitäten arbeitet. Die Fusion eines Lernansatzes mit einem solchen Framework wurde jedoch bisher nicht untersucht. In Kapitel 5 der Dissertation beschreiben wir ein Framework, das auf der Konferenz CVPR 2017 [Ta17] veröffentlicht wurde und die komplementäre Natur der beiden Ansätze (tief und traditionell) nutzt, um ihre jeweiligen Unzulänglichkeiten auszugleichen. Wir experimentieren speziell mit unserem zuvor vorgeschlagenen Netzwerk (FCRN) und LSD-SLAM.

Im Folgenden fassen wir die Vorteile durch die Hinzunahme der gelernten Komponente zusammen. (1) *Dichtere Rekonstruktion.* Da das Tiefenschätzungsnetzwerk dichte Ausgaben liefert, sind wir in der Lage, die Umgebung dicht zu rekonstruieren, auch ohne vorhandene Textur. Im Gegensatz dazu liefert traditionelles SLAM typischerweise Tiefenschätzungen entlang von Bildgradienten. (2) *Maßstabsschätzung in der realen Welt.* Beim traditionellen monokularen SLAM kann man die Tiefe nur relativ schätzen. Stattdessen kann die datengestützte Tiefenschätzung einen absoluten Maßstab für die 3D-Rekonstruktion liefern, der aus den Trainingsdaten gelernt wurde. Dies erhöht die Nutzbarkeit solcher Systeme in der Robotik und in Augmented-Reality-Anwendungen. (3) *Reine Rotationsbewegung.* Die monokulare Tiefenschätzung arbeitet mit einem einzigen Bild und ist somit nicht auf die Kamerabewegung angewiesen. Der kombinierte Ansatz ist robust gegenüber reinen Rotationsbewegungen oder langsamen Kamerabewegungen, da er sich nicht allein auf den Stereoabgleich verlässt. (4) *Semantik.* Geometrische Rekonstruktion kann mit semantischen Informationen ergänzt werden, die ebenfalls aus den Daten vorhergesagt werden (FCRN für semantische Segmentierung). Dies verleiht der 3D-Karte eine semantische Bedeutung und geht über Struktur und Aussehen hinaus, was in Anwendungen notwendig ist, bei denen intelligente Agenten innerhalb der Szene navigieren und agieren.

2.2 Objektzentrische Szenen

Objektzentrisches Verständniss ist ebenso wichtig wie das Verstehen von Szenen als Ganzes. Ein intelligentes System, das in der realen Welt eingesetzt wird, benötigt die Fähigkeit zu navigieren und seine Umgebung zu erkennen. Oft wird es aber auch erforderlich sein, Zielobjekte zu lokalisieren, zu isolieren und diese z. B. als interaktionsfähig oder nicht

interaktionsfähig zu erkennen oder Points of Interest und physikalische Randbedingungen für die Manipulation zu identifizieren. Solche Fähigkeiten sind in den schnell wachsenden Bereichen der autonomen Agenten und des robotischen Sehens unerlässlich.

In Kapitel 6 befassen wir uns mit objektzentrischen Problemen in spezifischen Anwendungen im Bereich des robotischen Sehens; unser Schwerpunkt ist die Lokalisierung. Die erste Anwendung ist die Lokalisierung und Segmentierung von gelenkigen chirurgischen Instrumenten in Bildern, die in MICCAI 2017 [La17] veröffentlicht wurde. Die Methode ist allgemein anwendbar und lässt sich tatsächlich von der menschlichen Posenschätzung in der Bilderverarbeitung inspirieren. Hier formen wir das Problem der Lokalisierung um, indem wir Ankerpunkte auf dem Zielobjekt als 2D-Heatmaps darstellen. Dies führt wiederum zu einer hochdimensionalen Ausgabe, die wir mit unserem zuvor vorgeschlagenen Netzwerk (FCRN) lernen können. Dieses Design erlaubt es auch, Lokalisierung und Segmentierung gleichzeitig mit einer einheitlichen Architektur zu lernen; wir zeigen, dass dieser Ansatz deutlich besser abschneidet als die direkte Regression kartesischer Koordinaten aus einem Bild. Die vorgeschlagene Methode erreichte auch den ersten Platz in der Segmentierung von laparoskopischen Instrumenten (EndoVis Challenge[3], 2017).

Als zweite Anwendung befassen wir uns mit der Lokalisierung von Greifpunkten für gängige Haushaltsgegenstände, der sogenannten Robotergreifpositionserkennung. Dies kann als Teil des visuellen Verarbeitungssystems eines Robotergreifers betrachtet werden. Diese Arbeit wurde in der ACCV 2018 Konferenz [Gh18] veröffentlicht und liefert Ergebnisse auf dem neuesten Stand der Technik und eine neue Perspektive auf dieses Problem. Inspiriert durch unsere frühere Forschung zur Vorhersage mehrerer Hypothesen [Ru17], besteht die Idee hier darin, mehrere realisierbare Greifalternativen vorhersagen zu lassen, wiederum in Anlehnung an die Heatmap-Darstellung. Dieser Ansatz liefert mehrere Optionen für den Roboterarm, die nach Vertrauenswürdigkeit geordnet werden.

3 Sprache

Die reale Welt ist multimodal; als Menschen nehmen wir sie nicht nur durch Bilder wahr, sondern z. B. auch durch Sprache, gesprochen oder geschrieben. Einige Probleme des visuellen Verständnis liegen in der Schnittmenge von Computer Vision und natürlicher Sprachverarbeitung. Um dies besser zu veranschaulichen, denken Sie an die folgenden Beispiele. Erstens: Viele Informationen, die im Web verfügbar sind, sind tatsächlich multimodal; visuelle Daten sind oft mit Beschreibungen (Alt-Text) verbunden, und Bilder können über Suchphrasen oder Schlüsselwörter abgerufen werden. Zweitens kann das Vorwissen aus der menschlichen Sprache genutzt werden, um visuelle Aufgaben zu leiten oder übergeordnete Argumentationsprobleme zu lösen. In der Tat kann das Wissen für eine Aufgabe oft von anderen Modalitäten übertragen werden, wie z.B. Sprache — ein Rezept zu kochen, nachdem man schriftliche Anweisungen verstanden hat.

Im zweiten Teil der Arbeit wird die Bedeutung von Sprache als Ausdrucksform für maschinelles Verstehen hervorgehoben, wobei insbesondere ihre Rolle für Kommunikation

[3] http://endovissub-instrument.grand-challenge.org

und Interaktion diskutiert wird. In intelligenten Agenten ebnet dies nun den Weg für übergeordnete, multidisziplinäre Fähigkeiten, die in vielen praktischen Szenarien mit Endbenutzern von entscheidender Bedeutung sind: Hilfe für sehbehinderte Benutzer, Benutzer-Maschine-Interaktion oder Erhöhung der Transparenz von Systemen durch Erklärungen.

3.1 Bildbeschreibung

Um die Kommunikation im Kontext des Szenenverständnis besser einordnen zu können, konzentrieren wir uns auf die Aufgabe, automatisch textuelle Beschreibungen von Bildern zu generieren, was auch als automatische Bildbeschreibung (Image Captioning) bezeichnet wird. Dies ist ein interessantes Problem, denn anstatt das visuelle Verständnis auf eine Repräsentation mit begrenztem Vokabular zu beschränken (wie bei der Bildklassifikation), erlaubt es die Welt frei zu bescrheiben. Die Erstellung von Bildunterschriften basiert nicht nur auf einem Bildmodell für das Verständnis der Szene, sondern auch auf einem Sprachmodell, das eine Szenendarstellung in Sätze umwandelt, die sowohl Grammatikregeln als auch semantische visuelle Entitäten berücksichtigen.

Bildbeschriftung hat zahlreiche Anwendungen. Sie ist ein entscheidender Schritt, um blinden Menschen den Zugang zu Computeranwendungen, dem Internet und sozialen Medien zu erleichtern oder ihnen zu helfen, ihre Umgebung besser zu verstehen. Es gibt bereits tragbare Geräte, die Blinden helfen sich auf der Straße zurechtzufinden und Menschen in ihrer Umgebung zu erkennen, indem sie visuelle Erkennung in Beschreibungen übersetzen. Ein weiteres Beispiel ist die Verwendung von natürlicher Sprache, um die Kommunikation zwischen dem Benutzer und dem System zu ermöglichen, was dazu beitragen kann, das Vertrauen und Engagement der Menschen zu gewinnen, wenn es um autonome Agenten wie selbstfahrende Autos geht. Schließlich können Bildbeschreibungen für die semantische Inhaltssuche in Bildern und Videos verwendet werden.

Deep Learning hat aufregende Fortschritte bei der Bildbeschriftung gemacht,aber es erfordert große Datensätze, die teuer und zeitaufwändig zu sammeln und zu beschriften sein können; dies wird normalerweise von menschlichen Annotatoren auf Plattformen wie Amazon Mechanical Turk (AMT) durchgeführt. Da die Annotatoren in der Regel pro Bild bezahlt werden, sind die Beschriftungen recht repetitiv und von geringem Aufwand, was zu Datensätzen führt, die einen manuellen Prozess der Bereinigung und Qualitätskontrolle benötigen. Trotz des Fortschritts ist das Mining einer großen Menge an Text, die bereits im Web verfügbar ist, für diese Aufgabe immer noch ein offenes Problem, und so werden die meisten Beschriftungsmodelle nur auf Basis begrenzter Datensätze entwickelt, die in der realen Welt schwer zu anzuwenden sind.

Kürzlich haben Forscher damit begonnen, über kuratierte Datensätze hinauszugehen und die Menge an Überwachung zu reduzieren, die sowohl für visuelle (z. B. Bildklassifizierung) als auch für linguistische (z. B. maschinelle Übersetzung) Aufgaben benötigt wird. Unüberwachtes Trainieren kann in der Tat von einer unbegrenzten Menge an unbeschrifteten Bildern (oder auch schwach beschrifteten Bildern, z. B. Tags) und großen Textkorpora (Bücher, Artikel, Beschreibungen) profitieren. In Kapitel 7 schlagen wir daher einen unüberwachten Bildbeschriftungsansatz vor und untersuchen, wie die beiden Modalitäten in Abwesenheit von Bild-Beschriftungspaaren abgeglichen werden können. Unser Ansatz

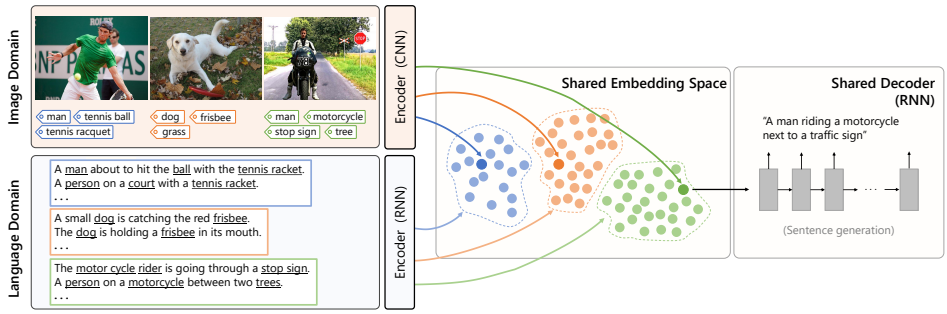


Abb. 2: Überblick über das Framework für unüberwachte Bildbeschriftungen. Bilder und Sätze stammen aus disjunkten Domänen und werden als *ungepaart* betrachtet.

wurde in ICCV 2019 [LRN19] veröffentlicht und hat den Stand der Technik bei der unpaarigen Bildbeschriftung vorangetrieben.

Das vorgeschlagene Trainingsschema besteht aus zwei Schritten (Abbildung 2). Der erste Schritt operiert nur auf der Sprachdomäne; ein rekurrentes "sequence-to-sequence"-Modell [SVL14] wird durch Rekonstruktion von Sätzen trainiert, um einen latenten Raum zu bilden. Die besondere Überlegung ist, dass der resultierende Einbettungsraum durch visuelles Vokabular strukturiert ist, was mit einem Triplettverlust erreicht wird. Wir zeigen, dass in diesem Raum Sätze, die ähnliche visuelle Inhalte beschreiben, ähnliche Repräsentationen haben. Anschließend betten wir Bildern in denselben Raum ein, wodurch die beiden Domänen (Bild- und Sprachdomäne) angeglichen werden. Um die gemeinsame Einbettung ohne Bild-Beschriftungs-Paare zu erreichen, ist eine Objekterkennung notwendig, um Pseudo-Korrespondenzen zwischen den beiden Domänen zu erzeugen. Wir zeigen jedoch, dass wir dank der visuellen Wortkookkurrenzen Bildbeschreibungen vorhersagen können, die weit über das feste Vokabular eines Objektdetektors hinausgehen. Da unser Ansatz auf disjunkte Domänen anwendbar ist, können wir auch über Bildunterschriften hinausgehen; wir zeigen, dass wir ein Modell trainieren können, das Fragen oder literarische Beschreibungen generiert wenn die Textdomäne durch z. B. Bücher ersetzt wird.

3.2 Interaktion

Bei der Aufgabe der Bildbeschreibung fließen visuelle Informationen in Form von Sprache vom System zum Benutzer. Als Nächstes befassen wir uns mit dem umgekehrten Problem: natürliche Sprache als Eingabe für ein visuelles System zu verwenden und so dem System gesprochenes oder geschriebenes menschliches Wissen zur Verfügung zu stellen.

Langfristig werden Agenten, die aus mehreren Teilsystemen für visuelles Verstehen bestehen, von denen jedes auf einem bestimmten Datensatz trainiert wurde, in realen Anwendungen eingesetzt werden und wahrscheinlich nicht perfekt in der Aufgabe sein, für die sie entwickelt wurden, z. B. aufgrund von Domänenverschiebungen. Es ist daher zu erwarten, dass die trainierten Modelle Fehler machen werden. In solchen Fällen, und insbesondere bei sicherheitskritischen Anwendungen, ist es unerlässlich, den Benutzer mit einzubeziehen, anstatt eine falsche Aktion durchzuführen. In der Praxis werden diese Modelle jedoch oft statisch innerhalb eines größeren Systems (Agenten) eingesetzt und als

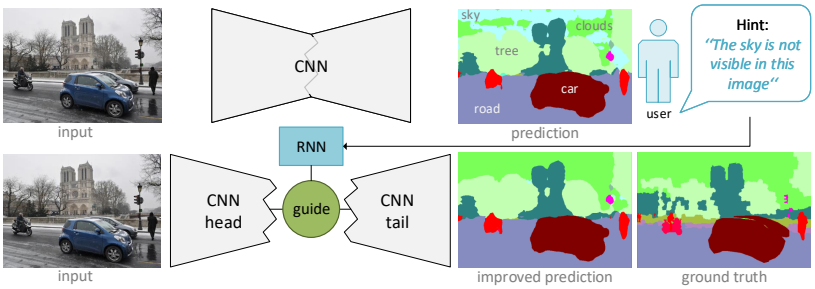


Abb. 3: Benutzer-Netzwerk-Interaktion für die semantische Bildsegmentierung. In diesem Beispiel hat das Modell den oberen Teil des Bildes fälschlicherweise als `sky` anstelle von `clouds` gekennzeichnet. Der Benutzer weist auf den Fehler hin und unser Modul korrigiert die Vorhersage.

Blackboxen behandelt. Benutzer, die keine Experten sind, haben keinen Zugang zum Innenleben des Systems. Das macht natürliche Schnittstellen für die Benutzerinteraktion zu einem wichtigen Bestandteil des visuellen Verständnisses; Dies ermöglicht es Nicht-Experten, mit komplexen Algorithmen zu interagieren, aber auch dem Agenten, Hilfe von seinem Benutzer zu erhalten, um sich an eine neue Umgebung anzupassen.

Motiviert durch dieses Szenario erweitern wir in Kapitel 8 tiefe Netzwerke für das visuelle Verstehen um eine Schnittstelle, die verbale (und auch nonverbale) Interaktion zwischen dem Benutzer und dem Netzwerk ermöglicht. Die Methode wurde in CVPR 2018 [Ru18] veröffentlicht. Das Hauptziel ist dabei die Bereitstellung von Hilfsinformationen für ein visuelles System. Die Hilfsinformationen (Hinweise) können aus verschiedenen Modalitäten stammen; wir experimentieren mit Benutzer-Klicks und Sätzen, die in natürlicher Sprache bereitgestellt werden. Ein Beispiel ist in Abbildung 3 dargestellt. Ein CNN wurde für die Aufgabe der semantischen Segmentierung vortrainiert. Der Benutzer betrachtet die Ausgabe und stellt Fehler fest; in diesem Beispiel ist der Himmel von Wolken bedeckt und der Benutzer möchte das zu seinem bevorzugten Ergebnis ändern, d. h., dass in der Vorhersage kein Himmel-Label erscheint.

Wir schlagen vor, dieses Problem durch räumlich-semantische Modulation der inneren Aktivierungen eines Netzwerks zu lösen. Wir bezeichnen dies als *Netzführung*. Wir zeigen, dass es möglich ist, die Ausgabe des Netzwerks bei Bedarf zu ändern, ohne seine Parameter zu verändern. Indem wir einen Menschen einbeziehen, um das Netzwerk zur Testzeit zu informieren, können wir eine Leistungssteigerung erreichen, ohne das Modell von seiner ursprünglichen Aufgabe abzulenken. Die Eingabe des Benutzers ist optional, so dass das Modell nicht darauf angewiesen ist und auch ohne den Benutzer normal arbeitet. Das Führungsmodul wird dann so trainiert, dass es die anfängliche Segmentierung immer entsprechend des gegebenen Hinweises verbessert. Obwohl das Modell nur mit einem simulierten Benutzer trainiert wurde, haben wir bei unseren Experimenten reale Benutzer einbezogen, die zur Testzeit mit dem Modell interagieren.

Es gibt mehrere andere Anwendungsfälle der vorgeschlagenen Idee, von denen einige in Folgearbeiten untersucht wurden. Zum Beispiel ist in risikoreichen Domänen wie der medizinischen Bildanalyse und Diagnose die Erfahrung von Medizinern ein Faktor, der berücksichtigt werden muss, vor allem wenn sie mit der Ausgabe des Systems nicht einver-

standen sind [Ja20]. Es muss ihnen ermöglicht werden, mit einem Algorithmus auf natürliche Art und Weise und ohne Informatikfachkenntnisse zu interagieren. Der Ansatz ist auch anwendbar, um das Netzwerk in Richtung des gewünschten Ergebnisses zu lenken, wenn mehrere Hypothesen erreichbar sind, zum Beispiel aufgrund von Mehrdeutigkeiten in den Daten [Ru17]. In einem anderen Beispiel kann die Möglichkeit der Interaktion von Benutzern mit dem Algorithmus mühsame Beschriftungsaufgaben beschleunigen, da sich die Annotatoren auf die Korrektur größerer Fehler konzentrieren können, anstatt detaillierte handgezeichnete Masken für jedes einzelne Objekt zu erstellen [UAF20].

4 Zusammenfassung und Gesellschaftlicher Einfluss

Bestehende KI-Systeme sind spezialisiert; sie werden programmiert, um bestimmte, genau definierte Aufgaben zu erledigen. KI-Systeme können jedoch auch modular sein, d. h. sie bestehen aus mehreren grundlegenden Blöcken, die sich jeweils auf ein bestimmtes Teilproblem konzentrieren, sodass sie für verschiedene Anwendungen nicht von Grund auf neu entwickelt werden müssen. In dieser Dissertation konzentrieren wir uns auf visuelle KI-Systeme und adressieren grundlegenden Probleme im Bereich des Szenenverständnisses, aufbauend auf den neuesten Durchbrüchen im tiefen Lernen.

Zunächst befassen wir uns mit der Extraktion geometrischer und semantischer Informationen eine Szene aus Bilddaten. Praktischen Anwendungen sind hier zahlreich und bedeutend. Zum Beispiel ist die Tiefenschätzung ein großer Schritt in Richtung kostengünstiger Geräte zur Unterstützung sehbehinderter Benutzer. Unser Modell wurde kürzlich von einem solchen System [Ba20] verwendet.

Als Nächstes wenden wir uns übergeordneten Problemen zu, die an der Schnittstelle von Sehen und Sprache liegen, um verbale Kommunikation und Interaktion mit dem visuellen System zu ermöglichen. Um die Herausforderungen der Zukunft anzugehen, die durch unüberwachtes oder lernendes Lernen angetrieben werden, verlassen sich unsere Algorithmen in diesem zweiten Teil nur auf deutlich reduzierte Lernüberwachung. Auch hier sind die Anwendungen in der realen Welt zahlreich: Technologien für Blinde, Chatbots, Verkörperter Agenten, Reduzierung des manuellen Beschriftungsaufwands und Software von der Fotobearbeitung bis zur medizinischen Bildanalyse. Zum Schluss noch ein Ausblick in die Zukunft: Die Entwicklung interaktionsfähiger KI-Systeme wird der Schlüssel zu Transparenz und Erklärbarkeit sein und entscheidend dazu beitragen, das Vertrauen der Gesellschaft in KI zu steigern.

Literaturverzeichnis

[Ba20] Bauer, Zuria; Dominguez, Alejandro; Cruz, Edmanuel; Gomez-Donoso, Francisco; Orts-Escolano, Sergio; Cazorla, Miguel: Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors. Pattern recognition letters, 137:27–36, 2020.

[Dh19] Dhamo, Helisa; Tateno, Keisuke; Laina, Iro; Navab, Nassir; Tombari, Federico: Peeking behind objects: Layered depth prediction from a single image. Pattern Recognition Letters, 125:333–340, 2019.

[EPF14] Eigen, David; Puhrsch, Christian; Fergus, Rob: Depth map prediction from a single image using a multi-scale deep network. Proc. NeurIPS, 27:2366–2374, 2014.

[ESC14] Engel, Jakob; Schöps, Thomas; Cremers, Daniel: LSD-SLAM: Large-scale direct monocular SLAM. In: Proc. ECCV. Springer, S. 834–849, 2014.

[Gh18] Ghazaei, Ghazal; Laina, Iro; Rupprecht, Christian; Tombari, Federico; Navab, Nassir; Nazarpour, Kianoush: Dealing with ambiguity in robotic grasping via multiple predictions. In: Proc. ACCV. Springer, S. 38–55, 2018.

[Go92] Goodale, Melvyn A; Milner, A David et al.: Separate visual pathways for perception and action. 1992.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proc. CVPR. S. 770–778, 2016.

[Ja20] Jacenków, Grzegorz; O'Neil, Alison Q; Mohr, Brian; Tsaftaris, Sotirios A: INSIDE: Steering Spatial Attention with Non-imaging Information in CNNs. In: Proc. MICCAI. Springer, S. 385–395, 2020.

[La16] Laina, Iro; Rupprecht, Christian; Belagiannis, Vasileios; Tombari, Federico; Navab, Nassir: Deeper depth prediction with fully convolutional residual networks. In: Proc. 3DV. IEEE, S. 239–248, 2016.

[La17] Laina, Iro; Rieke, Nicola; Rupprecht, Christian; Vizcaíno, Josué Page; Eslami, Abouzar; Tombari, Federico; Navab, Nassir: Concurrent segmentation and localization for tracking of surgical instruments. In: Proc. MICCAI. Springer, S. 664–672, 2017.

[La20] Laina, Iro: Semantics, Language and Geometry: Learning to Understand the Scene. Dissertation, Technische Universität München, 2020.

[LRN19] Laina, Iro; Rupprecht, Christian; Navab, Nassir: Towards Unsupervised Image Captioning with Shared Multimodal Embeddings. In: Proc. ICCV. S. 7414–7424, 2019.

[Ru17] Rupprecht, Christian; Laina, Iro; DiPietro, Robert; Baust, Maximilian; Tombari, Federico; Navab, Nassir; Hager, Gregory D: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proc. ICCV. S. 3591–3600, 2017.

[Ru18] Rupprecht, Christian; Laina, Iro; Navab, Nassir; Hager, Gregory D; Tombari, Federico: Guide me: Interacting with deep networks. In: Proc. CVPR. S. 8551–8561, 2018.

[SVL14] Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V: Sequence to sequence learning with neural networks. Proc. NeurIPS, 27:3104–3112, 2014.

[Ta17] Tateno, Keisuke; Tombari, Federico; Laina, Iro; Navab, Nassir: CNN-SLAM: Real-time dense monocular slam with learned depth prediction. In: CVPR. S. 6243–6252, 2017.

[UAF20] Uijlings, Jasper RR; Andriluka, Mykhaylo; Ferrari, Vittorio: Panoptic Image Annotation with a Collaborative Assistant. In: Proc. ACMM. S. 3302–3310, 2020.

Iro Laina schloss ihr Studium der Elektro- und Computertechnik an der Nationalen Technischen Universität Athen (NTUA) 2013 mit einem Diplom ab und erhielt 2016 ihren M.Sc. in Biomedical Computing an der Fakultät für Informatik der Technischen Universität München (TUM). Iro begann im April 2016 ihr Promotionsstudium unter der Betreuung von Prof. Dr. Nassir Navab (TUM). Seit Februar 2020 ist Iro Postdoctoral Research Assistant in der Visual Geometry Group an der University of Oxford und arbeitet mit Prof. Dr. Andrea Vedaldi.

Privatsphärensouverän Digital Interagieren: Ermöglichung von Privatsphärensouveränität für Endnutzende im digitalen Zeitalter[1]

Karola Marky[2]

Abstract: Durch die stetig fortschreitende Digitalisierung interagieren Menschen zunehmend mit Daten im digitalen Raum, die von (mobilen) Endgeräten generiert und auf (Cloud-)Servern gespeichert werden. Da diese Daten persönliche Informationen der Nutzenden enthalten können, spielt die Privatsphäre von Nutzenden eine zentrale Rolle in der Digitalisierung. Die vorliegende Dissertation untersucht verschiedene Unterstützungsmöglichkeiten für Nutzende bei der Formulierung, Umsetzung und Überprüfung von individuellen Privatsphäreentscheidungen im digitalen Raum, welche im Konzept *Privacy-Sovereign Interaction* zusammengefasst sind. Hierfür wurden die drei konkreten Untersuchungsgegenstände 1) Zwei-Faktor-Authentisierung, 2) Individuelle Verifizierbarkeit bei Internetwahlen und 3) Privatsphäre von Besuchern in IoT-Umgebungen in der Tiefe beleuchtet und in insgesamt fünfzehn empirischen Studien. Auf Basis der Studienergebnisse werden konkrete Anforderungen, Empfehlungen und Herausforderungen für Privacy-Sovereign Interaction im digitalen Raum abgeleitet.

1 Motivation und Einführung

Die stetig fortschreitende Digitalisierung resultiert in einer vermehrten Interaktion mit Daten im digitalen Raum. Aufgrund des möglichen Personenbezugs solcher Daten spielt individuelle Privatsphäre von Nutzenden eine zentrale Rolle in der Digitalisierung. Bereits durchgeführte Studien in der Literatur - bspw. [Em17] - zeigten bereits wiederholt auf, dass gängige Technologien häufig jedoch weder Fähigkeiten noch Wissen ihrer Nutzenden genügend berücksichtigen oder sogar keinerlei Funktionalität für individuellen Privatsphäreschutz anbieten. Als Konsequenz sind Nutzende häufig damit überfordert, individuelle Privatsphärepräferenzen durchzusetzen oder tun sich bereits schwer damit, Entscheidungen bezüglich ihrer Daten zu treffen. Hieraus resultiert eine Einschränkung der individuellen Souveränität von Nutzenden im digitalen Raum.

Die vorliegende Dissertation befasst sich mit Mechanismen und Prinzipien, die Nutzenden eine souveräne Interaktion mit ihren persönlichen Daten im digitalen Raum ermöglichen sollen. Die in der Dissertation vorgestellte Forschung ist in zwei grundlegenden Forschungsgebieten der Informatik angesiedelt: Einerseits Privatsphärenforschung als Teilgebiet der IT-Sicherheitsforschung und andererseits Gebrauchstauglichkeitsforschung als Teilgebiet der Human-Computer Interaction. Beide Gebiete werden im Forschungsgebiet

[1] Englischer Titel der Dissertation: "Privacy-Sovereign Interaction: Enabling Privacy-Sovereignty for End-Users in the Digital Era"

[2] Technische Universität Darmstadt, marky@tk.tu-darmstadt.de

der Usable Security und Privacy kombiniert. *Privacy-Sovereign Interaction* erweitert dieses Forschungsfeld, indem zusätzliche Aspekte neben der Gebrauchstauglichkeit in Tiefe untersucht werden. Die Ergebnisse der vorliegenden Dissertation umfassen insgesamt elf Paper. Alle Publikationen durchliefen dabei einen Peer-Review-Prozess, insgesamt vier Publikationen wurden bei der A*-gerankten[2] Konferenz *ACM CHI Conference on Human Factors in Computing Systems* vorgestellt.

1.1 Forschungsansatz

Privacy-Sovereign Interaction stellt Funktionalität für Nutzende bereit, die es diesen ermöglicht, mit persönlichen Daten gemäß individuellen Privatsphärepräferenzen zu interagieren. Dies umfasst ebenfalls den Schutz der Datenintegrität sowie digitalen Besitzes, der durch Daten repräsentiert ist, die alleinstehend Nutzenden nicht zugeordnet werden können - beispielsweise digitale Währungen. Den Nutzenden wird hierbei die höchste Vorherrschaft über ihre persönlichen Daten, die durch eine dritte Partei zu einem definierten Zeitpunkt generiert, gesammelt, gespeichert oder analysiert werden, ermöglicht. Diese Vorherrschaft ist lediglich durch andere Nutzenden sowie deren digitalen Besitz eingeschränkt. Die Interaktion ist so gestaltet, dass Nutzende sich lediglich minimale Kompetenzen aneignen müssen, welche zusätzlich durch sehr gute Gebrauchstauglichkeit und User Experience der entwickelten Technologien unterstützt werden.

Um das Gebiet der Privacy-Sovereign Interaction zu erforschen, wurde die oben vorgestellte Definition zunächst in drei Hauptteile aufgegliedert. Der erste Hauptteil betrifft die *User Agency*, welche Nutzenden eine zugreifbare Funktionalität bereitstellt, um Privatsphäre-bezogene Entscheidungen intentionsgemäß durchzusetzen. Die *User Competence* umfasst das Bewusstsein des Nutzenden über Privatsphärefunktionalitäten einer Technologie sowie die Kenntnisse über mögliche Konsequenzen der Nutzung und die korrekte Anwendung der bereitgestellten Funktionalität. Zur Teiloperationalisierung der User Competence wird die Gebrauchstauglichkeit gemäß ISO-Norm 9241-11 herangezogen. Diese beschreibt das Ausmaß, in dem eine Technologie verwendet werden kann, um spezifizierte Ziele effektiv, effizient und zufriedenstellend zu erreichen. Allerdings hat die Gebrauchstauglichkeit den engen Bezug zum konkreten aktiven Nutzungszeitraum einer Technologie. Die intensionsgemäßige Wirksamkeit des Handelns leitet sich jedoch ebenfalls aus Zusammenhängen ab, die diesen aktiven und konkreten Nutzungszeitraum verlassen. Diese Zusammenhänge werden durch die *User Experience*, welche in der ISO-Norm 9241-210 festgelegt ist, abgedeckt. Die User Experience betrachtet Effekte, die vor, während und nach der Nutzung einer Technologie entstehen. Dies ist von zentraler Wichtigkeit, da Vertrauen, *wahrgenommene* Effizienz und Verständlichkeit einer Technologie zu deren Akzeptanz beitragen.

[2] CORE Ranking: http://portal.core.edu.au/conf-ranks/

1.2 Untersuchungsgegenstände

Um die drei Bestandteile User Agency, User Competence, und User Experience im Rahmen von Privacy-Sovereign Interaction zu untersuchen, wurden für die Dissertation drei konkrete Untersuchungsgegenstände systematisch ausgewählt.

User Experience. Zur Untersuchung des User Experience Aspekts wurde der Untersuchungsgegenstand des *Schutzes* privater Daten und Besitzes gewählt. Aus Datenschutzsicht leitet sich dies wie folgt her: Persönliche Daten werden zunehmend auf den Servern von Dienstleistern gespeichert. Hieraus resultiert, dass Nutzende sich authentifizieren müssen, um auf ihre persönlichen Daten zuzugreifen. Ohne technischen Schutz durch IT-Sicherheitslösungen wäre es folglich nicht möglich, die Privatsphäre dieser Daten zu gewährleisten. Gleichzeitig steht die Stärke des technischen Schutzes häufig in Konflikt zur User Experience, da es aus Nutzendensicht von höchster Priorität ist, einen ausfallfreien und möglichst umfassenden Zugriff auf die persönlichen zu Daten haben. Ein aus IT-Sicherheitssicht starker Authentifizierungsmechanismus ist die *Zwei-Faktor-Authentisierung*. Diese wird vor allem im Kontext finanzieller Transaktionen eingesetzt und ist in vielen Ländern bereits verpflichtend für Online-Banking. Allerdings wurde bereits durch Studien nachgewiesen, dass Zwei-Faktor-Authentisierungslösungen deutliche Schwächen im Bereich der User Experience aufweisen, da Nutzende die Authentifizierungsvorgänge als zu zeitintensiv [We09] wahrnehmen und sich Personalisierung wünschen [WG17]. Die Dissertation untersucht daher die Forschungsfrage, wie Privacy-Sovereign Interaction zum Privatsphäreschutz für digitalen Besitz umgesetzt werden kann. Dies ist speziell auf den Untersuchsgegenstand der Zwei-Faktor-Authentisierung fokussiert.

User Competence. Im Bereich der User Competence existiert bereits eine Vielzahl an Arbeiten in der wissenschaftlichen Literatur, die sich mit den verschiedenen Umsetzungsmöglichkeiten für Privatsphäre-Einstellungsoberflächen befassen. Diese Einstellungsoberflächen sind bereits Teil alltäglicher Nutzung in vielen Bereichen. Weniger alltäglich, jedoch gleichzeitig von großer Bedeutung, sind Szenarien, in denen private Willensäußerungen digital erfasst werden. Internetwahlen stellen ein solches Szenario dar. Hierbei bezieht sich der Privatsphäreaspekt auf die Wahlentscheidungen der Wählenden, welcher in Demokratien durch das Wahlgeheimnis besonders geschützt ist. Neben der Wahrung des Wahlgeheimnisses ist Integrität eine große Herausforderung bei Internetwahlen, da fehlende Integrität im schlimmsten Falle den Ausgang einer Wahl beeinflussen kann. Aufgrund der Nutzung von privaten Geräten zur Stimmabgabe und des Internets können Wahlbehörden unmöglich sicherstellen, dass die Wahlinfrastruktur komplett frei von potentiellen Angreifern ist, die einen böswilligen Einfluss auf die Wahl nehmen wollen. Um diese Problematik zu adressieren, wurde das Konzept von Verifizierbarkeit erforscht.

In der vorliegenden Dissertation liegt der Fokus auf *Individueller Verifizierbarkeit*, die es Wählenden ermöglicht zu überprüfen, ob ihre Stimmen unverändert in der elektronischen Wahlurne registriert wurden. Die User Competence ist hierbei aus zwei primären Gründen von besonderer Wichtigkeit. Zunächst verursacht die individuelle Verifizierbarkeit eine Abweichung im Wahlprozess, welche in traditionellen Wahlkanälen - wie Wahllokalen oder bei Briefwahlen - nicht präsent ist. Des Weiteren müssen Wählende die Verifikati-

on eigenständig durchführen, da ihre Wahlentscheidung aufgrund des Wahlgeheimnisses nicht gegenüber Dritten preisgegeben werden darf. Darauf basierend untersucht die Dissertation die Forschungsfrage, wie Nutzende durch Privacy-Sovereign Interaction bei der Verifikation privater Daten unterstützt werden können. Dies ist speziell auf den Untersuchsgegenstand der individuellen Verifizierbarkeit fokussiert.

User Agency. User Agency schafft eine technologische Grundlage für Nutzende, um intentionsgemäß über *Privatsphäreaspekte* zu entscheiden und diese Entscheidungen zu realisieren. Viele Technologien verfügen bereits über technologische Grundlagen für Privatsphäreeinstellungen. Die Verbreitung von *Internet-of-Things (IoT) Geräten* führt jedoch zu neuen Privatsphäreherausforderungen, weil IoT-Geräte Daten jeder Person erfassen und verarbeiten können, die sich in ihrer Umgebung aufhält. Daraus resultiert, dass nicht nur die Privatsphäre direkter Besitzer dieser IoT-Geräte betroffen ist, sondern ebenfalls die jener Personen, die sich zeitweise in der Nähe von den IoT-Geräten aufhalten. Diese Personengruppe wird im Rahmen der Dissertation als Bystander bezeichnet.

Bisherige Forschungsarbeiten fokussieren primär auf die Besitzer der IoT-Geräte und jene, die diese als Angehörige desselben Haushaltes mitbenutzen. Privatsphäreeinstellungen sind häufig an ein Benutzerkonto gekoppelt und werden bei der Einrichtung des IoT-Gerätes gesetzt. Bystander haben jedoch keinen Zugriff auf solche Privatsphäreeinstellungen. Zur Erforschung des User-Agency-Aspektes fokussiert die Dissertation auf private IoT-Umgebungen und erarbeitet Grundlagen zur technologischen Umsetzung vom Privatsphäreschutz für Bystander. Hierauf basierend untersucht die vorliegende Dissertation die Forschungsfrage, wie Bystander in IoT-Umgebungen durch Privacy-Sovereign Interaction beim Schutz ihrer Privatsphäre unterstützt werden können.

2 Untersuchungsgegenstand: Zwei-Faktor-Authentisierung

Existierende Forschungsarbeiten zeigten, dass Nutzende Zwei-Faktor-Authentisierung effektiv nutzen können, jedoch existieren Effizienz- und Personalisierungsprobleme. Um diese Probleme umfassend zu untersuchen, wurde zunächst eine Interviewstudie mit 42 Versuchsteilnehmenden durchgeführt, um Anforderungen an die User Experience zu erarbeiten. Im zweiten Schritt wurden diese Anforderungen prototypisch in Form von personalisierbaren 3D-gedruckten Objekten umgesetzt und evaluiert.

Die Umsetzung durch personalisierbare 3D-gedruckte Objekte wurde bei der A*-gerankten Konferenz *ACM CHI Conference on Human Factors in Computing Systems* 2020 veröffentlicht.

2.1 Anforderungen an die User Experience

Interviewmethodik. Die Interviewstudie war in zwei Teile gegliedert. Im ersten Teil wurden die Versuchsteilnehmenden zu konkreten Erfahrungen mit Authentifizierungsverfahren befragt. Im zweiten Interviewteil wurden die Versuchsteilnehmenden hinsichtlich ihrer Erwartungen und Wünsche bezüglich der Interaktion mit Zwei-Faktor-Authentisierung

befragt. Hierbei wurde der Fokus daraufgelegt, dass die zwei Authentifizierungsfaktoren physisch voneinander getrennte Geräte sein müssen, um ausreichende Sicherheit zu garantieren.

Interviewresultate und Anforderungen. Die wichtigsten Kernergebnisse der Studie lassen sich wie folgt zusammenfassen: Versuchsteilnehmende möchten in der Nutzung von Zwei-Faktor-Authentisierung nicht an einen bestimmten Ort gebunden sein. Versuchsteilnehmende möchten nicht auf Laptops oder Personal Computers beschränkt sein und stattdessen primär *mobile Endgeräte* verwenden. Aus Sicherheitsgründen ist eine physische Trennung vom mobilen Endgerät und dem zweiten Authentifizierungsfaktor erforderlich, da ein Angreifer sonst lediglich Kontrolle über das Endgerät erlangen müsste, um den Nutzenden zu imitieren. Durch diese Trennung ist der zweite Authentifizierungsfaktor bei gängigen Lösungen zumeist auf Strom, Mobilfunknetz oder Netzwerk angewiesen. Viele Versuchsteilnehmenden berichteten, dass diese Notwendigkeit häufig in Erreichsbarkeitsproblemen resultierte. Daraus wird geschlussfolgert, dass der zweite Authentifizierungsfaktor im Idealfall *strom- und netzunabhängig* ist. Die nächste Ergebniskategorie umfasst Umsetzungsmöglichkeiten für den zweiten Authentifizierungsfaktor. Hierbei präferierten Versuchsteilnehmende *alltägliche Objekte*, da sie auf den Transport zusätzlicher Geräte verzichten möchten. Falls möglich, sollte der zweite Authentifizierungsfaktor in alltägliche Objekte wie Brillen, Accessoires oder Schlüsselanhänger *integrierbar* und *personalisierbar* sein. Auch die Interaktion mit diesen Objekten wurde von den Versuchsteilnehmenden berücksichtigt. Präferiert wurden *diskrete* Interaktionsmöglichkeiten, um den Authentifizierungsvorgang vor Anderen zu verschleiern oder *bedeckte* Interaktionen, beispielsweise in der Hosentasche.

2.2 Anforderungsumsetzung durch 3D-Auth

Resultierend aus den Interviewergebnissen wurden sieben konkrete Anforderungen an die Umsetzung von Zwei-Faktor-Authentisierung unter der Berücksichtigung der User Experience zusammengestellt. Im Diskussionsteil der Dissertation wird hergeleitet, wieso keine der gängigen Lösungen den gesamten Anforderungskatalog erfüllt. Die Hauptmängel gängiger Ansätze beziehen sich hierbei auf die Strom- und Netzunabhängigkeit sowie unzureichende Personalisierung.

3D-Auth. Zur Umsetzung des gesamten erarbeiteten Anforderungskatalogs, wurde das Konzept *3D-Auth* erarbeitet. 3D-Auth basiert auf individuell gestaltbaren 3D-gedruckten Objekten, die in Kombination mit einem mobilen Endgerät verwendet werden. Die Objekte sind aus zwei verschiedenen Materialien angefertigt. Eines ist ein simples nichtleitendes Material (z.B. PLA) und ein zweites leitfähiges, kapazitives Material, welches eine Erkennung der Objekte ohne Stromzufuhr durch handelsübliche Touchscreens basierend auf dem Prinzip der kapazitiven Kopplung ermöglicht. 3D-Auth-Objekte haben eine individuelle, interne, 3D-gedruckte Struktur aus dem kapazitiven Material, diese wird an der Objektunterseite durch Punkte aus dem kapazitiven Material von einem Touchscreen erkannt und repräsentiert so den Authentifizierungsfaktor ”Besitz”. Der zweite Authentifizierungsfaktor wird durch die Interaktion des Nutzenden mit dem Objekt umgesetzt

und repräsentiert den Authentifizierungsfaktor ”Wissen”. Durch die Interaktion werden weitere Punkte an der Objektunterseite in Touchpunkten aktiviert, woraus ein Authentifizierungsmuster resultiert. Basierend auf einer Touchscreeninteraktionsfläche von vier mal vier Zentimetern und dem Zugriff auf kapazitive Rohdaten, ergibt sich so ein Passwortraum von 33.554.431 möglichen Authentifizierungsmustern.

Entwicklung und Evaluation. Zur Entwicklung der Authentifizierungsinteraktionen wurden zwei aufeinander folgende Studien mit Experten durchgeführt. Der dadurch entstandene Interaktionsraum umfasst fünf Kategorien von atomaren Interaktionen. Pro Kategorie wurde ein 3D-Auth-Objekt prototypisch umgesetzt. Die Prototypen wurden in zwei Kontrollexperimenten mit 25 Nutzenden evaluiert. Insgesamt wurde 3D-Auth von den Versuchsteilnehmenden sehr positiv aufgefasst, was sich ebenfalls in den Messungen widerspiegelte. Die Versuchsteilnehmenden konnten insgesamt 80% der Interaktionen mit den Prototypen auf Anhieb korrekt ausführenund zeigten nach zehn Tagen eine korrekte Einprägung von 90% der Interaktionen.

3 Untersuchungsgegenstand: Individuelle Verifizierbarkeit

Der zweite Untersuchungsgegenstand im Fokus der vorliegenden Dissertation ist individuelle Verifizierbarkeit im Kontext von Internetwahlen. Bereits existierende Forschungsarbeiten zeigten, dass grundlegende Probleme von einzelnen individuell verifizierbaren Internetwahlprotokollen auf Seite der Nutzenden bestehen. Darunter wurde gezeigt, dass Nutzende Schwierigkeiten haben, den komplexen Protokollen erfolgreich zu folgen. Diese Komplexität schlägt sich ebenfalls in der Akzeptanz von Verifikationsprotokollen nieder, welche Wählende zum Teil als unnötigen Zusatz empfinden.

In diesem Teil der vorliegenden Dissertation werden zunächst individuell verifizierbare Internetwahlprotokolle durch eine systematische Literaturrecherche zusammengetragen und anschließend basierend auf Wählerinteraktionen kategorisiert, um eine spätere Evaluation mit 100 Wählenden zu ermöglichen. Zusätzlich werden verschiedene Umsetzungsmöglichkeiten für Code-Interaktionen mit 18 Wählenden als Nebenaspekt untersucht. In einem zweiten Schritt werden verschiedene Bedienoberflächen von sogenannten Code Sheet Systemen tiefergehend untersucht. Die Evaluation der Bedienoberflächen und die Untersuchung der Code-Interaktionen wurden bei der A*-gerankten Konferenz *ACM CHI Conference on Human Factors in Computing Systems* 2019 und 2020 vorgestellt.

3.1 Kategorisierung individuell verifizierbarer Systeme

Um eine Ausgangslage zu erarbeiten, wurde zunächst eine Literaturrecherche durchgeführt, um individuell verifizierbare Internetwahlprotokolle zusammen zu tragen. Die resultierenden Publikationen wurden durch Experten in fünf Kategorien auf Basis des Verifikationszeitpunkts und Wählendeninteraktionen eingeteilt. Die fünf Kategorien sind 1) Audit-or-Cast, 2) Tracking Data, 3) Verification Device, 4) Code Sheets, und 5) Delegation. Basie-

rend den Kategorien wurden vier Prototypen[3] individuell verifizierbarer Internetwahlsysteme für eine Nutzendenstudie umgesetzt, die den Kategorien entsprechen.

Evaluationsmethodik. Zur Evaluation der vier Kategorien wurde eine Studie mit 100 Versuchsteilnehmenden durchgeführt. Jede Versuchsperson wurde einer Kategorie zugelost und gebeten, zwei Stimmen für eine Bundestagswahl abzugeben. Innerhalb der Studie wurden die Wahloptionen der Versuchsteilnehmenden verfälscht, um einen Angriff zu simulieren. Die Effizienz wurde durch die Dauer des Vorgangs operationalisiert. Die subjektive Nutzendenzufriedenheit, Vertrauen und Verständlichkeit wurden durch Fragebögen erfasst.

Evaluationsresultate. Die Kernergebnisse dieser Evaluation lassen sich wie folgt zusammenfassen. In der Versuchsbedingung Audit-or-Cast wurden 28% der verfälschten Stimmen durch die Versuchsteilnehmenden detektiert. Bei Verification Devices lag diese Rate bei 64%, bei Tracking Data bei 84% und bei Code Sheets sogar bei 100%. Dies spiegelt sich ebenfalls in den Verständlichkeitsresultaten wieder. Generell zeigte die durchgeführte Studie, dass Audit-or-Cast lediglich in Expertenkreisen verwendet werden sollte, da das Konzept für Wählende ohne Fachwissen schwer nachvollziehbar und durchführbar ist. Code Sheets stellten die einzige Kategorie dar, in der eine Stimmabgabe ohne vorherige Verifikation nicht möglich ist. Dies könnte eine Erklärung für die Detektionsrate von 100% sein. Andererseits benötigten die Versuchsteilnehmenden bei der Interaktion mit Code Sheets signifikant länger. Auf Basis der Studienergebnisse wurden konkrete Empfehlungen für die Entwickler und Entscheidungstragende für Internetwahlen abgeleitet.

3.2 Verbesserung von Code Sheet Bedienoberflächen

Zur tiefergehenden Untersuchung von Code-Sheet-Bedienoberflächen wurden drei Studien durchgeführt. Zunächst wurde die Bedienoberfläche des Systems der Schweizer Post von zwölf Experten auf Schwachstellen untersucht. Im zweiten Schritt wurde eine verbesserte Bedienoberfläche erstellt und mit 36 Versuchsteilnehmenden in einer explorativen Studie evaluiert. Beide Studien dienten als Basis für die Erarbeitung eines Redesigns.

Das Redesign wurde zusammen mit der Originalbedienoberfläche in einer vergleichenden Studie mit 49 Versuchsteilnehmenden evaluiert. Es wurde gezeigt, dass die Versuchsteilnehmenden mit dem Redesign signifikant mehr verfälschte Stimmen detektieren, dem Redesign signifikant mehr vertrauen und es als verständlicher wahrnehmen. Allerdings dauerte die Interaktion mit dem Redesign verglichen dem Original länger. Neben diesen spezifischen Ergebnissen konnten zusätzliche Empfehlungen für weitere Verifikationsbedienoberflächen aus den Studienergebnissen abgeleitet werden.

[3] Die Kategorie Delegation wurde von der Evaluation ausgeschlossen, da hier der Wahlvorgang modifiziert ist, um die Delegation von individueller Verifizierbarkeit basierend auf mathematischen Beweisen zu ermöglichen.

4 Untersuchungsgegenstand: Privatsphäre In IoT-Umgebungen

Um die Privatsphäre von Bystandern in IoT-Umgebungen zu untersuchen, wurden zunächst drei explorative Studien durchgeführt. Die erste Studie diente zur Untersuchung allgemeiner Wahrnehmungen und Maßnahmen zum Schutz der Privatsphäre in Form von Bewältigungsstrategien. Die zweite Studie untersuchte mentale Modelle von Bystandern. Die letzte Studie untersuchte verschiedene Bekanntheitsgrade mit besuchten IoT-Umgebungen. Auf Basis dieser Studien wurden Herausforderungen für Privacy-Sovereign Interaction für Bystander abgeleitet. Als mögliche Lösung wurde ein Konzept zur Privatsphäreassistenz mit 1126 Versuchsteilnehmenden erforscht.

Die Resultate der ersten Studie wurden bei der Nordic Conference on Human-Computer Interaction (NordiCHI) 2020 veröffentlicht. Die dritte Studie wurde bei der B-gerankten International Conference on Mobile and Ubiquitous Multimedia (MUM) 2020 veröffentlicht. Die Untersuchung zur Privatsphäreassistenz befindet sich derzeit in Begutachtung. Als Nebenaspekt wurden Informationslevel untersucht, dies wurde bei der A*-gerankten Konferenz *ACM CHI Conference on Human Factors in Computing Systems* 2020 als Poster veröffentlicht.

4.1 Interviewstudien zu Smart Home Besuchen

Interviewstudie I. In der ersten Studie wurden 21 Teilnehmende über den Besuch eines Smart Homes befragt. Hierfür wurden den Versuchsteilnehmenden verschiedene auf dem Markt verfügbare Geräte präsentiert. Zunächst wurden die Teilnehmenden zu allgemeinen Wahrnehmungen eines solchen Besuches befragt. Es konnte gezeigt werden, dass die Teilnehmenden Privatsphäreaspekte als Nachteil für Besucher wahrnahmen. Im zweiten Teil ging es um konkrete Maßnahmen zum Privatsphäreschutz. Die drei Hauptergebnisse dieses Teils sind, dass Versuchsteilnehmende entweder die Auswirkungen auf ihre Privatsphäre fehleinschätzen, den Schutz der Privatsphäre aufgrund von Hilflosigkeit aufgaben oder hypothetische Maßnahmen, wie das Ausschalten eines Gerätes, erwägten.

Interviewstudie II. Um die Ursache der oben festgestellten Fehleinschätzungen zu erforschen, wurden die mentalen Modelle von fünfzehn Versuchsteilnehmenden untersucht. Diese wurden gebeten, den Datenfluss innerhalb eines besuchten Smart Homes zu skizzieren und dabei Geräte und Entitäten zu markieren, die Daten über den Versuchsteilnehmenden sammeln. Auf Basis der Ergebnisse wurde gezeigt, dass Fehleinschätzungen auf die Datensensitivität und Personengebundenheit bezogen sind, jedoch nicht zwingend mit der Technologiekenntnis des Versuchsteilnehmenden einhergehen.

Interviewstudie III. In der dritten Studie wurden 21 Versuchsteilnehmende zu verschiedenen IoT-Umgebungen befragt, die nach Bekanntheitsgrad variiert wurden. Die Ergebnisse zeigten, dass Hilflosigkeit im Kontext neuer, unbekannter Umgebungen eine große Rolle spielt.

Auf Basis der oben beschriebenen Studien wurde gezeigt, dass User Agency für Bystander bei gängigen Geräten nur schwer erreichbar ist. Versuchsteilnehmende wünschten sich

Möglichkeiten zum Privatsphäreschutz, priorisierten jedoch den jeweiligen Besuch in der entsprechenden IoT-Umgebung und die Interaktion mit dem Besitzer des Smart Homes.

4.2 Privatsphäreassistenz

Um Nutzende generell bei der Umsetzung ihrer persönlichen Privatsphärepräferenzen zu unterstützen, wurde in der Literatur das Konzept einer persönlichen Privatsphäreassistenz vorgeschlagen [Da18]. Im letzten Teil der vorliegenden Dissertation wird eine konkrete Umsetzungsmöglichkeit für Privatsphäreassistenz tiefergehend untersucht. Privatsphäreassistenten können basierend auf unterschiedlichen Konzepten Entscheidungen für Nutzende treffen: 1) Privatsphäreeinstellungen, die Nutzende vorher festlegen, 2) Privatsphäreprofile, die Nutzende repräsentieren, oder 3) Machine-Learning- Ansätze. Auf Basis der Literatur wurden Privatsphäreprofile ausgewählt, da diese einen guten Mittelweg zwischen Automatisierung und Kontrolle ermöglichen.

Auf Basis einer zuvor entwickelten Profilzuordnungsmethode wurden 1126 Versuchsteilnehmenden in einer Onlinestudie eines aus acht Privatsphäreprofilen zugeteilt. Anschließend wurden die Versuchsteilnehmenden mit 18 Szenarien konfrontiert, die zuvor durch Experten ausgewählt wurden. Pro Szenario wählten die Versuchsteilnehmenden einen aus drei möglichen Privatsphäreassistenten: 1) Benachrichtigung über IoT-Geräte, 2) Benachrichtigung über IoT-Geräte und Handlungsempfehlung, und 3) autonomer Assistent. Zusätzlich wählten die Versuchsteilnehmenden, ob sie die Datenerfassung im Szenario erlauben. Auf Basis dieser Studie konnte gezeigt werden, dass Privatsphäreprofile als Basis für Privatsphäreassistenz dienen können. Der dominanteste Aspekt, der eine Auswirkung auf die Entscheidung der Teilnehmenden hatte, war die Anzahl der Entscheidungen, die innerhalb von 24 Stunden getroffen werden müssen.

5 Integration der Ergebnisse und Gesellschaftliche Bedeutung

Personen werden in der digitalen Welt durch persönliche Daten repräsentiert. Solche Daten können auch in Form von digitalem Besitz vorliegen, wie beispielsweise digitale Währungen oder Daten über Eigentum in der analogen Welt. Um in beiden Welten - der analogen und der digitalen - souverän handeln zu können, benötigen Individuen die Vorherrschaft über persönliche Daten und digitalen Besitz. Im Gegensatz hierzu steht, dass das Internet als Infrastruktur ursprünglich nicht für den Schutz der Privatsphäre konzipiert wurde und folglich auch nicht für Privacy-Sovereign Interaction, da das ursprüngliche Ziel des Internets in der Vernetzung von Forschungseinrichtungen bestand. Heutzutage, im digitalen Zeitalter, hat sich der Internetzugang auf Privatpersonen und sogar auf physische Objekte (IoT-Geräte) ausgeweitet. Infolgedessen spielen Aspekte des individuellen Datenschutzes und folglich ebenfalls Privacy-Sovereign Interaction eine wichtige Rolle in der Digitalisierung.

Die Wichtigkeit von Privacy-Sovereign Interaction wird durch das Prinzip der informationellen Selbstbestimmung, welches in §8 der EU-Grundrechtecharta festgeschrieben ist,

gestützt. Basierend darauf schafft die allgemeine Datenschutzgrundverordnung (DSGVO) wichtige rechtliche Grundlagen für Datenschutz in Europa und unterstreicht ebenfalls dessen Wichtigkeit. Ein Blick auf aktuelle Umsetzungen der durch die DSGVO vorgeschriebenen Aspekte zeigt jedoch, dass der Fokus von Dienstleistungsanbietern vor allem auf der User Agency liegt und häufig auf diese beschränkt ist. Die zwei anderen Bausteine User Competence und User Experience werden hierbei häufig derart vernachlässigt, dass Nutzende in der Flut an Datenschutzbestimmungen aufgeben und ohne Lesen in diese einwilligen. Dies schmälert die informationelle Selbstbestimmung und folglich auch die Privatsphäre von Nutzenden fundamental. Die vorliegende Dissertation setzt an diesem Punkt mit neuartigen Lösungen an, die bisherige Konzepte wie Zwei-Faktor-Authentisierung, individuelle Verifizierbarkeit und Privatsphäreeinstellungen neu überdenken und dadurch Nutzenden individuelle Souveränität in Bezug auf ihre persönlichen Daten zurückgeben.

Literaturverzeichnis

[Da18] Das, Anupam; Degeling, Martin; Smullen, Daniel; Sadeh, Norman: Personalized Privacy Assistants for the Internet of Things: Providing Users with Notice and Choice. IEEE Pervasive Computing, 17(3):35–46, Jul 2018.

[Em17] Emami-Naeini, Pardis; Bhagavatula, Sruti; Habib, Hana; Degeling, Martin; Bauer, Lujo; Cranor, Lorrie; Sadeh, Norman: Privacy Expectations and Preferences in an IoT World. In: Proceedings of the Symposium on Usable Privacy and Security. USENIX Association, Berkeley, CA, USA, S. 399–412, 2017.

[Ma20] Marky, Karola: Privacy-Sovereign Interaction – Enabling Privacy-Sovereignty for End-Users in the Digital Era. Dissertation, Technische Universität Darmstadt, Darmstadt, Germany, 2020.

[We09] Weir, Catherine S; Douglas, Gary; Carruthers, Martin; Jack, Mervyn: User perceptions of security, convenience and usability for ebanking authentication tokens. computers & security, 28(1-2):47–62, 2009.

[WG17] Weidman, Jake; Grossklags, Jens: I like it, but i hate it: Employee perceptions towards an institutional transition to byod second-factor authentication. In: Proceedings of the Annual Computer Security Applications Conference. S. 212–224, 2017.

Karola Marky wurde 1988 in Kaiserslautern geboren. Ihr Studium der Angewandten Informatik schloss sie 2012 an der Technischen Universität Kaiserslautern ab. Nach Absolvierung des DAAD-Stipendienprogramms *Sprache und Praxis in Japan*, welches das Erlernen der japanischen Sprache und zwei Forschungspraktika bei der Japan Aerospace Exploration Agency und dem National Institute for Informatics beinhaltete, begann sie 2015 ein Masterstudium der Informatik an der Technischen Universität Darmstadt. Im Juli 2017 startete Frau Marky ihre Promotion, welche sie im Dezember 2020 bei Prof. Dr. Max Mühlhäuser im Fachgebiet Telekooperation der Technischen Universität Darmstadt abschloss. Während der Promotionszeit verbrachte sie sechs Monate bei der Forschungsgruppe GEIST an der Keio Universität in Japan.

Stabilität und Expressivität von tiefen generativen Modellen[1]

Lars Mescheder[2]

Abstract: In den letzten Jahren haben generative Modelle das maschinelle Lernen revolutioniert. Im Gegensatz zu rein diskriminativen Modellen können generative Modelle mit Unsicherheiten umgehen und leistungsfähige Modelle lernen, auch wenn keine annotierten Trainingsdaten verfügbar sind. Hierbei beeinträchtigen jedoch zwei Aspekte ihre Expressivität: (i) Einige der erfolgreichsten Ansätze werden nicht mehr mit Hilfe von Optimierungsalgorithmen trainiert, sondern mit Algorithmen, deren Dynamik bisher nicht gut verstanden wurde. (ii) Generative Modelle sind oft durch den Speicherbedarf der Ausgaberepräsentation begrenzt. In dieser Arbeit stellen wir Lösungen für beide Problemstellungen vor: Im ersten Teil der Arbeit stellen wir eine Konvergenztheorie und neuartige Regularisierer für Generative Adversarial Networks vor, die es uns erlauben die Stabilität des Trainings zu verbessern. Im zweiten Teil dieser Arbeit stellen wir neue implizite Ausgaberepräsentationen für generative und diskriminative Modelle in 3D vor. Durch diesen impliziten Ansatz können wir viele Techniken, die in 2D funktionieren, auf den 3D-Bereich ausdehnen ohne ihre Expressivität einzuschränken.

1 Einleitung

Was bedeutet es für einen Algorithmus, einen Datensatz zu verstehen? Aus der Perspektive des überwachten Lernens wird diese Frage durch die Anforderung beantwortet, dass der Algorithmus auf ungesehene Testfälle anwendbar sein muss. Während dieses Maß an Verständnis für die meisten praktischen Anwendungen des maschinellen Lernens ausreicht, ist es jedoch sowohl aus wissenschaftlicher als auch aus philosophischer Sicht unbefriedigend: Die Fähigkeit, eine sich wiederholende Aufgabe zu lösen, erfordert nicht unbedingt ein tiefes Verständnis über die Daten und der Algorithmus kann jede beliebige Abkürzung nutzen, welche die Aufgabe leichter lösbar macht. Zum Beispiel kann ein maschineller Lernalgorithmus, der Katzen und Hunde unterscheiden kann, einfach die Form des Tieres ignorieren und sich stattdessen nur auf Details wie z. B. die Textur des Fells konzentrieren [Ge19].

Ein vielversprechender alternativer Ansatz ist die Anforderung, dass ein maschinelles Lernmodell in der Lage sein sollte, neue Daten zu generieren. Im Gegensatz zum überwachten Lernen erlaubt diese Aufgabe dem Modell nicht, Abkürzungen zu nehmen, da es jede in der Datenverteilung vorhandene Beziehung modellieren muss. Daher muss ein generatives Modell ein viel umfassenderes Verständnis über die Datenverteilung entwickeln.

Generative Adversarial Networks (GANs) [Go14] stellen einen der vielversprechendsten Ansätze zur generativen Modellierung dar. GANs formulieren das generative Lernproblem

[1] Englischer Titel der Dissertation: *Stability and Expressiveness of Deep Generative Models* [Me20]

[2] Autonomous Vision Group, Universität Tübingen, lmescheder@gmail.com

als ein glattes Zwei-Spieler-Spiel um, was zu vielen theoretischen und praktischen Optimierungsproblemen führt. Im ersten Teil dieser Arbeit schlagen wir einen theoretischen Ansatz zum Verständnis der lokalen Konvergenz von GANs vor. Ausgehend von einem sehr einfachen Beispiel des GAN-Trainings, das analytisch verstanden werden kann, leiten wir Konvergenzkriterien her, indem wir die Jacobi-Matrix des Gradientenvektorfeldes analysieren. Unsere Analyse führt auch zu neuen Regularisierern für das GAN-Training. Experimentell stellen wir fest, dass unsere neuen Regularisierer sehr effektiv sind und sie uns - zum ersten Mal - ermöglichen, ein GAN mit einer Bildauflösung von 1024×1024 Pixeln ohne progressiv-wachsende Architekturen [Ka18] zu trainieren. Konsequenterweise werden unsere Regularisierer heutzutage häufig für das Training von GANs verwendet, einschließlich Style-GAN [KLA19, Ka20], das weithin als der State-of-the-Art in der generativen Modellierung anerkannt ist.

Ein weiteres wichtiges Problem von generativen Modellen, aber auch von vielen diskriminativen Modellen, ist die Dimensionalität des Ausgaberaums. Während generative Modelle in letzter Zeit bemerkenswerte Erfolge bei der Erzeugung realistischer hochauflösender Bilder erzielt haben, konnte dieser Erfolg im 3D-Bereich noch nicht repliziert werden. Einer der Hauptgründe dafür ist, dass bestehende Repräsentationen für 3D-Geometrie wie Voxel, Punktwolken und Polygonnetze entweder speicherineffizient sind, unter Diskretisierungsartefakten leiden oder nicht effizient aus Daten abgeleitet werden können. Im zweiten Teil dieser Arbeit schlagen wir einen neuartigen Ansatz für das 3D tiefe Lernen vor, der auf der direkten Vorhersage der kontinuierlichen 3D Okkupierungsfunktion eines Objekts basiert. Diese 3D-Repräsentation reduziert den Speicherbedarf während des Trainings drastisch, erfordert keine Diskretisierung und kann effizient aus Daten abgeleitet werden.

2 Die Trainingsdynamik von GANs

2.1 Hintergrund

Beim Lernen eines generativen Modells möchten wir ein neuronales Netz $G_\theta : \mathcal{Z} \to \mathcal{X}$ mit Parametervektor θ trainieren, das einen latenten Vektor $z \in \mathcal{Z}$ als Eingabe nimmt und ein Element aus einem hochdimensionalen Raum $\mathcal{X}$ (z.B. ein Bild) ausgibt. Für ein solches generatives Modell $G_\theta(\cdot)$, können wir $z \in \mathcal{Z}$ aus einer A-Priori-Verteilung p_0 (z.B. einer Normalverteilung) ziehen, um eine Stichprobe $G_\theta(z)$ zu erhalten. Daraus ergibt sich eine Wahrscheinlichkeitsverteilung p_θ auf $\mathcal{X}$, die wir die *Generatorverteilung* nennen. Nehmen wir an, dass wir in der Lage sind, unabhängige Stichproben aus einer empirischen *Datenverteilung* $p_\mathcal{D}$ (z.B. eine Bildverteilung) zu ziehen. Unser Ziel ist es, den Parametervektor θ von $G_\theta(\cdot)$ so anzupassen, dass $p_\theta \approx p_\mathcal{D}$.

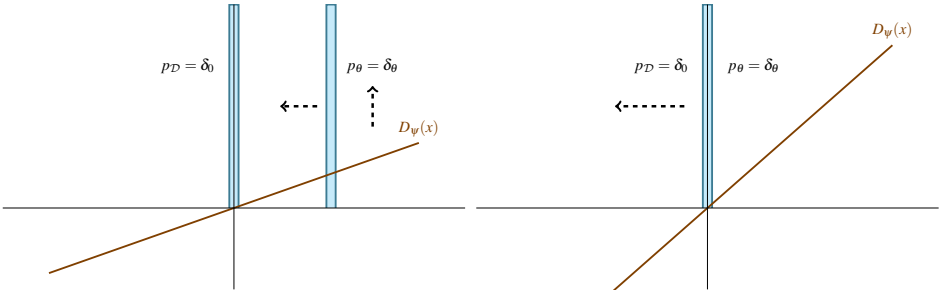
Dazu benötigen wir ein Abstandsmaß $\mathbb{D}(\cdot,\cdot)$ zwischen Wahrscheinlichkeitsverteilungen. Wir nennen ein solches Abstandsmaß eine *Divergenz* zwischen Wahrscheinlichkeitsverteilungen. Viele Divergenzen, die in der Praxis verwendet werden, können wir in der folgenden Form darstellen [Go14, NCT16, ACB17]:

$$\mathbb{D}(p_1 \,\|\, p_2) = \max_{D \in \mathcal{F}} \mathrm{E}_{x \sim p_1}\left[\varphi_1(D(x))\right] + \mathrm{E}_{x \sim p_2}\left[\varphi_2(-D(x))\right] \tag{1}$$

Hierbei bezeichnet $\mathcal{F} \subseteq \mathcal{X} \to \mathbb{R}$ eine Funktionsklasse und $\varphi_1, \varphi_2 : \mathbb{R} \to \mathbb{R}$ zwei Funktionen. Dabei diskriminiert die Funktion $D \in \mathcal{F}$ zwischen Stichproben aus p_1 und solchen aus p_2 und wird deshalb *Diskriminator* genannt. Der entscheidende Trick bei Generative Adversarial Networks ist nun, sowohl den Diskriminator $D(\cdot)$ als auch den Generator $G(\cdot)$ mit neuronalen Netzen zu parametrisieren. Dies führt zu einem hochdimensionalen, nichtkonvex-konkaven Min-Max-Problem, welches man in der Praxis meist durch simultanen oder alternierenden Gradientenabstieg zu lösen versucht [Go14].

Diese Trainingsprozedur ist jedoch ersteinmal nur heuristisch motiviert und es ist nicht klar, ob sie einen konvergenten Algorithmus liefert. In dieser Arbeit stellen wir eine Konvergenztheorie vor, die es erlaubt die Stabilitätseigenschaften des GAN-Trainings besser zu verstehen und neue Regularisierer herzuleiten.

2.2 Das Dirac-GAN



(a) $t = t_0$ (b) $t = t_1$

Abbildung 1: **Dirac-GAN.** (a) Zu Beginn schiebt der Diskriminator den Generator in Richtung der wahren Datenverteilung. (b) Wenn der Generator die Zielverteilung erreicht, ist die Steigung des Diskriminators am größten und drängt den Generator von der Zielverteilung wieder weg.

In diesem Abschnitt stellen wir ein minimales eindimensionales Beispiel für das GAN-Training vor, das wir *Dirac-GAN* nennen.

Um ein GAN vollständig zu spezifizieren, müssen wir die Datenverteilung $p_{\mathcal{D}}$, die durch den Generator erzeugte Verteilung p_θ und den Diskriminator D_ψ angeben. Die einfachste Datenverteilung $p_{\mathcal{D}}$ ist wohl durch eine einzige Zahl in $\mathbb{R}$ gegeben. Um die Diskussion noch weiter zu vereinfachen, setzen wir diese Zahl auf Null: $p_{\mathcal{D}} = \delta_0$. Der Generator erzeugt auch nur eine Zahl, die wir mit $\theta \in \mathbb{R}$ bezeichnen. Um zwischen zwei Zahlen zu unterscheiden, genügt ein linearer Diskriminator: $D_\psi(x) = \psi \cdot x$ mit $\psi \in \mathbb{R}$.

Interessanterweise konvergiert das GAN-Training nicht einmal lokal in diesem einfachen Szenario.

Lemma 1 *Beim Dirac-GAN hat die Jacobi-Matrix des Gradientenvektorfeldes am Gleichgewichtspunkt die beiden Eigenwerte $\pm\varphi_1'(0)\,i$, die beide auf der imaginären Achse liegen.*

Simultaner und alternierender Gradientabstieg ist daher nicht lokal konvergent am Gleichgewichtspunkt.

Unser einfaches Beispiel zeigt, dass naive gradientenbasierte GAN-Optimierung nicht immer zum Gleichgewichtspunkt konvergiert. Um die Instabilitäten zu verstehen, lohnt es sich das oszillatorische Verhalten genauer zu analysieren. Eine intuitive Erklärung für die Oszillationen ist in Abb. 1 zu sehen: Wenn der Generator weit von der wahren Datenverteilung entfernt ist, schiebt der Diskriminator den Generator in Richtung der wahren Datenverteilung. Gleichzeitig wird der Diskriminator sicherer, was die Steigung des Diskriminators erhöht (Abb. 1a). Wenn der Generator nun die Zielverteilung erreicht (Abb. 1b), ist die Steigung des Diskriminators am größten, wodurch der Generator von der Zielverteilung weggeschoben wird. Infolgedessen entfernt sich der Generator wieder von der wahren Datenverteilung und der Diskriminator muss seine Steigung von positiv zu negativ ändern. Nach einer Weile sind wir in einer ähnlichen Situation wie zu Beginn des Trainings, nur auf der anderen Seite der Datenverteilung. Dieser Prozess wiederholt sich unendlich oft und konvergiert nicht.

Das Phänomen kann auch bei komplexeren Beispielen auftreten: Solange die Datenverteilung auf einer niedrigdimensionale Mannigfaltigkeit konzentriert ist und die Klasse der Diskriminatoren groß genug ist, gibt es keinen Anreiz für den Diskriminator, Nullgradienten orthogonal zum Tangentenraum der Datenmannigfaltigkeit zu erzeugen und somit zum Gleichgewichtsdiskriminator zu konvergieren.

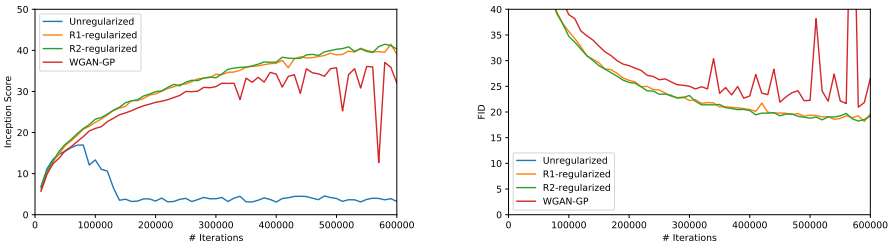
2.3 Regularisierung

Unsere Analyse des Dirac-GANs legt nahe, dass wir GAN-Training dadurch stabilisieren können, indem wir den Diskriminator für das Abweichen vom Gleichgewichtspunkt bestrafen. Der einfachste Weg, dies zu erreichen, ist den Gradienten zu regularisieren: Wenn die Generatorverteilung die wahre Datenverteilung erzeugt und der Diskriminator gleich Null auf der Datenverteilung ist, stellt der Regularisierer sicher, dass der Diskriminator keinen Gradienten orthogonal zur Datenverteilung erzeugen kann, ohne einen Verlust im GAN-Spiel zu erleiden. Dies führt zu dem folgenden Regularisierungsterm mit Parameter $\gamma > 0$:

$$R_1(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla D_\psi(x)\|^2 \right] \tag{2}$$

Im Hauptdokument [Me20] betrachten wir auch andere Varianten dieses Regularisierers, z.B. einen, der Diskriminatorgradienten auf der Generatorverteilung anstatt der wahren Datenverteilung bestraft (R_2-Regularisierung).

Wir sind nun bereit, unser Hauptkonvergenzergebnis für die regularisierte GAN-Trainingsdynamik zur formulieren.



(a) Inception Score (b) FID

Abbildung 2: **Quantitative Ergebnisse ImageNet**. Wir messen den Inception-Score [Sa16] und FID [He17] zwischen Generator- und Datenverteilung über die Anzahl der Iterationen.

Theorem 2 *Unter geeigneten technischen Annahmen*[3] *und für ausreichend kleine Lernraten sind sowohl simultaner als auch alternierender Gradientenabstieg für R_1/R_2-regularisiertes GAN-Training lokal konvergent. Außerdem ist die Konvergenzrate mindestens linear.*

2.4 Experimente

Wir wenden unsere neuen Regularisierer (R_1 und R_2) unter anderem auf die ImageNet [De09] und CelebA-HQ [Ka18] Datensätze an. ImageNet ist ein schwieriger Datensatz für generative Modelle, da er sehr vielfältig ist. CelebA-HQ führt aufgrund seiner hohen Auflösung (1024×1024 Pixel) zu einem schwierigen Lernproblem.

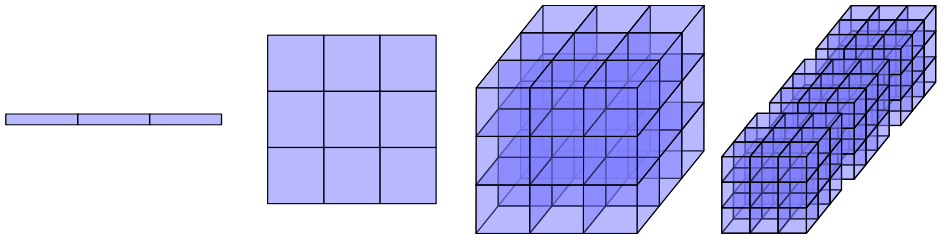
Quantitative Ergebnisse auf ImageNet und ein Vergleich zu Baselines [Go14, Gu17] sind in Abb. 2 zu sehen. Im Gegensatz zu unregularisiertem Training sind sowohl WGAN-GP als auch GAN-Training mit R_1/R_2-Regularisierung stabil. Außerdem führen unsere Regularisierer im Vergleich zu WGAN-GP zu besseren Endergebnissen, was unsere Theorie bestätigt. In Abb. 3 sehen wir qualitative Resultate auf CelebA-HQ. Wir sehen, dass unsere Regularisierer es ermöglichen, GANs mit einer sehr hohen Auflösung zu trainieren, ohne dass wir - anders als frühere Arbeiten [Ka18] - auf progressiv-wachsende Architekturen zurückgreifen müssen.

3 Tiefes Lernen im Funktionenraum

Wie wir im letzten Abschnitt gesehen haben, ermöglichen einfache Regularisierungstechniken ein stabiles Training von GANs für Bildverteilungen, selbst bei sehr hohen Bild-

[3] Wir nehmen unter anderem die Realisierbarkeit des Lernproblems an, also dass der Generator die Datenverteilung erzeugen kann. Wir nehmen auch an, dass der Diskriminator lokal zwischen der wahren und einer abweichenden, vom Generator erzeugten Verteilung unterscheiden kann. Siehe Kapitel 6 des Hauptdokuments für eine mathematische Ausformulierung dieser Annahmen.

Abbildung 3: **CelebA-HQ.** Stichproben für ein GAN, das auf dem CelebA-HQ-Datensatz [Ka18] mit einer Bildauflösung von 1024×1024 Pixeln trainiert wurde. Während des gesamten Trainings trainieren wir den Generator und Diskriminator auf voller Auflösung, d.h. wir verwenden keine der heuristisch motivierten Techniken aus [Ka18] zur Stabilisierung des Trainings.



(a) $3^1 = 3$ (b) $3^2 = 9$ (c) $3^3 = 27$ (d) $3^4 = 81$

Abbildung 4: **Der Fluch der Diskretisierung.** Bei Erhöhung der Dimensionalität von $\mathcal{S}$ wächst die Anzahl der Zellen in der Diskretisierung $\hat{\mathcal{S}}$ exponentiell. Selbst bei einer sehr geringen Auflösung von drei führt dies bereits zu 81 Zellen, wenn der zugrunde liegende Raum vierdimensional ist.

auflösungen. Unsere Welt ist jedoch nicht zwei-, sondern dreidimensional. Nützliche generative Modelle der realen Welt müssen daher auch mit drei oder mehr Dimensionen umgehen können. Leider gibt es, anders als in 2D, keine kanonische Ausgabedarstellung für diskriminative und generative Modelle in 3D. Der Grund dafür ist ein Phänomen, das wir in dieser Arbeit den *Fluch der Diskretisierung* nennen (Abb. 4): Wenn wir einen Ausgaberaum naiv diskretisieren, wächst der Speicherbedarf exponentiell mit der Dimensionalität des Raums. Während in zwei Dimensionen (z.B. bei Bildern) dieser Effekt noch nicht sehr gravierend ist, ändert sich dies in drei oder mehr Dimensionen, wo das Finden einer ausdrucksstarken und dennoch flexiblen Repräsentation, die sich in das tiefe Lernen integrieren lässt, eine schwierige Aufgabe ist.

In dieser Arbeit schlagen wir eine einfache Lösung für den Fluch der Diskretisierung vor: Indem wir die hochdimensionale Ausgabe des Netzes als Funktionenraum $\mathcal{V}^{\mathcal{S}}$ (Funktionen von $\mathcal{S}$ nach $\mathcal{V}$) uminterpretieren, sind wir in der Lage, die Diskretisierung beim Training der neuronalen Netze vollständig zu vermeiden. Unsere wichtigste Erkenntnis ist, dass ei-

ne Funktion $G_\theta : \mathcal{Z} \to \mathcal{V}^{\mathcal{S}}$ von $\mathcal{Z}$ in einen Funktionenraum $\mathcal{V}^{\mathcal{S}}$ äquivalent als Funktion von einem kartesischen Produkt von Räumen $\mathcal{S} \times \mathcal{Z}$ in einen niedrigdimensionalen Raum $\mathcal{V}$ beschrieben werden kann. Wir nennen den zugehörigen Operator, der eine Funktion mit hochdimensionalen Ausgaberaum auf eine Funktion mit niedrig dimensionalen Ausgaberaum abbildet, den *Funktionenraumoperator*.

3.1 Okkupierungsnetzwerke

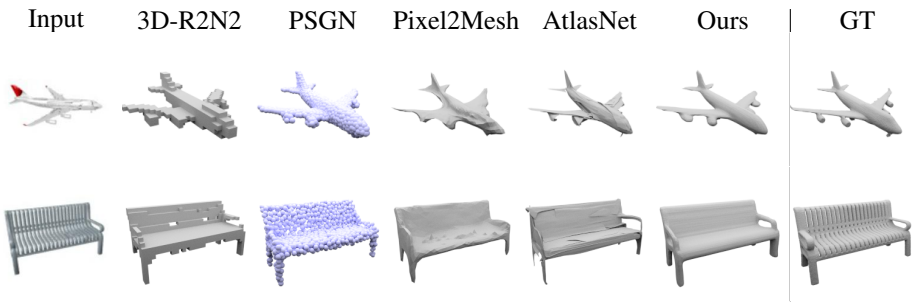


Abbildung 5: **Einzelbild-3D-Rekonstruktion.** Das Eingabebild ist in der ersten Spalte dargestellt, die anderen Spalten zeigen die Ergebnisse für unser Verfahren im Vergleich zu verschiedenen Baselines und der Ground Truth.

In diesem Kapitel stellen wir eine Anwendung des Funktionenraumoperators auf lernbasierte 3D-Rekonstruktion und generative 3D-Modellierung vor.

Bei der Verwendung von Voxeln stellen wir die 3D-Geometrie als diskretes 3D-Belegungsgitter $\hat{\mathcal{S}}$ dar. Im Idealfall möchten wir jedoch nicht nur an diskreten 3D-Orten über die Belegung nachdenken, sondern an *jedem* möglichen 3D-Punkt $s \in \mathcal{S} = [0,1]^3$, wobei $\mathcal{S}$ ein bestimmtes begrenzendes Volumen beschreibt. Wir nennen die resultierende Funktion $x : \mathcal{S} \to \{0,1\}$ die *Okkupierungsfunktion* des 3D-Objekts. Wir können daher die Aufgaben der lernbasierten 3D-Rekonstruktion und der generativen Modellierung als Lernen einer Funktion $\mathcal{Z} \times \mathcal{Y} \to \{0,1\}^{\mathcal{S}}$ mit $\mathcal{Z}$ einem latenten Raum, $\mathcal{Y}$ einem Raum von Beobachtungen (z.B. Bilder, Punktwolken, etc.) und $\mathcal{S} = [0,1]^3$ interpretieren. Durch Anwendung des Funktionenraumoperators können wir dies äquivalent als Lernen einer Funktion $\mathcal{S} \times \mathcal{Z} \times \mathcal{Y} \to \{0,1\}$ beschreiben. Während die Approximation einer Funktion $\mathcal{Z} \times \mathcal{Y} \to \{0,1\}^{\hat{\mathcal{S}}}$ (wie bei Voxel-Darstellungen) unter dem Fluch der Diskretisierung leidet, können wir die Funktion $\mathcal{S} \times \mathcal{Z} \times \mathcal{Y} \to \{0,1\}$ mit einem neuronalen Netzwerk $f_\theta(\cdot)$ approximieren, das ein Tripel (s,z,y) auf eine reelle Zahl abbildet, welche die Wahrscheinlichkeit der Belegung darstellt:

$$f_\theta : \mathcal{S} \times \mathcal{Z} \times \mathcal{Y} \to [0,1] \tag{3}$$

Wir nennen dieses Netzwerk ein Okkupierungsnetzwerk (ONet). Anders als andere Ansätze führt unser Ansatz zu hochauflösenden geschlossenen Oberflächen ohne Selbstüberschneidungen und benötigt kein Polygonnetz aus der gleichen Objektklasse als Vorlage.

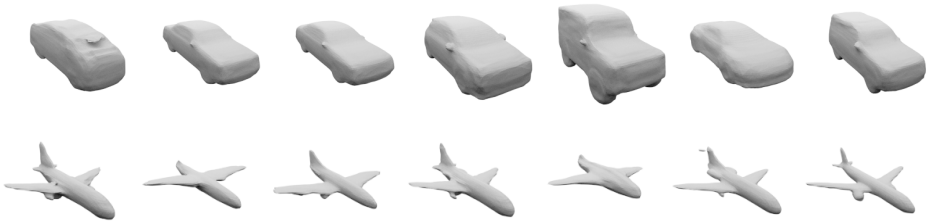


Abbildung 6: **Generative Modelle.** Zufällige Stichproben unserer generativen Modelle, die auf den Kategorien “Auto“ und “Flugzeug“ des ShapeNet-Datensatzes trainiert wurden. Wir sehen, dass unsere Modelle in der Lage sind, die Verteilung von 3D-Objekten zu erfassen und überzeugende neue Samples zu produzieren.

3.2 Experimente

Wir betrachten unter anderem das schwierige Lernproblem, aus einem einzelnen Bild eines Objekts das zugehörige 3D Objekt zu rekonstruieren. Hierzu nutzen wir den ShapeNet Datensatz [Ch15].

Qualitative Ergebnisse für unser Modell und die Baselines sind in Abb. 5 dargestellt. Wir stellen fest, dass alle Methoden in der Lage sind, die 3D-Geometrie des Eingabebildes zu erfassen. Allerdings erzeugt 3D-R2N2 [Ch16] eine sehr grobe Darstellung und kann daher nicht alle Details abbilden. Im Gegensatz dazu erzeugt PSGN [FSG17] eine detailgetreue Ausgabe, der es jedoch an Konnektivität mangelt. Daher sind bei PSGN zusätzliche verlustbehaftete Nachbearbeitungsschritte erforderlich, um ein finales Polygonnetz zu erzeugen. Pixel2Mesh [Wa18] ist in der Lage, überzeugende Polygonnetze zu erstellen, übersieht aber oft Löcher bei komplizierteren Topologien. In ähnlicher Weise erfasst AtlasNet [Gr18] die Geometrie gut, erzeugt aber Artefakte in Form von Selbstüberschneidungen. Im Gegensatz zu den Baselines ist unsere Methode in der Lage, komplexe Topologien zu erfassen, erzeugt geschlossene Oberflächen und bewahrt die meisten Details.

Einige Stichproben unserer rein-generativen Modelle sind in Abbildung 6 dargestellt. Wir sehen, dass unsere generativen Modelle überzeugende neue 3D Objekte generieren können.

4 Fazit

In dieser Arbeit haben wir generative Modelle aus zwei Perspektiven betrachtet.

Als Erstes haben wir die Stabilität von GANs analysiert. Wir haben eine Theorie über die lokale Konvergenz des GAN-Trainings hergeleitet und diese benutzt, um Regularisierer für das GAN-Training abzuleiten. Wir haben herausgefunden, dass (i) GAN-Training unter Eigenwerten der Jacobi-Matrix nahe der imaginären Achse leidet, (ii) naive GAN-Optimierung nicht immer konvergent ist, (iii) null-zentrierte Gradientenregularisierung das Training stabilisieren kann und (iv) unsere Regularisierer es ermöglichen, hochauflösende

GANs zu trainieren, ohne dass ein Training mit progressiv-wachsenden Architekturen erforderlich ist.

Im zweiten Teil der Arbeit haben wir den Funktionenraumoperator vorgeschlagen und ihn auf das 3D tiefe Lernen angewendet. Dies führt zu einer Modellklasse, die wir Okkupierungsnetzwerke nennen. Dabei haben wir herausgefunden, dass (i) Okkupierungsnetzwerke in der Lage sind, hochdimensionale 3D-Geometrie zu repräsentieren, (ii) implizite Repräsentationen effizient aus Daten wie 2D-Bildern gelernt werden können und (iii) implizite Darstellungen zum Lernen von generativen Modellen in 3D verwendet werden können.

Obwohl wir natürlich nicht alle theoretischen and praktischen Fragen beantwortet haben, haben die in dieser Arbeit vorgestellten Forschungsarbeiten bereits wesentlich zur Entwicklung der nächsten Generation von tiefen generativen Modellen beigetragen [KLA19, Ka20, Mi20].

Literatur

[ACB17] Arjovsky, Martin; Chintala, Soumith; Bottou, Léon: Wasserstein generative adversarial networks. In: Proc. of the International Conf. on Machine learning (ICML). 2017.

[Ch15] Chang, Angel X.; Funkhouser, Thomas A.; Guibas, Leonidas J.; Hanrahan, Pat; Huang, Qi-Xing; Li, Zimo; Savarese, Silvio; Savva, Manolis; Song, Shuran; Su, Hao; Xiao, Jianxiong; Yi, Li; Yu, Fisher: , ShapeNet: An Information-Rich 3D Model Repository, 2015.

[Ch16] Choy, Christopher Bongsoo; Xu, Danfei; Gwak, JunYoung; Chen, Kevin; Savarese, Silvio: 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In: Proc. of the European Conf. on Computer Vision (ECCV). 2016.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; jia Li, Li; Li, Kai; Fei-fei, Li: ImageNet: A large-scale hierarchical image database. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2009.

[FSG17] Fan, Haoqiang; Su, Hao; Guibas, Leonidas J.: A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017.

[Ge19] Geirhos, Robert; Rubisch, Patricia; Michaelis, Claudio; Bethge, Matthias; Wichmann, Felix A.; Brendel, Wieland: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: Proc. of the International Conf. on Learning Representations (ICLR). 2019.

[Go14] Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron C.; Bengio, Yoshua: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS). 2014.

[Gr18] Groueix, Thibault; Fisher, Matthew; Kim, Vladimir G.; Russell, Bryan; Aubry, Mathieu: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2018.

[Gu17] Gulrajani, Ishaan; Ahmed, Faruk; Arjovsky, Martín; Dumoulin, Vincent; Courville, Aaron C.: Improved Training of Wasserstein GANs. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.

[He17] Heusel, Martin; Ramsauer, Hubert; Unterthiner, Thomas; Nessler, Bernhard; Hochreiter, Sepp: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.

[Ka18] Karras, Tero; Aila, Timo; Laine, Samuli; Lehtinen, Jaakko: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: Proc. of the International Conf. on Learning Representations (ICLR). 2018.

[Ka20] Karras, Tero; Laine, Samuli; Aittala, Miika; Hellsten, Janne; Lehtinen, Jaakko; Aila, Timo: Analyzing and improving the image quality of StyleGAN. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2020.

[KLA19] Karras, Tero; Laine, Samuli; Aila, Timo: A style-based generator architecture for generative adversarial networks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2019.

[Me20] Mescheder, Lars Morten: , Stability and Expressiveness of Deep Generative Models, 2020.

[Mi20] Mildenhall, Ben; Srinivasan, Pratul P; Tancik, Matthew; Barron, Jonathan T; Ramamoorthi, Ravi; Ng, Ren: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proc. of the European Conf. on Computer Vision (ECCV). 2020.

[NCT16] Nowozin, Sebastian; Cseke, Botond; Tomioka, Ryota: f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In: Advances in Neural Information Processing Systems (NeurIPS). 2016.

[Sa16] Salimans, Tim; Goodfellow, Ian J.; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec; Chen, Xi: Improved Techniques for Training GANs. In: Advances in Neural Information Processing Systems (NeurIPS). 2016.

[Wa18] Wang, Nanyang; Zhang, Yinda; Li, Zhuwen; Fu, Yanwei; Liu, Wei; Jiang, Yu-Gang: Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018.

Lars Mescheder ist angewandter Wissenschaftler im Bereich des maschinellen Lernens und Computersehens. Nach seinem Grundstudium der Mathematik und Informatik an der Technischen Universität Braunschweig war er von 2016 bis 2020 als Doktorand am Max-Planck-Institut für Intelligente Systeme in Tübingen tätig. Dort war er Teil der Autonomous Vision Group unter Prof. Andreas Geiger und wurde von Microsoft Research und dem Intel Network on Intelligent Systems gefördert. In seiner Forschung konzentriert sich Lars Mescheder auf tiefe generative Modelle und 3D Repräsentationen. Aktuell ist er angewandter Wissenschaftler bei Amazon Tübingen und entwickelt dort Algorithmen für Amazon Scout, einen autonomen Auslieferungsroboter.

Beweisbar Gesetzmäßigkeiten in Daten finden und ausnutzen[1]

Stefan Neumann[2]

Abstract: Im letzten Jahrzehnt gab es immensen Fortschritt und Wachstum in den Bereichen der Künstlichen Intelligenz und des Maschinellen Lernens. Ermöglicht wurde diese Entwicklung durch spezialisierte neue Hardware, die immer größere Verfügbarkeit von Daten und Durchbrüche bei der Entwicklung von Algorithmen, die *Gesetzmäßigkeiten in Daten finden und ausnutzen*. Obwohl wir uns in der Praxis täglich vom großen Erfolg dieser Algorithmen überzeugen können, ist unser theoretisches Verständnis von ihnen jedoch weiterhin eingeschränkt. Allerdings wären formale Garantien für diese Algorithmen wünschenswert, weil sie wichtige Einblicke in die Stärken und die Grenzen dieser Algorithmen bieten. Diese Dissertation verkleinert die Kluft zwischen Theorie und Praxis, indem wir Algorithmen enwickeln, die *beweisbar* Gesetzmäßigkeiten in Daten finden und ausnutzen.

1 Einleitung

In den letzten Jahren gab es enorme Fortschritte auf den Gebieten der Künstlichen Intelligenz, des Data Mining und des Maschinellen Lernens. Dies hat sowohl in der Industrie als auch im akademischen Bereich zum immensen Wachstum dieser Bereiche geführt. Treiber dieser Entwicklung waren spezialisierte neue Hardware, die immer größere Verfügbarkeit von Daten und Durchbrüche bei der Entwicklung von Algorithmen, die *Gesetzmäßigkeiten in Daten finden und ausnutzen*.

Viele in der Praxis eingesetzte Algorithmen sind allerdings *Heuristiken*, also Algorithmen ohne mathematisch beweisbare Qualitätsgarantien. Daher können wir uns zwar vom Erfolg dieser Algorithmen in Anwendungen überzeugen, aber unser theoretisches Verständnis von ihnen und den zugrunde liegenden Probleme der Informatik bleibt beschränkt. Man kann sagen, dass der Fortschritt bei der Entwicklung von praktischen Algorithmen in den letzten Jahren so rasant war, dass sich die Kluft zwischen dem, was praktisch möglich ist, und unserem theoretischen Verständnis stark vergrößert hat.

Dennoch ist ein solides theoretisches Verständnis der praktischen Methoden höchst wünschenswert: Die rigorose mathematische Analyse eines Algorithmus liefert wichtige Einblicke in dessen Stärken und deckt zudem seine Grenzen auf. Formale Garantien für einen Algorithmus helfen uns daher, dessen Qualität zu bewerten und sie geben uns uns ein besseres Verständnis dafür, wie viel Vertrauen wir in seine Ausgabe haben können. Letzteres ist besonders wichtig, wenn die Daten sensible Attribute enthalten (etwa Alter oder Geschlecht) oder wenn die Ausgabe des Algorithmus für kritische Entscheidungen verwendet wird (zum Beispiel um zu entscheiden, ob ein Patient eine Krankheit hat oder nicht).

[1] Englischer Titel der Dissertation: “Provably Finding and Exploiting Patterns in Data” [Ne20]

[2] Königliche Technische Hochschule (KTH), Stockholm, Schweden, neum@kth.se

Es ist jedoch kein Zufall, dass nur wenige praktische Algorithmen theoretische Garantien besitzen. Viele der Probleme, die von diesen Algorithmen gelöst werden, sind NP-schwer und oft sogar NP-schwer zu approximieren. Da die Algorithmen in der Praxis aber gute Ergebnisse liefern, suggeriert dies, dass die klassische Worst-Case-Analyse, die üblicherweise bei der Analyse von Algorithmen eingesetzt wird, für diese Klasse von Problemen zu pessimistisch ist: Es ist zwar möglich, künstliche Eingaben zu entwickeln, an denen ein Algorithmus scheitert, aber in der Praxis kommt dies nie vor, weil die praktischen Eingabedaten viel mehr Struktur und *Gesetzmäßigkeiten* (engl. "pattern", auf deutsch auch "Muster") enthalten als die künstlich erzeugten Eingaben. Um also Algorithmen mit theoretischen Garantien zu erhalten, werden wir *über die Worst-Case-Analyse hinaus* gehen und dabei die Gesetzmäßigkeiten *finden und ausnutzen*.

In dieser Dissertation präsentieren wir Algorithmen, die *beweisbar* Gesetzmäßigkeiten in Daten finden und ausnutzen. Dadurch wird die Kluft zwischen dem, was praktisch möglich ist, und unserem theoretischen Verständnis verringert. Die Ergebnisse dieser Arbeit können in die folgenden Kategorien eingeteilt werden:

1. *Gesetzmäßigkeiten nachweislich finden:* Wir entwickeln Algorithmen, die beweisbar eine Menge von versteckten Clustern in bipartiten Zufallsgraphen finden. Dieses Problem muss etwa bei der Analyse von Online-Shopping-Daten gelöst werden, wo man Gruppen ("Cluster") von Produkten finden möchte, die häufig zusammen gekauft werden. Wir präsentieren den ersten Algorithmus, der nachweislich *winzige* versteckte Cluster findet. Dieses Resultat liefert eine theoretische Begründung für den Erfolg von Heuristiken, die in der Praxis angewendet werden. Weiterhin präsentieren wir den ersten Datenstromalgorithmus, der einen bipartiten Graphen nur zwei Mal sequentiell liest und dabei eine Menge von versteckten Clustern findet. Dieser Algorithmus skaliert auf viel größere Datensätze als vorhandene Methoden.

2. *Komplexität des Findens von Gesetzmäßigkeiten:* Wir betrachten häufig verwendete Subroutinen von Data Mining-Algorithmen und erhalten neue Härteresultate für deren Komplexität. Für das approximative Zählen der Dreiecke in einem Graphen und für die Approximation der Häufigkeit von Itemsets in Transaktionsdatenbanken zeigen wir, dass existierende Algorithmen nicht signifikant verbessert werden können, außer gängige Hypothesen der Komplexitätstheorie sind falsch. Außerdem präsentieren wir eine Hierarchie der Enumerationskomplexität von mehreren Maximal Frequent Pattern Mining-Problemen und wir bestimmen eine Bedingung, unter der diese Hierarchie zusammenbricht.

3. *Gesetzmäßigkeiten nachweislich ausnutzen:* Wir formulieren ein Modell, das die Gesetzmäßigkeiten von Kommunikationsanfragen in Rechenzentren abbildet, und entwickeln Algorithmen, die diese Gesetzmäßigkeiten in den Daten ausnutzen und damit den Datenverkehr im Rechenzentrum reduzieren. Wir beweisen, dass der kompetitive Faktor unserer Algorithmen asymptotisch optimal ist. Wir zeigen außerdem, dass verteilte Union-Find-Datenstrukturen mit unserem Modell implementiert werden können. Unsere Algorithmen basieren auf einer neuen Technik, die effiziente ganzzahlige lineare Programmierung ("integer linear programing", ILP) mit der manuellen Berechnung optimaler ILP-Lösungen kombiniert.

Um unsere Ergebnisse zu erhalten, verwenden wir die sogenannte Beyond-Worst-Case-Analyse, die Annahmen an die Eingabedaten macht, um die Eigenschaften zu modellieren, die man von Daten aus der Praxis erwarten würde. Zum Beispiel analysieren wir die Algorithmen aus Punkt 1 für Zufallsgraphen statt für beliebige Graphen. Für die Ergebnisse von Punkt 3 präsentieren wir ein Modell für die Gesetzmäßigkeiten der Kommunikationsanfragen in einem Rechenzentrum. Damit erzielen wir nachweislich bessere Resultate als die bestmöglichen, die keine Annahme an die Kommunikationsanfragen machen. Die Härteresultate aus Punkt 2 werden unter Verwendung der klassischen Worst-Case-Analyse hergeleitet, aber sie dienen als Motivation, diese Probleme in Zukunft mit Hilfe der Beyond-Worst-Case-Analyse zu untersuchen: Unsere hergeleiteten unteren Grenzen sind *scharf* ("tight"), sie zeigen also, dass die Laufzeiten der vorhandenen Algorithmen bezüglich der Worst-Case-Analyse nicht signifikant verbessert werden können. Folglich ist weiterer Fortschritt bei diesen Problemen nur dann möglich, wenn man gewisse Gesetzmäßigkeiten bei den Daten annimmt und diese in der Analyse ausnützt.

2 Gesetzmäßigkeiten nachweislich finden

Ein beliebtes Problem im Unüberwachten Lernen ("unsupervised learning") ist das Finden von Gesetzmäßigkeit in bipartiten Graphen. Dieses Problem ist auch bekannt als *Biclustering* oder *Co-Clustering* und wird seit den 1970er Jahren untersucht [Ha72]. Das Problem wird in verschiedenen Bereichen der Informatik studiert, etwa in Data Mining [Dh01, Mi08], im Maschinellen Lernen [LCX15, ZA19] und der Bioinformatik [MO04].

In diesem Problem ist die Eingabe ein bipartiter Graph $G = (U \cup V, E)$ und die Gesetzmäßigkeiten sind Cluster $U_1, \ldots, U_k \subseteq U$ und $V_1, \ldots, V_k \subseteq V$, sodass jeder *Bicluster* (U_i, V_i) bestimmte Kriterien erfüllt. Beliebte Kriterien sind, dass die Bicluster (U_i, V_i) Subgraphen induzieren, die entweder einen vollständigen bipartiten Graphen ("Biclique") bilden [Or77, GGM04] oder "viele" Kanten relativ zur globalen Dichte ("density") des Graphen enthalten [ZA19, RPG16] (siehe unten für das Problem, das wir untersuchen).

In Anwendungen stehen die beiden Seiten des bipartiten Graphen für Objekte von verschiedenen Typen und eine Kante kennzeichnet die Interaktion der entsprechende Objekte. Beispielsweise kann man Daten von Online-Shops analysieren, indem man die Knoten auf der linke Seite U von G mit den Kunden des Online-Shops identifiziert und die Knoten auf der rechten Seite V von G mit Produkten identifiziert. Eine Kante (u, v) kennzeichnet, dass Kunde u das Produkt v gekauft hat. Jeder Bicluster (U_i, V_i) entspricht daher einer Gruppe von Produkten V_i, die von ähnlichen Kunden U_i gekauft werden.

Reale Datensätze weisen häufig zwei entscheidende Eigenschaften auf: Auf der einen Seite des bipartiten Graphen sind die Knotengrade beschränkt und auf der anderen Seite des bipartiten Graphen enthalten die Cluster V_j nur sehr wenige Knoten. In der Praxis sind diese Annahme etwa im obigen Online-Shop-Beispiel erfüllt: Fast alle Kunden kaufen höchstens einige Hundert verschiedene Produkte, daher ist der Grad der Knoten in U beschränkt. Außerdem wird in Experimenten oft beobachtet, dass die Produkt-Cluster V_j meist aus weniger als 50 Produkten bestehen. Typische solche Produkt-Cluster sind die

7 Harry Potter-Bücher oder die 23 Filme aus dem Marvel Cinematic Universe. Daher sind die Cluster V_j viel kleiner als V, denn V enthält häufig Millionen Produkte. Man kann also sagen, dass die Cluster V_j *winzig* sind im Vergleich zur Größe von V.

Zwei der größten Herausforderungen in diesem Bereich waren folgende: (1) Während viele praktische Algorithmen winzige Cluster V_j finden können, konnten die besten theoretischen Ergebnisse nur das Finden von mittelgroßen Clustern garantieren (Einzelheiten siehe unten). Können wir diese Lücke schließen? (2) Moderne praktische Algorithmen liefern sehr gute Ergebnisse für Graphen mit Tausenden Knoten, aber sie skalieren nicht auf Graphen mit mehreren Hunderttausend oder gar Millionen Knoten. Können wir effizientere Algorithmen erhalten, indem wir ausnutzen, dass die Graphen in der Praxis sehr dünnbesetzt sind? *Wir beantworten beide Fragen positiv.*

Das Argument für Zufallsgraphen. Die Tatsache, dass frühere Algorithmen für diese Probleme meist Heuristiken waren, ist kein Zufall. Tatsächlich sind die meisten Probleme zum Finden von Clustern in bipartiten Graphen NP-schwer wenn man *Worst-Case-Eingaben* betrachtet [Or77]. Zudem gibt es starke untere Schranken für die Laufzeit von Approximationsalgorithmen [CIK16].

Um diese Härteergebnisse zu umgehen, betrachten wir ein gängiges Zufallsgraphenmodell (siehe unten für die formale Definition). Dieses Modell wurde in verschiedenen Wissenschaftsbereichen studiert, vom Maschinellem Lernen über die Theoretische Informatik bis hin zu Mathematik und Physik [Ab18]. Dank diesem Modell kann man Algorithmen entwickeln, die garantieren, dass sie eine Menge von *versteckten* Clustern $U_1,\dots,U_k$ und $V_1,\dots,V_k$ finden. Darüber hinaus wurden verschiedene Heuristiken (ohne theoretische Garantien) mit Hilfe von ähnlichen Modellen entwickelt [RPG16, Ru17] und diese Algorithmen liefern hervorragende Ergebnisse in der Praxis. Das deutet darauf hin, dass die Annahmen des Zufallsgraphenmodells die Realität gut abbilden.

Die formale Definition des Zufallsgraphenmodell ist wie folgt. Sei $G = (U \cup V, E)$ ein bipartiter Graph mit k *versteckten* Clustern $U_1,\dots,U_k \subseteq U$ und $V_1,\dots,V_k \subseteq V$ auf jeder Seite des Graphen. Ferner seien $p,q \in [0,1]$ Wahrscheinlichkeiten mit $p > q$. Wir nehmen an, dass Kanten (u,v) zwischen Knoten $u \in U_i$ und $v \in V_i$ mit Wahrscheinlichkeit p eingefügt werden und Kanten (u,v) mit $u \in U_i$ und $v \in V_j$ mit $i \neq j$ mit Wahrscheinlichkeit q eingefügt werden. Dadurch gibt es "relativ viele" Kanten zwischen den Knoten jedes Biclusters (U_i,V_i) im Vergleich zur globalen Dichte des Graphen, wo es "relativ wenige" Kanten gibt. Das Problem ist nun wie folgt: Bei Eingabe eines Zufallsgraphen, der entsprechend der oben beschriebenen Verteilung erzeugt wurde, muss der Algorithmus die verstecken Cluster $U_1,\dots,U_k$ und $V_1,\dots,V_k$ finden.

Winzige Cluster finden. Entsprechend unserer Argumentation oben sind die rechten Cluster V_j in der Praxis winzig. Während praktische Algorithmen höchst erfolgreich bei der Identifizierung dieser Cluster sind, konnten vorhandene Algorithmen mit theoretischen Garantien [Mc01, LCX15] nur mittelgroße Cluster der Größe $|V_j| = \Omega(\sqrt{n})$ finden, wobei $n = |V|$ die Anzahl der Knoten ist. Dies ist in dem oben beschriebenen praktischen Szenario unrealistisch (etwa für $|V| \geq 10^6$ erhalten wir $\sqrt{|V|} \geq 10^3$). Daher gibt es eine große Kluft zwischen den praktischen Ergebnissen und ihrer theoretischen Begründung.

Wir schließen diese Lücke zwischen Theorie und Praxis, indem wir einen praktischen Algorithmus präsentieren, der beweisbar winzige Cluster findet. Für alle $\varepsilon > 0$ zeigen wir, dass man Cluster V_j der Größe $|V_j| = O(n^\varepsilon)$ finden kann, wenn einige Bedingungen an die Paramater p, q und die Größe der linken Cluster U_i erfüllt sind. Vorherige Algorithmen benötigen $\varepsilon \geq \frac{1}{2}$. Die Garantien gelten auch dann, wenn der Graph extrem dünnbesetzt ist und der Grad jedes Knoten nur polylogarithmisch in der Gesamtzahl der Knoten ist.

Darüber hinaus zeigen Experimente, dass unser Algorithmus auf künstlich erzeugten Daten bessere Ergebnisse erzielt als existierende Heuristiken und dass die in realen Daten gefundenen Cluster von hoher Qualität ist. Somit kombiniert der Algorithmus theoretische Garantien mit praktischer Leistung.

Effiziente Datenstromalgorithmen. Eine zweite wichtige Frage in diesem Bereich ist es, vorhandene Methoden skalierbarer zu machen. Wir untersuchen dieses Problem im Datenstrommodell, bei dem die Eingabe ein Datenstrom der linken Knoten $u \in U$ zusammen mit all ihren inzidenten Kanten ist. Das Ziel ist, einen Algorithmus zu entwickeln, der den Graphen nur wenige Male sequentiell liest und der zudem wenig Speicher benötigt.

Wir präsentieren den ersten Datenstromalgorithmus ("streaming algorithm"), der den Graphen *nur einmal* sequentiell liest und dabei die rechten Cluster V_i findet. Der Algorithmus ist extrem speichereffizient: Der verwendete Speicherplatz ist lediglich $O(ks \log m)$, wobei $m = |U|$ die Anzahl der Knoten im Datenstrom ist, k ist die Anzahl der Cluster und s ist eine obere Schranke für die Knotengrade der Knoten in U (oben hatten wir argumentiert, dass s in realen Datensätzen klein ist, etwa weil Kunden in Online-Shops nur wenige Hundert Produkte kaufen). Damit ist die Speicherplatznutzung nur einen $O(\log m)$-Faktor höher als nötig, um überhaupt die Cluster V_i zu speichern (dies würde $O(ks)$ Speicher benötigen). Für eine Version des Algorithmus beweisen wir, dass er den informationstheoretisch optimalen Speicherplatz verwendet und zudem die versteckten Cluster findet, wenn der Graph anhand des Zufallsgraphenmodells erzeugt wurde. Nach einem zweiten sequentiellen Lesens des Datenstroms kann der Algorithmus außerdem die linken Cluster finden.

Zudem erweitern wir den Algorithmus, sodass er Boolesche Matrixfaktorisierungen (BMF) berechnen kann, was ein wichtiges Problem in den Bereichen des Data Mining [Mi08, HPM18] und des Maschinellen Lernens [RPG16, Ru17] ist. Es ist zudem der erste Datenstromalgorithmus für BMF, den es jemals gab.

In Experimenten mit realen Datensätzen ist der Algorithmus um Größenordnungen schneller und speichereffizienter als ein klassischer Algorithmus, der nicht im Datenstrommodell arbeitet. Insbesondere verwendet unser Algorithmus in keinem Datensatz mehr als 500 MB RAM, selbst wenn die Graphen Millionen von Knoten und Kanten enthalten. Weiterhin hat er hat die wünschenswerte Eigenschaft, dass seine Laufzeit linear in der Anzahl der Kanten des Graphen skaliert. Zudem findet der Algorithmus Cluster von hoher Qualität, die innerhalb von Faktor 2 von der Qualität des klassischen Algorithmus liegen.

3 Die Komplexität des Findens von Gesetzmäßigkeiten

Als nächstes entwickeln wir ein besseres Verständnis der Komplexität von fundamentalen Problemen im Zusammenhang mit der Suche nach Gesetzmäßigkeiten in Daten.

Komplexität der Approximation von Häufigkeiten. Einige der am häufigsten verwendeten Subroutinen in Data Mining-Algorithmen dienen der Berechnung der *Häufigkeit* von Gesetzmäßigkeiten. Wir fokussieren uns auf zwei wichtige Subroutinen: Das Zählen von Dreiecken in Graphen und die Bestimmung von Häufigkeiten in Transaktionsdatenbanken.

Wenn etwa die Eingabe ein Graph G ist und die gesuchte Gesetzmäßigkeit ein Dreieck ist, ist das Ziel, die Anzahl von Dreiecken $\#\Delta(G)$ in G zu zählen. Das Zählen von Dreiecken in Graphen ist von großer Wichtigkeit bei der Analyse von sozialen Netzwerken und in der Bioinformatik bei der Analyse von chemischen Strukturen.

In Transaktionsdatenbanken ist das Ziel, die Häufigkeit von *Itemsets* (Elementemengen) zu berechnen. Formal ist dieses Problem wie folgt definiert: Sei $[d] = \{1, \dots, d\}$ eine Menge von Elementen ("items"), dann ist ein *Itemset* eine Teilmenge von $[d]$. Eine $m \times d$ *Transaktionsdatenbank* $\mathscr{D}$ ist eine (Multi-)Menge $\mathscr{D} = \{T_1, \dots, T_m\}$, wobei jedes T_i ein Itemset über $[d]$ ist. Wir nennen $\#\mathsf{supp}(T) = |\{i : T \subseteq T_i\}|$ die *Häufigkeit* ("support") des Itemset T, $\#\mathsf{supp}(T)$ ist also die Anzahl der *Transaktionen* $T_i \in \mathscr{D}$ mit $T \subseteq T_i$. In der Praxis tritt dieses Problem in Online-Shops auf: Wenn die Elemente in $[d]$ die Produkte eines Online-Shops sind, dann entspricht eine Transaktion T_i den von einem Kunden gekauften Produkten und $\#\mathsf{supp}(T)$ zählt die Anzahl der Kunden, die die Produkte in T gekauft haben.

Obwohl $\#\mathsf{supp}(T)$ und $\#\Delta(G)$ in der Praxis häufig berechnet werden müssen, werden die Subroutinen zur Berechnung dieser Größen in der Praxis weiterhin als langsam betrachtet. Wenn wir also die Berechnung von $\#\mathsf{supp}(T)$ oder $\#\Delta(G)$ beschleunigen könnten, würde dies viele praktische Algorithmen beschleunigen. Leider implizieren gängige Komplexitätshypothesen, dass Algorithmen die *exakten* Werte von $\#\mathsf{supp}(T)$ und $\#\Delta(G)$ nicht wesentlich schneller berechnen können als mittels erschöpfender Suche [Wi05]. Da die exakte Berechnung von $\#\mathsf{supp}(T)$ und $\#\Delta(G)$ häufig zu langsam ist, greifen einige Algorithmen auf die *approximative* Berechnung dieser Größen zurück, um eine schnellere Laufzeit zu erhalten [MTV94, Li16, RU14]. Nun ist es eine natürliche Frage, wie schnell diese approximative Berechnung durchgeführt werden kann.

Daher untersuchen wir die feinkörnige ("fine-grained") Komplexität der *Approximation* von $\#\mathsf{supp}(T)$ und $\#\Delta(G)$. Feinkörnige Komplexität [Wi18] ist ein relativ neuer Bereich in der Komplexitätstheorie, der darauf abzielt, scharfe untere Schranken für die Laufzeit von Problemen aus der Informatik zu beweisen. Diese unteren Schranken werden mit Hilfe von Hypothesen bewiesen, die stärker sind als die Standardannahme $\mathsf{P} \neq \mathsf{NP}$. Wir betrachten *Lückenversionen* ("gap version") von $\#\mathsf{supp}(T)$ und $\#\Delta(G)$. Für einen gegebenen Schwellenwert ("threshold") K muss entschieden werden, ob $\#\mathsf{supp}(T) \geq K$ oder $\#\mathsf{supp}(T) \leq K/3$. Falls $K/3 < \#\mathsf{supp}(T) < K$, darf der Algorithmus irgendeine Antwort zurückgeben. Dieses Problem wurde bereits in der Vergangenheit aus der Perspektive von oberen Schranken untersucht [MTV94, Li16, RU14], da man effiziente Subroutinen für diese Probleme verwenden kann, um vorhandene Algorithmen zu beschleunigen.

Für die Lückenversion von $\#\mathsf{supp}(T)$ erhalten wir eine scharfe untere Schranke, die wir nun formal beschreiben. Angenommen die Eingabe sind eine Transaktionsdatenbank mit m Transaktionen und ein Schwellenwert K. Ein einfacher Algorithmus zur Lösung der Lückenversion von $\#\mathsf{supp}(T)$ wählt zufällig $O((m/K)\log m)$ Transaktionen und approximiert anhand dieser Zufallsstichprobe $\#\mathsf{supp}(T)$. Wir zeigen, dass dieser Algorithmus nicht wesentlich verbessert werden kann: Sofern eine gängigen Komplexitätshypothese korrekt ist, kann die Lückenversion von $\#\mathsf{supp}(T)$ nicht wesentlich schneller gelöst werden kann als in Zeit[3] $(m/K)^{1-o(1)}$. Damit liefern wir eine vollständige Erklärung der Komplexität der Lückenversion von $\#\mathsf{supp}(T)$ bis auf $m^{o(1)}$-Faktoren.

Wir erhalten scharfe untere Schranken für die Lückenversion von $\#\Delta(G)$, bei der wir Dreiecke in einem Graphen mit n Knoten zählen. In der Lückenversion von $\#\Delta(G)$ müssen wir entscheiden, ob $\#\Delta(G) \geq K$ oder $\#\Delta(G) \leq K/3$. Wir zeigen, dass ein Algorithmus, der die Lückenversion von $\#\Delta(G)$ schneller als in Zeit $(n^3/K)^{1-o(1)}$ entscheidet und keine schnelle Matrixmultiplikation verwendet, zu einem Durchbruch im Bereich der kombinatorischen Algorithmen führt. Weiterhin zeigen wir, dass unsere untere Schranke für die Lückenversion von $\#\Delta(G)$ untere Schranken für die approximative Berechnung wichtiger Graphmetriken impliziert. Zum Beispiel beweisen wir untere Schranken für die approximative Berechnung des Clustering-Koeffizienten oder der Transitivität eines Graphen.

Wir präsentieren scharfe obere und untere Schranken für eine Lückenversion von SAT, bei der wir entscheiden müssen, ob eine SAT-Formel mindestens K erfüllende Zuweisungen besitzt *oder gar keine*. Wir zeigen unter Annahme einer gängigen Komplexitätshypothese, dass dieses Problem nicht schneller als in Zeit $(2^n/K)^{1-o(1)}$ gelöst werden kann.

Auflisten maximaler Gesetzmäßigkeiten. Eine weitere häufig verwendete Subroutine in Data Mining-Algorithmen ist es, alle *maximalen* Gesetzmäßigkeiten zu finden. Für eine Transaktionsdatenbank $\mathscr{D}$ und einen Schwellenwert K heißt ein Itemset T *häufig*, wenn $\#\mathsf{supp}(T) \geq K$; andernfalls heißt es *selten*. Außerdem ist T *maximal häufig*, wenn T häufig ist und für alle Itemsets T' mit $T \subsetneq T'$ gilt, dass T' selten ist. Nun besteht das Problem darin, alle maximal häufigen Itemsets auszugeben. Dieses Problem wurde auch für andere Datentypen untersucht, bei denen die Datenbanken aus Graphen oder Sequenzen bestehen.

Beim Auflisten aller maximal häufigen Itemsets ("maximal frequent pattern mining") müssen alle maximal häufigen Itemsets berechnet und ausgegeben werden. Da es potentiell exponentiell (in $|\mathscr{D}|$) viele maximal häufige Itemsets gibt und die Ausgabe viel größer als die Eingabe werden kann, ist es dienlich, die Komplexität dieses Problems nicht nur in der Größe der Eingabe zu messen, sondern in der Größe der Eingabe *und Ausgabe* [KK14]. Dies führt zu Aufzählungs- und Erweiterbarkeitsproblemen. Ein typisches Problem in diesem Bereich wäre beispielsweise, eine gegebene Menge von maximal häufigen Itemsets zu erweitern: gegeben eine Menge von maximal häufigen Itemsets, berechne ein weiteres maximal häufiges Itemset, das nicht in der Eingabemenge enthalten ist, oder gebe aus, dass kein solches maximal häufiges Itemset existiert. Solche Probleme werden im Bereich der *Aufzählungskomplexität* ("enumeration complexity") studiert [JPY88].

[3] Wenn ein Problem "nicht schneller als in Zeit $T^{\alpha-o(1)}$" gelöst werden kann, entspricht dies der folgenden formalen Aussage: "Für alle $\varepsilon > 0$ gibt es ist keinen Algorithmus mit Laufzeit $O(T^{\alpha-\varepsilon})$." Mit anderen Worten: Es gibt keinen Algorithmus mit einer polynomiell schnelleren Laufzeit als $O(T^{\alpha})$.

Wir entwickeln eine *Hierarchie* der Aufzählungskomplexität von Data Mining-Problemen für verschiedene Datentypen. Die Datentypen, die wir berücksichtigen, umfassen Transaktions-, Sequenz- und Graphdatenbanken von verschiedene Graphklassen (von Bäumen bis zu allgemeinen Graphen). Wir zeigen, dass die Hierarchie zusammenbricht (also alle Probleme die gleiche Komplexität haben), wenn wir das Problem leicht verallgemeinern und die Menge der erlaubten Itemsets leicht einschränken dürfen.

4 Gesetzmäßigkeit nachweislich ausnutzen

Wir betrachten zudem Gesetzmäßigkeiten von Kommunikationsanfragen in Rechenzentren. In der Praxis kann empirisch gezeigt werden, dass es diese Gesetzmäßigkeiten gibt und dass Rechenzentren, die sich an diese Gesetzmäßigkeit anpassen, geringere Kosten haben [Ha14]. Daher entwickeln wir Algorithmen, die die Gesetzmäßigkeiten *ausnutzen*.

Wir entwickeln ein formales Modell, das die Struktur der Gesetzmäßigkeit in Rechenzentren erfasst. Wir nehmen an, dass das Rechenzentrum ℓ Server enthält, auf die n Workloads verteilt sind. Zu Beginn ist jeder Workload einem Server zugewiesen und jeder Server hat eine Kapazität, um $(1+\varepsilon)k$ Workloads zu speichern, wobei $k = n/\ell$ und $\varepsilon > 0$ eine Konstante ist. Anschließend beginnen die Workloads mittels Kommunikationsanfragen über das Netzwerk zu kommunizieren. Diese Folge von Kommunikationsanfragen erhält der Algorithmus während seiner Laufzeit (“online”).

Unsere Annahme für die Kommunikationsanfragen ist, dass sie *kleine Gesetzmäßigkeiten induzieren*: Wir betrachten die Workloads als Knoten eines Graphen und die Kommunikationsanfragen als Kanten. Unser Modell geht davon aus, dass, nachdem wir alle Kommunikationsanfragen bearbeitet haben, die Zusammenhangskomponenten im resultierenden Graphen höchstens k Knoten enthalten (es kann also jede Komponente auf einem einzelnen Server gespeichert werden). Diese Zusammenhangskomponenten sind die *Gesetzmäßigkeiten* der Workloads, die durch die Kommunikationsanfragen induziert werden. Weiterhin gehen wir davon aus, dass nach Abschluss der Kommunikationsanfragen alle Knoten aus derselben Komponente auf demselben Server gespeichert sein müssen.

Die Kosten für Kommunikationsanfragen und für das Verschieben von Workloads sind wie folgt. Wenn die kommunizierenden Workloads dem selben Server zugewiesen sind, hat diese Kommunikationsanfrage Kosten 0; andernfalls betragen die Kosten 1. Der Algorithmus kann sich jederzeit entscheiden, Workloads zwischen den Servern zu verschieben, wobei die Kapazitätsbeschränkungen der Server eingehalten werden müssen; für jedes Verschieben muss der Algorithmus $\alpha > 1$ bezahlen. Wir analysieren den Algorithmus mittels einer *Kompetitivitätsanalyse* (“competitive analysis”): wir teilen die von unserem Algorithmus gezahlten Kosten durch die Kosten des optimalen Offline-Algorithmus, der alle Kommunikationsanfragen im Voraus kennt und minimale Kosten hat.

Für dieses Modell entwickeln wir Algorithmen und zeigen, dass ihre Kompetitivität (bis auf konstante Faktoren) optimal ist. Das Hauptergebnis ist ein *randomisierter* Algorithmus mit Kompetitivität (“competitive ratio”) $O(\log \ell + \log k)$. Wir beweisen zudem eine passende untere Schranke von $\Omega(\log \ell + \log k)$ für randomisierte Algorithmen. Darüber hinaus

erhalten wir einen *deterministischen* Algorithmus mit Kompetitivität $O(\ell \log k)$ und eine passende untere Schranke von $\Omega(\ell \log k)$. Als Anwendung unserer Algorithmen zeigen wir, dass sie zur Implementierung von verteilten Union-Find-Datenstrukturen mit beinahe optimaler Netzwerkkommunikation (bis auf $O(1)$-Faktoren) verwendet werden können.

Ohne die Annahme an die Gesetzmäßigkeiten hätte der bestmögliche deterministische Algorithmus eine Kompetitivität von $\Omega(k)$ für $\ell = O(1)$ [Av19]. Dies zeigt, dass das *Ausnutzen* der Gesetzmäßigkeiten notwendig ist, um bessere Algorithmen zu erhalten.

Um diese Ergebnisse zu erzielen, entwickeln wir eine neue Technik, die effiziente ganzzahlige lineare Programmierung ("integer linear programming", ILP) mit der manuellen Wartung optimaler ILP-Lösungen kombiniert: Wenn neue Kommunikationsanfragen eingehen, verwenden wir ein ILP, um die Workloads den Servern zuzuweisen (dies ähnelt klassischen Algorithmen für Prozess-Scheduling). Wir können jedoch unsere Garantien nicht erreichen, wenn wir bei jeder Kommunikationsanfrage das ILP neu lösen. Stattdessen identifizieren wir bestimmte Typen von Kommunikationsanfragen, nach denen wir manuell eine optimale ILP-Lösung erhalten können ohne einen Workload zu verschieben.

Literaturverzeichnis

[Ab18] Abbe, Emmanuel: Community Detection and Stochastic Block Models: Recent Developments. J. Mach. Learn. Res., 18(177):1–86, 2018.

[Av19] Avin, Chen; Bienkowski, Marcin; Loukas, Andreas; Pacut, Maciej; Schmid, Stefan: Dynamic Balanced Graph Partitioning. In: SIAM J. Discrete Math. 2019.

[CIK16] Chandran, L. Sunil; Issac, Davis; Karrenbauer, Andreas: On the Parameterized Complexity of Biclique Cover and Partition. In: IPEC. S. 11:1–11:13, 2016.

[Dh01] Dhillon, Inderjit S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: KDD. S. 269–274, 2001.

[GGM04] Geerts, Floris; Goethals, Bart; Mielikäinen, Taneli: Tiling Databases. In: DS. S. 278–289, 2004.

[Ha72] Hartigan, John A.: Direct Clustering of a Data Matrix. J. Am. Stat. Assoc., 67(337):123–129, 1972.

[Ha14] Hamedazimi, Navid; Qazi, Zafar; Gupta, Himanshu; Sekar, Vyas; Das, Samir R; Longtin, Jon P; Shah, Himanshu; Tanwer, Ashish: Firefly: A reconfigurable wireless data center fabric using free-space optics. In: SIGCOMM. Jgg. 44, S. 319–330, 2014.

[HPM18] Hess, Sibylle; Piatkowski, Nico; Morik, Katharina: The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization. In: SDM. S. 405–413, 2018.

[JPY88] Johnson, David S.; Papadimitriou, Christos H.; Yannakakis, Mihalis: On Generating All Maximal Independent Sets. Inf. Process. Lett., 27(3):119–123, 1988.

[KK14] Kimelfeld, Benny; Kolaitis, Phokion G.: The Complexity of Mining Maximal Frequent Subgraphs. ACM Trans. Database Syst., 39(4):32:1–32:33, 2014.

[LCX15] Lim, Shiau Hong; Chen, Yudong; Xu, Huan: A Convex Optimization Framework for Bi-Clustering. In: ICML. S. 1679–1688, 2015.

[Li16] Liberty, Edo; Mitzenmacher, Michael; Thaler, Justin; Ullman, Jonathan: Space Lower Bounds for Itemset Frequency Sketches. In: PODS. S. 441–454, 2016.

[Mc01] McSherry, Frank: Spectral Partitioning of Random Graphs. In: FOCS. S. 529–537, 2001.

[Mi08] Miettinen, Pauli; Mielikäinen, Taneli; Gionis, Aristides; Das, Gautam; Mannila, Heikki: The Discrete Basis Problem. IEEE Trans. Knowl. Data Eng., 20(10):1348–1362, 2008.

[MO04] Madeira, Sara C.; Oliveira, Arlindo L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. Comput. Biology Bioinform., 1(1):24–45, 2004.

[MTV94] Mannila, Heikki; Toivonen, Hannu; Verkamo, A. Inkeri: Efficient Algorithms for Discovering Association Rules. In: AAAI Workshops. Technical Report WS-94-03. S. 181–192, 1994.

[Ne18] Neumann, Stefan: Bipartite Stochastic Block Models with Tiny Clusters. In: NeurIPS. S. 3871–3881, 2018. Ein extended abstract von diesem Paper erschien bei der INFORMATIK 2019 unter dem Titel "Finding Tiny Clusters in Bipartite Graphs" in der Session *Best of Data Science Made in Germany, Austria and Switzerland*.

[Ne20] Neumann, Stefan: Provably Finding and Exploiting Patterns in Data. Dissertation, University of Vienna, 2020.

[Or77] Orlin, James: Contentment in graph theory: covering graphs with cliques. Indagationes Mathematicae, 80(5):406–424, 1977.

[RPG16] Ravanbakhsh, Siamak; Póczos, Barnabás; Greiner, Russell: Boolean Matrix Factorization and Noisy Completion via Message Passing. In: ICML. S. 945–954, 2016.

[RU14] Riondato, Matteo; Upfal, Eli: Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees. ACM Trans. Knowl. Discov., 8(4):20:1–20:32, 2014.

[Ru17] Rukat, Tammo; Holmes, Christopher C.; Titsias, Michalis K.; Yau, Christopher: Bayesian Boolean Matrix Factorisation. In: ICML. S. 2969–2978, 2017.

[Wi05] Williams, Ryan: A new algorithm for optimal 2-constraint satisfaction and its implications. Theor. Comput. Sci., 348(2-3):357–365, 2005.

[Wi18] Williams, Virginia Vassilevska: On some fine-grained questions in algorithms and complexity. In: Proc. ICM. 2018.

[ZA19] Zhou, Zhixin; Amini, Arash A.: Analysis of spectral clustering algorithms for community detection: the general bipartite setting. J. Mach. Learn. Res., 20:47:1–47:47, 2019.

Stefan Neumann ist Postdoktorand an der Königlichen Technischen Hochschule (KTH) in Stockholm, Schweden, in der Gruppe von ACM Fellow Aristides Gionis. Er promovierte an Universität Wien unter der Betreuung von ACM Fellow Monika Henzinger; während dieser Zeit besuchte er ACM Fellow Eli Upfal für sechs Monate an der Brown University in Providence, RI, USA. Für seine Dissertation hat er den *Award of Excellence* von der österreichischen Bundesregierung erhalten. Sein Paper [Ne18] wurde bei der INFORMATIK 2019 als *Best of Data Science Made in Germany, Austria and Switzerland* präsentiert. Vor seiner Promotion erhielt er einen Master in Informatik von der Universität des Saarlandes und dem Max Planck Institut für Informatik und einen Bachelor in Mathematik von der Universität Jena.

Ausdrucksstärke und Entscheidbarkeit gewichteter Automaten und gewichteter Logiken[1]

Erik Paul[2]

Abstract: Die Dissertation untersucht gewichtete Automaten als Erweiterung des fundamentalen Modells der endlichen Automaten sowie gewichtete Logiken als quantitative Erweiterung der monadischen Logik zweiter Stufe. Als erstes Resultat zeigen wir Zerlegungssätze für eine generische gewichtete Logik, welche sich als gewichtete Verallgemeinerungen in die Familie der Feferman-Vaught-Sätze für die klassische monadische Logik zweiter Stufe einreihen. Im zweiten Resultatkomplex beweisen wir vier Entscheidbarkeitsresultate für das Automatenmodell der Max-Plus-Baumautomaten. Wir zeigen, dass die Äquivalenz endlich mehrdeutiger Max-Plus-Baumautomaten entscheidbar ist. Hierbei heißt ein Baumautomat endlich mehrdeutig, falls die Anzahl der Läufe des Automaten auf jedem Baum durch eine globale Konstante beschränkt ist. Für diese endlich mehrdeutigen Automaten zeigen wir des Weiteren, dass es entscheidbar ist, ob sich ein gegebener Automat auch durch einen Automaten beschreiben lässt, der höchstens einen Lauf auf jedem Baum zulässt, sowie, dass es für einen solchen eindeutigen Automaten entscheidbar ist, ob dieser sich als Maximum endlich vieler deterministischer Automaten darstellen lässt oder sogar zu einem deterministischen Automaten äquivalent ist. Das letzte Resultat verbindet Automaten und Logiken. Wir zeigen, dass sich Quantitative Monitorautomaten durch eine gewichtete Logik beschreiben lassen.

1 Einführung

Automatentheorie, begründet in der Mitte des 20. Jahrhunderts, stellt heute eines der wichtigsten Teilgebiete der theoretischen Informatik dar. Eines der fundamentalsten Berechnungsmodelle, welches die Automatentheorie untersucht, sind *endliche Automaten*. Diese ermöglichen es uns, endliche Darstellungen für potenziell unendlichen Mengen von Zeichenketten anzugeben. Derartige Mengen von Zeichenketten oder *Wörtern* werden auch als *Sprachen* bezeichnet, und die von endlichen Automaten definierbaren Sprachen als *reguläre* Sprachen. Nach ihrer Konzeptualisierung haben sich endliche Automaten aufgrund ihrer guten Entscheidbarkeitseigenschaften schnell zu einem zentralen Forschungsgegenstand entwickelt. So wurde etwa bewiesen, dass reguläre Grammatiken, reguläre Ausdrücke und monadische Logik zweiter Stufe (MSO-Logik) die gleiche Ausdrucksstärke wie endliche Automaten besitzen. Aktuell finden endliche Automaten und temporale Logiken Anwendung in der *Modellprüfung* (model checking) im Rahmen von Software-Verifikationsmethoden. Für diese sind Entscheidbarkeitsresultate essenziell. Andererseits wurden erfolgreiche Bestrebungen unternommen, die fundamentale Idee, auf der das Konzept der endlichen Automaten beruht, von Zeichenketten auf andere Strukturen zu erweitern. So zum Beispiel auf *unendliche Worte* durch Büchi und Muller, *endliche Bäume*

[1] Englischer Titel der Dissertation: “Expressiveness and Decidability of Weighted Automata and Weighted Logics”

[2] Universität Leipzig, epaul@informatik.uni-leipzig.de

durch Doner sowie Thatcher und Wright, *unendliche Bäume* durch Rabin, *Schachtelworte* durch Alur und Madhusudan und *Bilder* durch Blum und Hewitt. Parallel hierzu führte Schützenberger das Modell *gewichteter Automaten* ein, welches die Beschreibung quantitativer Eigenschaften von Sprachen erlaubt. In Folgearbeiten wurden viele dieser Erweiterungen zusammengeführt und die so entstehenden Modelle untersucht und verglichen. Beispielsweise wurden reguläre Ausdrücke, reguläre Grammatiken und logische Charakterisierungen für reguläre Baum- bzw. Bildsprachen entwickelt, genauso wie *gewichtete reguläre Ausdrücke* [Sc61] und *gewichtete Logiken* [DG07] für gewichtete Sprachen von Zeichenketten. Die Dissertation [Pa20a] untersucht zwei dieser Erweiterungen und ihre Beziehung zueinander: gewichteten Automaten und gewichteten Logiken.

2 Feferman-Vaught-Sätze

Im ersten Teil der Dissertation erarbeiten wir gewichtete Erweiterungen des Feferman-Vaught-Satzes [FV59], eines der fundamentalen Resultate der klassischen Modelltheorie. Der klassische Feferman-Vaught-Satz besagt, dass die Auswertung eines Satzes der Prädikatenlogik erster Stufe auf einem verallgemeinerten Produkt relationaler Strukturen zurückgeführt werden kann auf die Auswertung von Sätzen erster Stufe auf den Komponenten, aus denen das Produkt besteht, sowie die Auswertung eines Satzes zweiter Stufe auf der Indexmenge der Komponenten. Fasst man eine Formel erster Stufe als Frage auf, die von einer Struktur mit ja oder nein beantwortet wird, je nachdem ob die Struktur die Formel erfüllt, kann man den Feferman-Vaught-Satz als Zerlegungssatz verstehen. Eine Frage über ein Produkt von Strukturen lässt sich so in Fragen über die einzelnen Komponenten des Produkts zerlegen, dass sich die Antworten auf diese zu einer Antwort auf die ursprüngliche Frage kombinieren lassen. Der Satz selbst hat seine Ursprünge in Arbeiten von Mostowski, wurde von Feferman und Vaught erstmals formuliert und später von verschiedenen Autoren verallgemeinert, insbesondere zu analogen Zerlegungssätzen für Sätze der monadischen Logik zweiter Stufe.

In der Dissertation erarbeiten wir Erweiterungen des Feferman-Vaught-Satzes für eine generische gewichtete Logik zweiter Stufe und einige ihrer Einschränkungen. Als Grundlage für unsere Logik dient die gewichtete Logik, die von Droste und Gastin zur Charakterisierung gewichteter Automaten entwickelt worden ist [DG07]. *Gewichtet* heißt für unsere Logik, dass die Auswertung einer Formel keinen Boole'schen Wahrheitswert liefert, sondern ein Gewicht aus einem Halbring, also einer Struktur mit zueinander verträglicher Addition und Multiplikation. Einfache Beispiele für Halbringe sind die natürlichen Zahlen, der *tropische* Halbring oder *Min-Plus-Halbring*, d. h. die reellen Zahlen erweitert durch $-\infty$ mit Minimum und Plus als Operationen, sowie der *Kapazitätshalbring*, d. h. die reellen Zahlen mit Minimum und Maximum als Operationen. Die möglichen Interpretation dieser Gewichte sind vielfältig, so können diese zum Beispiel als Vielfachheiten, Wahrscheinlichkeiten, Kapazitäten oder Profite aufgefasst werden. Konkreter lassen sich etwa für endliche Graphen Formeln angeben, welche die Größe der größten Clique, die Anzahl der Cliquen einer festen Größe oder die Größe des *minimalen Schnitts* des Graphen beschreiben.

Formal unterscheidet unsere Logik zwei Arten von Formeln. Die erste Art besteht aus allen Formeln der klassischen monadischen Logik zweiter Stufe. Deren Auswertung erfolgt wie üblich, das Ergebnis der Auswertung wird anschließend von wahr bzw. falsch in die Eins bzw. die Null des Halbrings umgewandelt. Zur zweiten Art von Formeln gehören zunächst alle Formeln der ersten Art und alle Elemente des Halbrings. Des Weiteren dürfen alle Formeln der zweiten Art miteinander multipliziert und addiert sowie mittels vier gewichteter Quantoren quantifiziert werden. Die Auswertung der Quantoren ist hierbei an die der klassischen Quantoren angelehnt. Während die Existenzquantoren erster und zweiter Stufe eine Disjunktion über das gesamte Universum der Struktur bzw. über dessen Potenzmenge modellieren, modellieren hier Summenquantoren erster und zweiter Stufe eine Summe über das gesamte Universum bzw. über dessen Potenzmenge. In analoger Weise modellieren Produktquantoren erster und zweiter Stufe jeweils ein Produkt über das gesamte Universum bzw. über dessen Potenzmenge.

Der klassische Feferman-Vaught-Satz unterscheidet nicht zwischen endlichen und unendlichen Strukturen. Da in Boole'scher Logik beliebige unendliche Disjunktionen und Konjunktionen wohldefiniert sind, gilt dies auch für die Auswertung von Existenz- und Allquantoren als potenziell unendliche Disjunktionen und Konjunktionen. In Halbringen hingegen sind unendliche Summen und Produkte im Allgemeinen nicht wohldefiniert, in endlichen Körpern konvergiert etwa eine unendliche Summe aus Einsen nicht. Um analog zum klassischen Satz auch Aussagen für unendliche Strukturen gewinnen zu können, betrachten wir daher auch sogenannte *bi-vollständige* Halbringe, d. h. Halbringe, in denen beliebige Summen und Produkte definiert sind. Beispiele solcher Halbringe sind die natürlichen Zahlen, erweitert um ein Element ∞, der Min-Plus-Halbring, eingeschränkt auf die nicht-negativen reellen Zahlen, sowie der Kapazitätshalbring. Für all unsere Resultate gibt es also zwei Varianten. Für endliche Strukturen gelten die Aussagen für alle Halbringe, für unendliche Strukturen für alle bi-vollständigen Halbringe. Unsere Resultate lassen sich wie folgt zusammenfassen.

- Wir zeigen einen Feferman-Vaught-Satz für disjunkte Vereinigungen von Strukturen für alle Formeln unserer gewichteten Logik, in denen kein Produktquantor zweiter Stufe vorkommt und in denen der Produktquantor erster Stufe nur Formeln quantifiziert, die selbst keinen gewichteten Quantor enthalten. Überraschenderweise ist das so definierte Fragment unserer Logik genau das, welches über Wörtern die gleiche Ausdrucksstärke wie gewichtete Automaten besitzt [DG07].
- Wir zeigen, dass die Einschränkungen auf den Produktquantoren notwendig sind. Genauer geben wir konkrete Halbringe und Formeln an und zeigen, dass für diese keine Zerlegung der Form des Feferman-Vaught-Satzes möglich ist. Die Formeln, die wir hierfür verwenden, treten bereits in [DG07] und [DR09] als Formeln auf, die nicht von gewichteten Automaten modelliert werden können. Während es in diesen Artikeln mit elementaren Argumenten möglich war, die Nicht-Erkennbarkeit der Formeln zu zeigen, ist dies in unserem Fall schwieriger. Um die Nicht-Zerlegbarkeit unserer Formeln zu zeigen, verwenden wir den Satz von Ramsey [Ra30].
- Für Produkte von Strukturen zeigen wir einen Feferman-Vaught-Satz auf dem Fragment unserer Logik, welches nur Quantoren erster Stufe und keine Produktquanto-

ren verwendet. Die Einschränkung auf die Logik erster Stufe ist für Produkte von Strukturen bereits im Fall der klassischen Logik nötig. Auch hier zeigen wir, dass sich Formeln, die einen Produktquantor enthalten, im Allgemeinen nicht zerlegen lassen.

- Wir geben Bedingungen an den Halbring an, unter denen die obigen Einschränkungen auf den Produktquantoren nicht notwendig sind. Genauer zeigen wir dies für *schwach bi-aperiodische* Halbringe und *De Morgan Algebren*.
- Wir zeigen, dass unsere Feferman-Vaught-Sätze auch für verallgemeinerte Vereinigungen und Produkte gelten, sofern diese durch *Übersetzungsschemata* (translation schemes) [MPS90] oder Courcelles *Transduktionen* [Co94] definiert sind. Übersetzungsschemata sind hierbei ein Werkzeug der Modelltheorie, welches die "Übersetzung" von Strukturen einer Signatur in Strukturen einer anderen Signatur ermöglicht. Sie können zum Beispiel verwendet werden, um zwischen *Texten* und Bäumen sowie zwischen Schachtelworten, *alternierenden Texten* und *Hecken* zu übersetzen.

Die größten Herausforderungen bei der Erarbeitung dieser Ergebnisse waren einerseits, die korrekten Einschränkungen unserer allgemeinen Logik zu ermitteln und zu zeigen, dass diese Einschränkungen auch notwendig sind. Andererseits erforderten alle Betrachtungen des Produktquantors neue Überlegungen, da es für diesen keine Entsprechung in der klassischen Logik gibt. Während der Allquantor der monadischen Logik zweiter Stufe sich mittels Negation durch den Existenzquantor darstellen lässt, ist eine Darstellung des Produktquantors durch den Summenquantor im Allgemeinen nicht möglich.

Veröffentlicht wurden die in diesem Abschnitt beschriebenen Resultate 2018 beim 43. Internationalen Symposium zu Mathematischen Grundlagen der Informatik (International Symposium on Mathematical Foundations of Computer Science, MFCS) [DP18].

3 Max-Plus-Baumautomaten

Im zweiten Teil der Dissertation erweitern wir vier Entscheidbarkeitsresultate von *Max-Plus-Wortautomaten* auf *Max-Plus-Baumautomaten*. Ein Max-Plus-Wortautomat ist ein endlicher Automat über Zeichenketten, dessen Transitionen mit reellen Zahlen oder dem Gewicht $-\infty$ gewichtet sind. Jedem Lauf des Automaten wird ein Gewicht zugeordnet, indem die Gewichte aller Transitionen des Laufs aufsummiert werden. Jedem Wort wird das Maximum der Gewichte aller Läufe auf dem Wort zugewiesen. Auf diese Weise definiert ein Max-Plus-Wortautomat eine Funktion, die Wörter auf reelle Zahlen oder $-\infty$ abbildet. Allgemeiner sind Max-Plus-Wortautomaten und die eng verwandten Min-Plus-Automaten gewichtete Automaten über dem Max-Plus-Halbring bzw. dem Min-Plus-Halbring. Erstmals eingeführt wurden Min-Plus-Automaten von Simon als Werkzeug, um die Entscheidbarkeit der *endlichen-Potenzeigenschaft* (finite power property) zu zeigen. Seit ihrer Einführung erfreuen sich Max- und Min-Plus-Automaten eines fortwährenden Forschungsinteresses und kommen in verschiedensten Kontexten zum Einsatz, so etwa

zur Bestimmung der Sternhöhe einer Sprache, zum Beweis der Terminierung bestimmter Termersetzungssysteme und zur Modellierung diskreter Ereignissysteme. Auch finden sie im Rahmen von natürlicher Sprachverarbeitung Verwendung, wo vor dem Hintergrund numerisch stabiler Berechnungen Wahrscheinlichkeiten häufig als Log-Wahrscheinlichkeiten im Min-Plus-Halbring berechnet werden.

Während sich viele Eigenschaften endlicher Automaten sehr einfach entscheiden lassen, etwa die Determinisierbarkeit, die Leerheit oder die Äquivalenz zweier Automaten, sind vergleichbare Eigenschaften bei Max-Plus-Wortautomaten schwer bis gar nicht entscheidbar. In der Dissertation betrachten wir vier Entscheidungsprobleme für Max-Plus-Automaten: das *Äquivalenzproblem*, das *Eindeutigkeitsproblem*, das *Determinisierbarkeitsproblem* und das *endliche Determinisierbarkeitsproblem*. Die ersten drei dieser Probleme betrachten wir für *endlich mehrdeutige*, das vierte für *eindeutige* Automaten. Hierbei nennen wir einen Automaten endlich mehrdeutig, falls die Anzahl der Läufe auf jeder Eingabe durch eine globale Konstante beschränkt ist. Gibt es sogar nur höchstens einen Lauf auf jeder Eingabe, nennen wir den Automaten eindeutig. Als Spezialfall eindeutiger Automaten betrachten wir *deterministische* Max-Plus-Automaten. Dies sind Automaten, in denen höchstens ein Zustand initial ist und in denen für jeden Zustand und jedes Eingabesymbol höchstens eine Transition in einen Folgezustand ein reelles Gewicht trägt. Der Grad der Mehrdeutigkeit eines Max-Plus-Automaten lässt sich in Polynomialzeit entscheiden. Des Weiteren bilden die Klassen von Funktionen, die durch deterministische, eindeutige und endlich mehrdeutige Automaten charakterisiert werden, eine strikt aufsteigende Hierarchie. Die betrachteten Probleme sind wie folgt definiert.

Beim *Äquivalenzproblem* soll entschieden werden, ob zwei gegebene Max-Plus-Automaten äquivalent sind, d. h. ob beide jeder Eingabe das gleiche Gewicht zuordnen. Dieses Problem wurde von Krob für Max-Plus-Wortautomaten als im Allgemeinen unentscheidbar bewiesen [Kr94], eingeschränkt auf endlich mehrdeutige Automaten konnten Hashiguchi et al. hingegen die Entscheidbarkeit des Problems zeigen [HIJ02]. Das *Determinisierbarkeitsproblem* fragt, ob ein gegebener Max-Plus-Automat zu einem deterministischen Max-Plus-Automat äquivalent ist. Für den allgemeinen Fall ist dieses Problem bis heute offen, für bestimmte Teilklassen von Max-Plus-Wortautomaten konnte aber die Entscheidbarkeit gezeigt werden, insbesondere für eindeutige Max-Plus-Wortautomaten durch Mohri [Mo97]. Für das *Eindeutigkeitsproblem* ist zu entscheiden, ob ein gegebener Max-Plus-Automat äquivalent zu einem eindeutigen Max-Plus-Automaten ist. Dieses Problem konnte bisher nur für endlich mehrdeutige und *polynomiell mehrdeutige* Max-Plus-Wortautomaten als entscheidbar gezeigt werden [Kl04, KL09]. Bei letzteren Automaten ist die Anzahl der Läufe auf jeder Eingabe polynomiell in der Größe der Eingabe beschränkt. Das *endliche Determinisierbarkeitsproblem* schließlich fragt, ob sich ein gegebener Max-Plus-Automat auch als endliches Maximum deterministischer Max-Plus-Automaten darstellen lässt. Die Klasse von Funktionen, die eine derartige Darstellung zulassen, liegt strikt zwischen den Klassen der von deterministischen und von endlich mehrdeutigen Automaten charakterisierten Funktionen und ist nicht vergleichbar mit der Klasse der von eindeutigen Automaten charakterisierten Funktionen. Die Entscheidbarkeit dieses Problems ist bisher nur für endlich mehrdeutige Max-Plus-Wortautomaten bekannt [Ba13].

In der Dissertation betrachten wir diese vier Probleme für Max-Plus-Baumautomaten. Während ein Max-Plus-Wortautomat eine Funktion auf den Wörtern über einem endlichen Alphabet definiert, beschreibt ein Max-Plus-Baumautomat eine Funktion auf den *Bäumen* über einem endlichen *Rangalphabet*, also einem Alphabet, in dem jedem Buchstaben eine feste Anzahl erwarteter Vorgänger zugeordnet wird. Max-Plus-Baumautomaten bilden als gewichtete Baumautomaten über dem Max-Plus-Halbring eine Verallgemeinerung sowohl von Max-Plus-Wortautomaten als auch von (ungewichteten) endlichen Baumautomaten. Anwendung finden Max-Plus-Baumautomaten etwa in Form *probabilistischer kontextfreier Grammatiken* im Rahmen der natürlich Sprachverarbeitung. Hier zeigen wir, dass die zuvor beschriebenen entscheidbaren Eigenschaften von Max-Plus-Wortautomaten auch für Max-Plus-Baumautomaten entscheidbar sind. Genauer zeigen wir, dass das Äquivalenzproblem, das Eindeutigkeitsproblem und das Determinisierbarkeitsproblem für endlich mehrdeutige Max-Plus-Baumautomaten entscheidbar sind sowie, dass das endliche Determinisierbarkeitsproblem für eindeutige Automaten entscheidbar ist.

Für das Äquivalenzproblem konstruieren wir eine Reduktion auf das Erfüllbarkeitsproblem linearer Ungleichungssysteme mit ganzzahligen Lösungen, welches entscheidbar ist. Dies entspricht dem Ansatz aus [HIJ02], statt der dort verwendeten *Kreiszerlegungen* (cycle decompositions) verwenden wir jedoch den Satz von Parikh [Pa66], was auch eine beträchtliche Vereinfachung des bestehenden Beweises darstellt. Für das Eindeutigkeitsproblem verallgemeinern wie die *Dominanzeigenschaft* (dominance property) [Kl04] von Wörtern auf Bäume und zeigen, dass ein endlich mehrdeutiger Max-Plus-Baumautomat diese genau dann erfüllt, wenn er äquivalent zu einem eindeutigen Automaten ist. Insbesondere beschreiben wir, sofern der Automat die Dominanzeigenschaft besitzt, einen konkreten Algorithmus, um einen äquivalenten eindeutigen Automaten zu konstruieren. Um die Entscheidbarkeit des Determinisierbarkeitsproblems für endlich mehrdeutige Automaten zu zeigen, kombinieren wir zunächst Resultate aus [BMV10] und [Mo97], um die Entscheidbarkeit dieses Problems für eindeutige Max-Plus-Baumautomaten zu zeigen. Die Entscheidbarkeit für endlich mehrdeutige Automaten folgt hieraus mit dem vorhergehenden Resultat. Auch dieses Resultat ist effektiv mittels der Konstruktion in [BMV10]. Für das endliche Determinisierbarkeitsproblem verallgemeinern wir die *Gabeleigenschaft* (fork property) [BK13] von Wörtern auf Bäume und zeigen, dass diese von einem Max-Plus-Baumautomaten genau dann erfüllt wird, wenn dieser sich nicht als endliches Maximum deterministischer Max-Plus-Baumautomaten darstellen lässt. Auch hier erhalten wir einen Algorithmus, der zu einem eindeutigen Max-Plus-Baumautomaten, welcher die Gabeleigenschaft nicht erfüllt, endlich viele deterministische Max-Plus-Baumautomaten konstruiert, deren Maximum äquivalent zum Ausgangsautomaten ist.

Die Resultate zu endlich mehrdeutigen Automaten wurden 2017 beim 42. Internationalen Symposium zu Mathematischen Grundlagen der Informatik (International Symposium on Mathematical Foundations of Computer Science, MFCS) veröffentlicht [Pa17a]. Das Resultat zum endlichen Determinisierbarkeitsproblem wurde 2019 beim 36. Internationalen Symposium zu Theoretischen Aspekten der Informatik (International Symposium on Theoretical Aspects of Computer Science, STACS) veröffentlicht [Pa19].

Die größte Herausforderung bei der Erarbeitung der in diesem Abschnitt beschriebenen Ergebnisse war es, die für Max-Plus-Wortautomaten existierenden Beweisansätze an die nichtlineare Struktur von Bäumen anzupassen. Grundsätzlich lassen sich die Beweiszüge aller oben beschriebenen Resultate wie folgt zusammenfassen. Zunächst wird aus dem oder den Ausgangsautomaten ein neuer Automat konstruiert und für diesen eine Bedingung formuliert. Anschließend wird gezeigt, dass diese Bedingung genau dann erfüllt ist, wenn der Ausgangsautomat die zu entscheidende Eigenschaft erfüllt sowie, dass diese Bedingung selbst entscheidbar ist. Um etwa die Determinisierbarkeit eines eindeutigen Max-Plus-Automaten zu entscheiden, kann man prüfen, ob dieser die *Zwillingseigenschaft* (twins property) erfüllt [Mo97]. Während es sich elementar zeigen lässt, dass die Zwillingseigenschaft für eindeutige Automaten entscheidbar ist und nur Automaten mit dieser Eigenschaft determinisierbar sind, erfordert es den Beweis der Korrektheit eines Determinisierungsalgorithmus für Automaten, welche die Zwillingseigenschaft erfüllen. Um die Eindeutigkeit eines endlich mehrdeutigen Max-Plus-Automaten zu entscheiden, wird dieser in endlich viele eindeutige Automaten zerlegt und deren Produktautomat auf die Dominanzeigenschaft überprüft. Auch hier kann die Entscheidbarkeit der Eigenschaft elementar gezeigt werden; dass diese Eigenschaft die Äquivalenz des Automaten zu einem eindeutigen Automaten charakterisiert, erfordert hingegen in Hin- und Rückrichtung komplexere Beweise. Für das Äquivalenzproblem wird ein Automat mit vektoriellen Gewichten konstruiert und auf eine Eigenschaft überprüft, deren einfachste bekannte Reduktion auf ein entscheidbares Problem jene auf die Erfüllbarkeit diophantischer Ungleichungssysteme ist. Zwar lassen sich die benannten Konstruktionen und Eigenschaften oft formal direkt auf Baumautomaten übertragen, sind dann aber nicht mehr notwendig oder hinreichend oder der Beweis ihrer Entscheidbarkeit wird signifikant schwerer. Dies hat die Entwicklung neuer Beweismethoden erfordert, die Potenzial für die Betrachtung weiterer Probleme bieten. Besonders hervorzuheben sind hierbei zwei Beweismethoden, die fundamental für die Entwicklung des folgenden, nicht in der Dissertation enthaltenen Resultats waren.

Als Fortführung der Untersuchung des endlichen Determinisierbarkeitsproblems für eindeutige Max-Plus-Baumautomaten konnten wir zeigen, dass dieses Problem auch für endlich mehrdeutige Max-Plus-Baumautomaten entscheidbar ist. Dies verallgemeinert einerseits das entsprechende Resultat für Max-Plus-Wortautomaten [Ba13] sowie auch unser eigenes Resultat für eindeutige Automaten, macht Letzteres aber nicht obsolet, da wir auf dessen Entscheidbarkeit reduzieren. Die erste wichtige, im Rahmen der Dissertation entwickelte Beweismethode, die wir für die Lösung dieses Problems eingesetzt haben, ist die Kombination der Anwendung des Satzes von Parikh auf Automaten mit vektoriellen Gewichten und der Reduktion auf die Erfüllbarkeit diophantischer Ungleichungssysteme. Diese ermöglicht die Entwicklung eleganter Entscheidbarkeitsaussagen für Eigenschaften gewichteter Max-Plus-Automaten, die zunächst sehr komplex scheinen. Die wichtigste Beweismethode liegt in der Weiterentwicklung der Anwendung von *Überdeckungsautomaten* in Beweisen. Ein Überdeckungsautomat entsteht allgemein als Konstruktion aus einem Ausgangsautomaten, dessen Zustände erweitert werden, um Eigenschaften des Laufes zu speichern, auf dem sich die Zustände befinden. Die wohl bekannteste Überdeckungskonstruktion ist die der *Schützenberger-Überdeckung*, welche von einer Konstruktion Schützenbergers inspiriert ist, erstmals von Sakarovitch explizit gemacht wurde und bei vielen Resultaten zu Max-Plus-Automaten zum Einsatz gekommen ist. Die

Entscheidbarkeit der endlichen Determinisierbarkeit für endlich mehrdeutige Max-Plus-Baumautomaten erforderte die Entwicklung sehr komplexer Überdeckungsautomaten, um die im Baumfall neu auftretenden Phänomene behandeln zu können.

Unser Resultat zum endlichen Determinisierbarkeitsproblems endlich mehrdeutiger Max-Plus-Baumautomaten wurde 2020 beim 47. Internationalen Kolloquium zu Automaten, Sprachen und Programmierung (International Colloquium on Automata, Languages, and Programming, ICALP) veröffentlicht und dort als einer von zwei studentischen Beiträgen ausgezeichnet [Pa20b].

4 Monitor-Logiken

Im letzten Teil der Dissertation entwickeln wir eine Logik für unendliche Wörter, welche die gleiche Ausdrucksstärke wie *quantitative Monitorautomaten* besitzt. In diesem 2016 durch Chatterjee, Henzinger und Otop eingeführten Berechnungsmodell [CHO16] kann bei jedem Zustandsübergang eines Laufs des Automaten einer von endlich vielen Zählern (monitor counter) gestartet werden. Zu Beginn mit Null initialisiert kann dieser in den Folgetransitionen inkrementiert, dekrementiert oder gestoppt werden. An jeder Transition kann höchstens ein Zähler gestartet werden, jeder Zähler muss nach endlich vielen Transitionen gestoppt werden und ein Zähler muss erst gestoppt werden, bevor er erneut gestartet werden kann. Der Zählerwert beim Stoppen eines Zählers wird dem Buchstaben des Wortes zugeordnet, an dessen Transition er gestartet wurde. Aus der hieraus entstehenden unendlichen Folge von Zählerwerten wird dem Lauf durch eine *Bewertungsfunktion* ein Gewicht zugeordnet. Das Gewicht eines unendlichen Wortes berechnet sich schließlich aus dem Infimum der Gewichte aller Läufe des Automaten auf dem Wort. Quantitative Monitorautomaten sind äquivalent zu einer Teilklasse *geschachtelter gewichteter Automaten* (nested weighted automata), einem Automatenmodell dessen Leerheits- und Universalitätsprobleme für viele Bewertungsfunktionen entscheidbar sind. Des Weiteren bilden quantitative Monitorautomaten eine Verallgemeinerung gewichteter Büchi-Automaten und ihrer Erweiterung mit Bewertungsfunktionen.

Quantitative Monitorautomaten sind sehr ausdrucksstark. Als Beispiel stelle man sich ein Warenlager vor, welches in regelmäßigen Abständen beliefert wird. Zwischen den Lieferungen kann in jedem Zeitschritt ein Artikel des Lagers durch eine Anfrage abgerufen werden. Eine fortlaufende Abfolge von Lieferungen und Anfragen kann als unendliche Folge über dem Alphabet $\{\mathsf{Lieferung}, \mathsf{Anfrage}\}$ aufgefasst werden. Von Interesse können nun etwa die minimale, die maximale oder die durchschnittliche Anzahl an Anfragen zwischen zwei Lieferungen einer solchen Folge sein. All diese Eigenschaften lassen sich mit Hilfe quantitativer Monitorautomaten beschreiben. So kann ein Automat, der als Eingabe eine Folge aus Lieferung und Anfrage erhält, bei jeder Lieferung einen Zähler starten, der die Anfragen zählt bis er bei der nächsten Lieferung wieder gestoppt wird. Eine passende Bewertungsfunktion berechnet dann die gewünschte Eigenschaft, für die durchschnittliche Anzahl an Anfragen zum Beispiel das *Cesàro-Mittel*. Ein solches Verhalten kann weder mit gewichteten Büchi-Automaten noch mit deren Erweiterung durch Bewertungsfunktionen beschrieben werden.

In der Dissertation entwickeln wir eine gewichtete Logik und zeigen, dass diese die gleiche Ausdrucksstärke wie quantitative Monitorautomaten besitzt. Im Zuge dessen beweisen wir auch verschiedene Abschlusseigenschaften quantitativer Monitorautomaten und, dass bei diesen die Büchi- und Muller-Akzeptanzbedingungen äquivalent sind. Die entwickelte Logik besitzt drei gewichtete Quantoren, welche jeweils die Zähleroperationen, die Bewertungsfunktion und die Infimumbildung modellieren. Die größte Herausforderung bei der Erarbeitung dieses Resultats war einerseits, geeignete Quantoren zu finden, und andererseits, die passenden Einschränkungen an diese zu finden. Ohne Einschränkungen an die Quantoren ist die Logik ausdrucksstärker als quantitative Monitorautomaten, was wir auch formal beweisen. Eine Neuheit ist hierbei, dass die Auswertung des Quantors für die Zähleroperationen von einer MSO-definierbaren Bedingung abhängt. Des Weiteren ist unser Resultat effektiv, d. h. für eine Formel unserer Logik zeigen wir explizit, wie sich ein quantitativer Monitorautomat konstruieren lässt, der die gleiche Funktion wie die Formel beschreibt. Andererseits zeigen wir, wie sich für jeden quantitativen Monitorautomaten eine Formel unserer Logik konstruieren lässt, die das gleiche Verhalten wie der Automat aufweist.

Das in diesem Abschnitt beschriebene Resultat wurde 2017 beim 42. Internationalen Symposium zu Mathematischen Grundlagen der Informatik (International Symposium on Mathematical Foundations of Computer Science, MFCS) veröffentlicht [Pa17b].

Literaturverzeichnis

[Ba13] Bala, Sebastian: Which finitely ambiguous automata recognize finitely sequential functions? In (Chatterjee, Krishnendu; Sgall, Jiří, Hrsg.): MFCS. Jgg. 8087 in LNCS. Springer, S. 86–97, 2013.

[BK13] Bala, Sebastian; Koniński, Artur: Unambiguous automata denoting finitely sequential functions. In (Dediu, Adrian-Horia; Martín-Vide, Carlos; Truthe, Bianca, Hrsg.): LATA. Jgg. 7810 in LNCS. Springer, S. 104–115, 2013.

[BMV10] Büchse, Matthias; May, Jonathan; Vogler, Heiko: Determinization of weighted tree automata using factorizations. JALC, 15(3/4):229–254, 2010.

[CHO16] Chatterjee, Krishnendu; Henzinger, Thomas A.; Otop, Jan: Quantitative monitor automata. In (Rival, Xavier, Hrsg.): SAS. Jgg. 9837 in LNCS. Springer, S. 23–38, 2016.

[Co94] Courcelle, Bruno: Monadic second-order definable graph transductions: A survey. Theor. Comput. Sci., 126(1):53–75, 1994.

[DG07] Droste, Manfred; Gastin, Paul: Weighted automata and weighted logics. Theor. Comput. Sci., 380(1-2):69–86, 2007.

[DP18] Droste, Manfred; Paul, Erik: A Feferman-Vaught decomposition theorem for weighted MSO logic. In (Potapov, Igor; Spirakis, Paul; Worrell, James, Hrsg.): MFCS. Jgg. 117 in LIPIcs. LZI, S. 76:1–76:15, 2018.

[DR09] Droste, Manfred; Rahonis, George: Weighted automata and weighted logics with discounting. Theor. Comput. Sci., 410(37):3481–3494, 2009.

[FV59] Feferman, Solomon; Vaught, Robert L.: The first order properties of products of algebraic systems. Fund. Math., 47(1):57–103, 1959.

[HIJ02] Hashiguchi, Kōsaburō; Ishiguro, Kenichi; Jimbo, Shūji: Decidability of the equivalence problem for finitely ambiguous finance automata. IJAC, 12(3):445–461, 2002.

[Kl04] Klimann, Ines; Lombardy, Sylvain; Mairesse, Jean; Prieur, Christophe: Deciding unambiguity and sequentiality from a finitely ambiguous max-plus automaton. Theor. Comput. Sci., 327(3):349–373, 2004.

[KL09] Kirsten, Daniel; Lombardy, Sylvain: Deciding unambiguity and sequentiality of polynomially ambiguous min-plus automata. In (Albers, Susanne; Marion, Jean-Yves, Hrsg.): STACS. Jgg. 3 in LIPIcs. LZI, S. 589–600, 2009.

[Kr94] Krob, Daniel: The equality problem for rational series with multiplicities in the tropical semiring is undecidable. IJAC, 4(3):405–426, 1994.

[Mo97] Mohri, Mehryar: Finite-state transducers in language and speech processing. Comput. Linguist., 23(2):269–311, 1997.

[MPS90] Mycielski, Jan; Pudlák, Pavel; Stern, Alan S.: A lattice of chapters of mathematics (interpretations between theorems). Mem. Amer. Math. Soc. 426. AMS, 1990.

[Pa66] Parikh, Rohit Jivanlal: On context-free languages. J. ACM, 13(4):570–581, 1966.

[Pa17a] Paul, Erik: The equivalence, unambiguity and sequentiality problems of finitely ambiguous max-plus tree automata are decidable. In (Larsen, Kim G.; Bodlaender, Hans L.; Raskin, Jean-François, Hrsg.): MFCS. Jgg. 83 in LIPIcs. LZI, S. 53:1–53:13, 2017.

[Pa17b] Paul, Erik: Monitor logics for quantitative monitor automata. In (Larsen, Kim G.; Bodlaender, Hans L.; Raskin, Jean-François, Hrsg.): MFCS. Jgg. 83 in LIPIcs. LZI, S. 14:1–14:13, 2017.

[Pa19] Paul, Erik: Finite sequentiality of unambiguous max-plus tree automata. In (Niedermeier, Rolf; Paul, Christophe, Hrsg.): STACS. Jgg. 126 in LIPIcs. LZI, S. 55:1–55:17, 2019.

[Pa20a] Paul, Erik: Expressiveness and Decidability of Weighted Automata and Weighted Logics. Dissertation, Universität Leipzig, 2020.

[Pa20b] Paul, Erik: Finite sequentiality of finitely ambiguous max-plus tree automata. In (Czumaj, Artur; Dawar, Anuj; Merelli, Emanuela, Hrsg.): ICALP. Jgg. 168 in LIPIcs. LZI, S. 137:1–137:15, 2020.

[Ra30] Ramsey, Frank P.: On a problem of formal logic. Proc. London Math. Soc., series 2, 30:264–286, 1930.

[Sc61] Schützenberger, Marcel-Paul: On the definition of a family of automata. Inform. Control, 4(2-3):245–270, 1961.

Erik Paul wurde am 24. Juni 1989 in Schkeuditz geboren. Im Jahr 2009 begann er ein Studium der Mathematik an der Universität Leipzig, welches er im Sommer 2015 mit Auszeichnung abschloss. In seiner Diplomarbeit beschäftigte er sich mit der Darstellung gewichteter Baumautomaten durch gewichtete Logiken. Die Arbeit an diesem Thema setzte er im Rahmen einer Promotion ab Herbst 2015 als Stipendiat im DFG Graduiertenkolleg “Quantitative Logiken und Automaten” (QuantLA) fort. Seit der Verteidigung seiner Dissertation im Sommer 2020 ist er wissenschaftlicher Mitarbeiter der Abteilung Automaten und Sprachen am Institut für Informatik der Universität Leipzig.

Rechnergestützte Fertigung unter Berücksichtigung der Wahrnehmung: Erweiterung der Bandbreite digital hergestellter Objekte[1]

Michal Piovarci[2]

Abstract: Haptisches und visuelles Feedback sind für die Beurteilung der Qualität und des Angebots von Objekten von wesentlicher Bedeutung. Einer der Vorteile der additiven Fertigung ist die Erstellung von Objekten mit personalisierten taktilen und visuellen Eigenschaften. Diese Personalisierung geschieht durch das Abscheiden von funktionell gradierten Werkstoffen in mikroskopischer Auflösung. Die originalgetreue Wiedergabe von Objekten aus der realen Welt auf einem 3D-Drucker ist jedoch eine Herausforderung. Viele verfügbare Werkstoffe und die Freiheit bei der Materialabscheidung machen es schwierig, den Raum druckbarer Objekte zu untersuchen. Interessanterweise haben unsere Berührungs- und Sehwahrnehmungen ähnlich wie die Fertigungshardware angeborene Einschränkungen, die durch biologische Einschränkungen vorgegeben sind. In dieser Arbeit gehen wir davon aus, dass es möglich ist, diese Einschränkungen zu nutzen, um die *wahrnehmbare* Spektrum eines 3D-Druckers zu erhöhen, indem die numerische Optimierung mit wahrnehmungsgeprägten Erkenntnissen kombiniert wird. Anstatt exakte Replikate zu optimieren, suchen wir nach sinngemäß äquivalenten Lösungen. Wir zeigen Anwendungen für das Fertigung von konformen Objekten, die Haptik von Zeichenwerkzeugen, sowie Objekten mit räumlich variierendem Glanz.

1 Einführung

Die additive Fertigung ist eine leistungsfähige Technologie, die es uns ermöglicht, den haptischen und visuellen Eindruck eines Objekts durch die räumliche Abscheidung funktionell gradierter Werkstoffe zu steuern. Dies ist von großer Bedeutung für alle, von einem Käufer, der Schuhe kauft bis zu einem Arzt, der ein Geschwulst abtastet, die ihren Tastsinn nutzen, um die Welt zu erforschen. Haptisches und visuelles Feedback vermitteln unterschiedliche Eigenschaften für jedes Objekt, das sich für unterschiedliche Anwendungen eignet. Somit besteht eine enge Verbindung zwischen der Art und Weise, wie sich ein Objekt anfühlt und wie wir es verwenden. Daher ist die Reproduktion des gewünschten haptischen und visuellen Verhaltens entscheidend, um bestehende, reale Konstrukte zu replizieren und neue zu entwerfen.

Um ein Objekt mit individuellen haptischen und visuellen Eigenschaften zu entwerfen, muss ein Designer die neuen Möglichkeiten [SR13] vollständig ausschöpfen. Die hohe Maß an Freiheitsgraden macht jedoch eine effiziente Untersuchung des Bauraumes schwierig. Daher ist es notwendig, neue Algorithmen zu schaffen, die dem Planer helfen, das Potenzial der additiven Fertigung voll auszuschöpfen. Diese Idee ähnelt der Verbesserung von Kamera- und Anzeigegeräten mit rechnergestützten Techniken, um sie an die

[1] Englischer Titel der Dissertation: "Perception-aware Computational Fabrication: Increasing The Apparent Gamut of Digital Fabrication"

[2] Università della Svizzera italiana, michal.piovarci@gmail.com

Hardware-Grenzen zu bringen [Lu10, Ma13]. Im Zusammenhang mit der Fertigung bezeichnen wir mit rechnergestützte Fertigung die Reihe von Techniken, die Berechnungen verwenden, um die gesamten Fähigkeiten von Fertigungsgeräten zu nutzen.

Computational Fabrication ist ein zielorientierter Konstruktionsprozess [Ch13] (Abb. 1). Die Eingabe ist eine Beschreibung des Zielverhaltens. Das Ergebnis der Technik stellt ein druckfähiges Design für das Fertigungsgerät dar, das ein Objekt erzeugt, das dem vorgeschriebenen Verhalten entspricht. Dieses Ziel wird durch die Optimierungsschleife Parametrisieren-Simulieren-Auswerten erreicht. Der Algorithmus wird mit einem parametrischen Design aus einem hinreichend expressiven Bauraum initialisiert. Als nächstes wird die numerische Simulation eingesetzt, um die Eigenschaften der digitalen Konstruktion abzuschätzen. Schließlich wird der Unterschied zwischen gewünschtem und simuliertem Verhalten ausgewertet, was zu einer Aktualisierung des parametrischen Modells führt. Solche Techniken wurden erfolgreich eingesetzt, um Objekte mit Zielreflexion [Ma09], elastischem Verhalten [Sc15] oder Klang [Li16] zu optimieren.

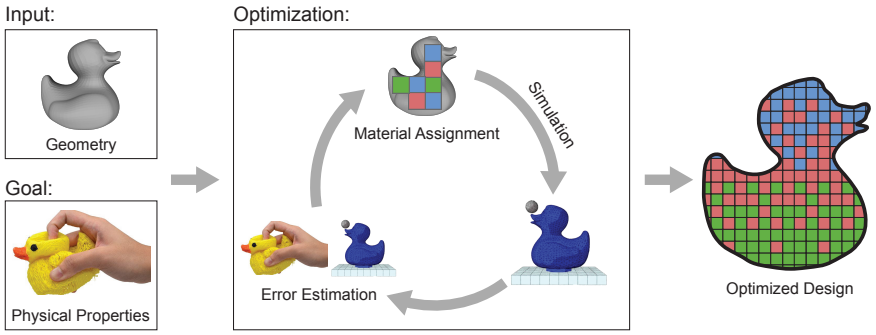


Abb. 1: Phasen der rechnergestützten Fertigung Die Eingabegeometrie wird unterteilt, und es werden Werkstoffe zugewiesen. Anhand numerischer Simulation schätzen wir die physikalischen Eigenschaften der Zuordnung. Letztendlich vergleichen wir den aktuellen Status mit den Zieleigenschaften und verwenden die Fehlerschätzung, um die Optimierung voranzutreiben.

Die direkte Anwendung der rechnergestützten Fertigung für die Reproduktion des Zielverhaltens ist eine anspruchsvolle Aufgabe. Der Optimierungsprozess hängt von einer effizienten Abschätzung physikalischer Eigenschaften durch numerische Simulation ab, um den Bauraum zu erkunden. Dies führt im Allgemeinen zu einem hochgradig nichtkonvexen und Derivat-freiem Optimierungsverfahren. Das Problem ist in unserem Umfeld noch komplizierter. Die wahrgenommenen haptischen und visuellen Empfindungen hängen von der feinskaligen Interaktion zwischen der Oberfläche des Objekts und unseren Fingern oder einfallenden Lichtstrahlen ab. Wie bei vielen komplexen Konstruktionsaufgaben wird die zugrunde liegende Physik derzeit nicht vorhersehbar entsprechend der Komplexität der physikalischen Phänomene, die Interaktion, das Maß, in dem sie auftreten, und die Grenzen der Fertigungsprozesse bestimmen, modelliert [MPC98]. Darüber hinaus beschränken die Hardware-Grenzen den Raum von Objekten, die auf einem bestimmten Drucker reproduziert werden können, oder die so genannte Spektrum des Geräts. Die Spektrum wird durch die Auswahl der verfügbaren Werkstoffe, die Notwendigkeit von Stützkonstruktionen zum Drucken von Überhängen oder die Auflösung, mit der Druck

durchgeführt werden kann, begrenzt. Zwar bringen uns Computertechniken näher an den Rand des Möglichen, doch können sie die Hardware-Einschränkungen grundsätzlich nicht überwinden und können nur innerhalb der Bandbreite eines bestimmten Geräts arbeiten. Um diese Probleme zu lösen, machen wir die interessante Beobachtung, dass das menschliche sensorielle System, ähnlich wie Hardware-Beschränkungen, seine eigenen Unvollkommenheiten hat. Zum Beispiel verarbeiten unsere Finger angewandte Reize nicht linear [Sk11], wir reagieren empfindlich auf Vibrationen in einem relativ engen Bereich [ITR06], unsere Augen haben eine begrenzte Sehschärfe [Fe96] usw.

In dieser These [Pi20] gehen wir davon aus, dass es möglich ist, die Grenzen des menschlichen sensoriellen Systems zu nutzen, um den Fertigungsprozess zu verbessern. Durch die Kombination von Berechnung mit numerischen Wahrnehmungsmodellen maskieren wir die rechnerischen und Hardware-basierten Grenzen der Fertigung und erhöhen die *wahrnehmbare* Spektrum eines 3D-Druckers auf effiziente Art und Weise. Anstatt exakte Replikate zu optimieren, suchen wir nach sinngemäß gleichwertigen Lösungen, die auf aktueller Hardware herstellbar sind. Um dieses Ziel zu erreichen, erweitern wir die standardmäßige Optimierungsschleife Parametrisieren-Simulieren-Auswerten, indem wir den Evaluierungsschritt durch eine wahrnehmbare Fehlermetrik ersetzen. Um diese Metrik wiederherzustellen, setzen wir auf die Entwicklung psychophysikalischer Experimente [Fe60], die untersuchen, wie wahrgenommene haptische Eigenschaften mit physikalischen Attributen von hergestellten digitalen Konstruktionen zusammenhängen. Solche Experimente erfordern große Mengen an Exemplaren und Teilnehmern, was die Verwendung der Experimente im Zusammenhang mit der Fertigung, in die Reize nicht digital verteilt werden können, erschwert. Wir lösen dieses Problem, indem wir neue experimentelle Konstruktionen vorschlagen, die Anzahl der Teilnehmer begrenzen. Wir unterstützen die Experimente mit numerischer Optimierung, die automatisch eine rechnergestützte Fehlermetrik generiert, die sich auf wahrgenommene Größen physikalischer Eigenschaften bezieht. Um die Fehlermetrik während der Optimierung zu bewerten, müssen wir die physikalischen Attribute neu generierter Konstruktionen schätzen. Da eine solche Simulation oft rechnerisch unlösbar ist, schlagen wir zwei Alternativen vor. Die erste Möglichkeit besteht darin, numerische Simulation mit wahrnehmungsgeprägten Erkenntnissen zu kombinieren, um eine wahrnehmungsbewusste Vergröberung durchzuführen, die den Rechenaufwand auf relevante Phänomene konzentriert. Die zweite Möglichkeit ist ein rein datengesteuertes Fertigung-in-der-Schleife-Modell, die Fertigungseinschränkungen und den Untersuchungs-Nutzung-Abgleich implizit verarbeitet. Diese beiden Ansätze ermöglichen eine effiziente numerische Abschätzung physikalischer Eigenschaften, die ansonsten schwierig zu simulieren sind und häufig zu Forschungsproblemen führen. Wir zeigen die Anwendung der vorgeschlagenen Methodik auf zwei Probleme: Die Gestaltung von Objekten mit vorgeschriebener Konformität und die Replikation des haptischen Feedbacks von traditionellen Zeicheninstrumenten. Im letzten Teil der Dissertation erweitern wir unsere Untersuchung über die Haptik hinaus – genauer gesagt auf die Erscheinungsreproduktion. Wir schlagen ein komplettes System zur Modifizierung des Oberflächenglanzes von Objekten durch räumliche Rastereffekte (Halftoning) von Lacken vor. Der Kern unseres Ansatzes ist ein Vorhersagemodell, das sowohl das Aussehen als auch die wahrgenommene Qualität von Lack-Rasterdrucken abschätzen kann. Der Schwerpunkt dieser Dissertation liegt auf einer neuen Methodik, die Wahrnehmung mit rechnerischer Fertigung verbindet, um

Objekte mit vorgeschriebenem haptischen Feedback zu entwerfen. Wir validieren dieses Konzept in vier Szenarien und zeigen Möglichkeiten für zukünftige Arbeiten auf.

2 Rechnergestützte Fertigung unter Berücksichtigung der Wahrnehmung

Ziel der haptischen Fertigung ist es, Objekte herzustellen, die gewünschte Haptik *feel* aufweisen. Um dieses Ziel zu erreichen, möchten wir komplexe wahrnehmbare Empfindungen untersuchen, die von mehreren potenziell gekoppelten physikalischen Phänomenen gesteuert werden. Stellen Sie sich vor, dass wir möchten das Gefühl des Berührens eines Lederflickens reproduzieren, was in der Automobilindustrie ein häufiges Problem darstellt [SC11]. Die Haptik von Leder wird nicht durch ein einziges Attribut bestimmt, sondern vielmehr durch mehrere Hinweise (e.g., Rauheit, Klebrigkeit, Wärme), die sich ein Gesamtperzept bilden. Wie sich die einzelnen Reize kombinieren, ist zunächst unbekannt und der Versuch, lederähnliches Material vollständig auf einem 3D-Drucker zu reproduzieren, ist technologisch anspruchsvoll, da viele der physikalischen Phänomene, die bei der haptischen Erkundung auftreten, im atomaren Maßstab auftreten. Ähnlich wie beim Standard-Ansatz Parametrisieren-Simulieren-Auswerten gehen wir dieses Problem numerisch an, indem wir eine Minimierung formulieren:

$$\arg\min_{\mathbf{p}} P(f(\mathbf{p}), T), \tag{1}$$

wobei $\mathbf{p}$ die Konstruktionsparameter für die Problemdomäne sind, $f()$ eine Funktion ist, die Konstruktionsparameter physikalischen Messungen zuordnet e.g., Schätzung der Rauheit einer prozedural generierten Textur, T das gewünschte haptische Zielverhalten ist und $P()$ eine perzeptuelle Fehlermetrik ist.

Die zentrale Herausforderung liegt in der Formulierung der Fehlermetrik. Zu diesem Zweck schlagen wir vor, einen so genannten Wahrnehmungsraum zu bauen. Ein Wahrnehmungsraum ist eine mehrdimensionale Einbettung, in der euklidische Abstand zwischen den Proben der wahrgenommenen Differenz entspricht. In dieser Arbeit schlagen wir numerische Methoden zur Wiederherstellung eines solchen Raums vor und zeigen, wie Erkenntnisse aus dem Wahrnehmungsraum verwendet werden können, um die numerische Simulation zu beschleunigen, Objekte mit vorgeschriebenem haptischem Verhalten zu entwerfen und einen hochwertigen Oberflächenglanz zu erzeugen.

3 Ein interaktionsbewusstes, perzeptuelles Modell für nicht-lineare elastische Objekte

Die Konstruktion elastischer Objekte ist sowohl rechnerisch, da zu viel Konstruktionsraums zur Verfügung steht, um das globale Optimum zu finden, als auch physisch durch die begrenzte Menge an Werkstoffen, die auf 3D-Druckern verfügbar sind, eingeschränkt. Wir zeigen, dass diese Einschränkungen durch die Kombination von Berechnung und Wahrnehmung umgangen werden können. Kern unseres Ansatzes ist die Konstruktion eines so

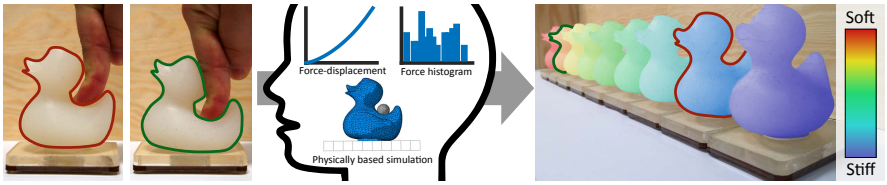


Abb. 2: Musterenten (links) mit gewünschten Elastizitätseigenschaften (z. B. gemessen) hat, berücksichtigt unser System mehrere mögliche Werkstoffe, die für die Replikation der Enten verwendet werden können (rechts), und wählt Werkstoffe aus, die am besten den Compliance-Eigenschaften entsprechen, wenn sie von einem Beobachter untersucht werden (rote und grüne Umrisse). Darüber hinaus können wir alle möglichen Werkstoffe nach ihrer wahrgenommenen Compliance sortieren, wie von unserem Modell vorhergesagt. Die Messungen werden durch Farben von steif (blau) bis weich (rot) angezeigt.

genannten Wahrnehmungsraumes, in dem die euklidische Distanz zwischen Reizen mit einer wahrgenommenen Differenz korrespondiert. Um einen solchen Raum zurückzugewinnen, schlagen wir vor, die beiden Optionen der erzwungenen Experimentkonstruktion zu verwenden. Den Teilnehmern wird ein Referenzmuster und zwei mögliche Reproduktionen präsentiert. Ihre Aufgabe ist es, die Reproduktion auszuwählen, die sich eher wie die Referenz anfühlt. Um das Ergebnis einer solchen Studie in einen Wahrnehmungsraum umzuwandeln, verwenden wir nicht-metrische multidimensionale Skalierung (NMDS). Wir sind der Meinung, dass ein eindimensionaler Raum ausreicht, um die Wahrnehmung von Compliance zu erklären. Die Achse des wiedergewonnenen Wahrnehmungsraumes ist jedoch unbekannt. Wir analysieren weiterhin mehrere mögliche Rechenmodelle und zeigen, dass die Wahrnehmung vom gewünschten Verhalten, der Objektgeometrie und hauptsächlich von der Interaktion mit dem Muster abhängt. Um diese Ergebnisse zu verallgemeinern, schlagen wir ein Rechenmodell vor, die wahrgenommene Compliance vorhersagt. Das Modell kombiniert numerische Simulation, die Form- und Werkstoffeigenschaften verarbeitet, mit einem datengesteuerten Prädiktor für die Interaktion basierend auf der lokalen Compliance. Wir zeigen, dass das vorgeschlagene Modell die Ähnlichkeit von Objekten vorhersagen kann, Werkstoffersatz anbietet, die Herstellungskosten senkt, intuitivere Benutzeroberflächen entwickelt und vor allem qualitativ hochwertigere Reproduktionen berechnet als modernste Methoden.

4 Wahrnehmungsbewusste Modellierung und Fertigung digitaler Zeichenwerkzeuge

Während die Fallstudie der Compliance-Reproduktion ermutigende Ergebnisse liefert, zeigt sie auch die Schwächen des Ansatzes auf, die Anwendung auf komplexere Probleme beschränkt. Die erste Herausforderung liegt in der psychophysikalischen Konstruktion des Experiments. Indem wir Drillingsmuster in Betracht ziehen, wird die kombinatorische Komplexität für jeden neuen Reiz, der in der Studie enthalten ist, erhöht. Darüber hinaus erfordert sogar ein eindimensionaler Wahrnehmungsraum erhebliche experimentelle Anstrengungen, um die maßgebende Achse zu erklären. Um diese Einschränkungen aus-

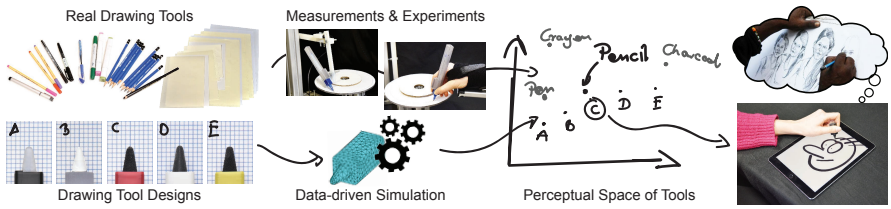


Abb. 3: Wir schlagen ein System für die Fertigung von digitalen Zeichenwerkzeugen vor, das Gefühl von echten Werkzeugen imitieren. Zu diesem Zweck messen wir Eigenschaften verschiedener echter Zeichenwerkzeuge, studieren deren Wahrnehmung und entwerfen einen wahrnehmungsbewussten Raum von Zeichenwerkzeugen. Später entwickeln wir eine Simulationstechnik, die es uns ermöglicht, neue Konstruktionen in den Raum einzubetten und die paarweise Ähnlichkeit zwischen ihnen und den Werkzeugen, die wir replizieren wollen, zu bewerten. Dies treibt den Konstruktionsprozess verschiedener digitaler Werkzeuge voran.

zuräumen, schlagen wir neuartige wahrscheinlichkeitsbasierten Rechenmodell. Das vorgeschlagene Modell ermöglicht den Einsatz von Erforschungs- und Nutzungsstrategien und ist sicherer in Bezug auf fehlende Muster, was seine Leistung für Studien mit begrenzter Teilnehmerzahl weiter verbessert. Darüber hinaus erzeugt das Modell automatisch einen Wahrnehmungsraum, der durch physikalische Messungen erklärt wird. Mit den neuen Studienkonstruktionen und Optimierungen konstruieren wir einen Wahrnehmungsraum aus Zeichenwerkzeugen, der durch wahrgenommene Reibung und wahrgenommene Schwingung erklärt wird. Um unser Modell auf die Optimierung digitaler Konstruktionen anzuwenden, verwenden wir die numerische Simulation. Aufgrund der Komplexität der maßgebenden physikalischen Phänomene und der Größenordnung, in der sie auftreten, fehlen uns ausreichend leistungsfähige Vorhersagemodelle. Um die Berechnung zu beschleunigen, schlagen wir stattdessen vor, die wahrnehmungsbewusste Vergröberung des numerischen Modells zu verwenden. Die zentrale Beobachtung ist, dass wir die haptische Interaktion nur bis zu der Auflösung simulieren müssen, die vom Mensch spürbar ist. Wir zeigen dies durch die Entwicklung eines exponentiellen Euler-Simulators, die elastodynamischen Gleichungen nur in dem Bereich berechnet, der für einen Menschen mit einem Zeichenwerkzeug wahrnehmbar ist. Wir eine Anwendung des Simulators und der Fehlermetrik bei der Konstruktion von digitalen Taststiften, die eher traditionellen Zeichenwerkzeugen als handelsüblichen Alternativen ähneln.

5 Fertigung-in-der-Schleife Co-Optimierung von Oberflächen und Taststiften für das Zeichnen von Haptik

In einer Folgearbeit validieren wir die Möglichkeit, die *wahrnehmbare* Spektrum eines 3D-Druckers zu vergrößern, indem wir anstelle direkter Nachbildungen wahrnehmbar äquivalente Lösungen herstellen. Wir bleiben im Bereich der Zeichenwerkzeuge und haben uns zum Ziel gesetzt, das Verhalten traditioneller Zeicheninstrumente mit digitalen Taststiften nachzuahmen. Dies stellt eine Herausforderung dar, da die Interaktion zwischen einem Zeicheninstrument und einem Substrat in einer geringeren Größenordnung geschieht, die kleiner als auf einem kommerziellen 3D-Drucker ist. Um diese Einschränkung

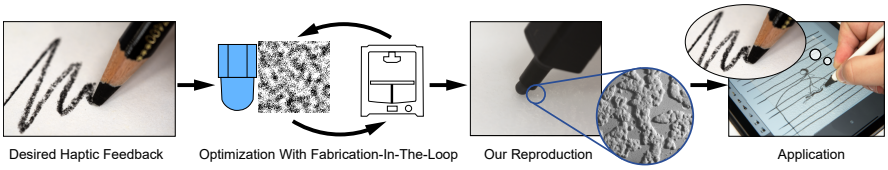


Abb. 4: Wir bieten eine datengesteuerte Methode an, um das haptische Feedback von Zeichenwerkzeugen zu imitieren. Unsere Methode verwendet das Modell der Fertigung-in-der-Schleife, das durch unser datengesteuertes stellvertretendes Modell ermöglicht wird, das automatisch Kompromisse zwischen Erforschung und Nutzung abwickelt und die Anzahl der gedruckten Muster minimiert. Die endgültigen Taststift-Oberflächen-Kombinationen können auf handelsüblicher Hardware hergestellt und direkt in aktuelle digitale Zeichenlösungen integriert werden.

zu umgehen, nutzen wir unseren Wahrnehmungsraum der Zeichenwerkzeuge. Wir beginnen damit, eine Parametrisierung des Problems zu formulieren, die aus der gemeinsamen Interaktion zwischen einem Taststift und einer Oberfläche besteht. Wir zeigen, dass eine effiziente Abschätzung der Interaktion für eine qualitativ hochwertige haptische Reproduktion entscheidend ist. Da wir ausdrücklich verlangen, das gekoppelte Verhalten abzuschätzen, können wir uns nicht auf unseren datengesteuerten Simulator verlassen. Interessanterweise beobachteten wir, dass beim Erstellen eines präziseren numerischen Modells die Geschwindigkeit der Berechnung der Fertigungszeit annäherte. Das hat uns dazu gebracht, eine Optimierung der Fertigung-in-der-Schleife zu nutzen. Die zentrale Herausforderung bei der Anwendung einer solchen Optimierung in der Praxis liegt in der effizienten Auswahl des Konstruktionsraumes, die Zeitkomplexität berücksichtigt, die bei der Bewertung der objektiven Funktion durch physikalisches Herstellen und Messen von Mustern entsteht. Zu diesem Zweck schlagen wir vor, ein probabilistisches stellvertretendes Modell zu verwenden, das uns Vertrauensbereiche für die vorhergesagte wahrgenommene Reibung und wahrgenommene Schwingung gibt. Wir verwenden diese Vertrauensschätzungen, um eine Akquisitionsfunktion zu formulieren, die den Konstruktionsraum durch Maximierung der Verbesserung jedes Musters in Richtung unserer Zielvorgabe abtastet. Wir zeigen die Vorteile unserer Methode durch die Reproduktion mehrerer herkömmlicher Zeicheninstrumente. Die erzeugten Instrumente sind vollständig 3D-gedruckt und weisen dennoch ein haptisches Feedback ähnlich wie Werkstoffe auf, die sich weit außerhalb der Druckskala befinden. Die Qualität der Reproduktionen wird in einer blinden Nutzerstudie mit Gelegenheitsnutzern und einer Befragung mit professionellen Künstlern nachgewiesen.

6 Hin zu räumlich variierender Glanzreproduktion für den 3D-Druck

Im letzten Szenario wenden wir die Methodik auf den schnell wachsenden Bereich der Erscheinungsreproduktion an. Die Erscheinungsreproduktion ist ein neuer Bereich, und dementsprechend gibt es noch keine standardisierte Methode, um das Erscheinungsbild der Objekte vollständig zu reproduzieren. Die größte Herausforderung besteht darin, dass die aktuellen Fertigungstechniken zur Modifizierung der Oberflächenreflexion (Glanz) auf die Modifizierung der Oberflächenmikrogeometrie oder die Abscheidung von Werkstoffen mit unterschiedlichen Erscheinungseigenschaften angewiesen sind. Dies führt zu ei-

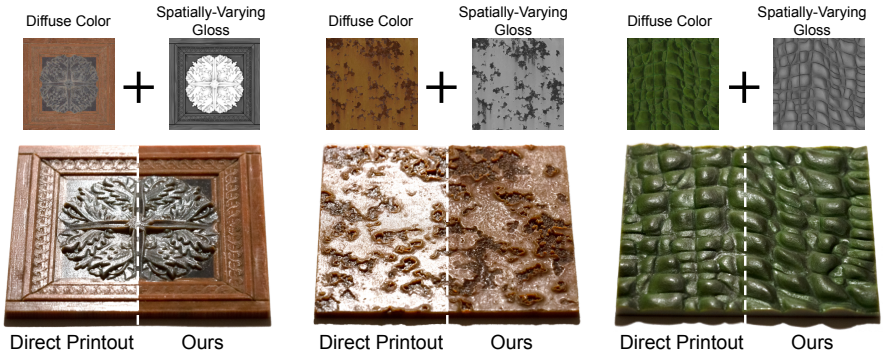


Abb. 5: Die Eingabe in unser System ist eine diffuse Farbe und räumlich variierenden Glanz. Zuerst reproduzieren wir die Farbe mit handelsüblichen Tintenstrahldruckern (linke Hälften). Als Nächstes verwenden wir unseren benutzerdefinierten Drucker als Nachbearbeitungsschritt, um Lacke zu sprühen, der Eingangsreflexion entsprechen (rechte Hälften).

ner Kopplung von Objektglanz mit diffuser (Untergrund-)Farbe. Um diesen Einfluss zu entkoppeln, schlagen wir vor, Objekte zu erstellen, bei denen die Farbinformation separat mit hoher Auflösung gedruckt und der Glanz in einem Nachbearbeitungsschritt durch räumliche Abscheidung von Lacken verändert wird. Die Lacke sind als transluzent formuliert und modifizieren den Oberflächenglanz nur durch diffusive Partikel. Leider erhöhen diese Partikel die Viskosität der Lacke so, dass sie nicht mit handelsüblichen Tintenstrahldruckern gespritzt werden können. Um die Produktion mit solch anspruchsvollen Werkstoffen zu ermöglichen, schlagen wir ein kundenspezifisches Druckgerät vor, das hochviskose Werkstoffe spritzen kann. Wir bestimmen das vorgeschlagene Gerät, indem wir ein Kalibrierverfahren beschreiben. Als Nächstes quantifizieren wir den durch erzielbaren Glanzumfang mithilfe unserer optimierten Parametern und wählen wir drei Lackgrundierungen als Basismaterialien für die Halbtonierung aus. Wir unterstützen das Druckgerät mit einem Vorhersagemodell, das auf der Simplex-Interpolation basiert und das Aussehen eines Dithered-Lackgemisches abschätzen kann. Um die Sichtbarkeit von Dithering-Artefakten zu minimieren, schlagen wir Qualitätsparameter vor, die Eigenschaften des menschlichen visuellen Systems und der physikalischen Vermischung von Lacken auf der Objektoberfläche imitieren. Wir zeigen die Möglichkeiten unseres Setups, indem wir mehrere 2D- und 3D- Beispiele mit räumlich variierender Farbe und Glanz fertigen.

7 Schlussfolgerung

In dieser Dissertation schlagen wir vor, die Grenzen des menschlichen sensorischen Systems zu nutzen, um die rechnerischen und Hardware-Grenzen der Fertigung zu überwinden und die *wahrnehmbare* Spektrum eines 3D-Druckers effektiv zu steigern. Wir zeigen die Machbarkeit der Idee auf und stellen numerische Verfahren vor, um die menschliche Wahrnehmung in die Gestaltung von Objekten mit gewünschten haptischen und visuellen Eigenschaften einzubinden. Unser Entwurf und die Bewertung von Wahrnehmungsstudien bieten ein robustes Verfahren zur Abschätzung der physikalischen Phänomene, die wahr-

genommenen Wechselwirkungen steuern. Die Anwendung von Wahrnehmungsräumen für die Fertigung hängt von der Fähigkeit ab, die physikalischen Eigenschaften digitaler Konstruktionen effizient abzuschätzen. Zu diesem Zweck schlagen wir vor, die wahrnehmungsbewusste Vergröberung des numerischen Modells und eine vollständig datengesteuerte Optimierung der Fertigung-in-der-Schleife zu verwenden. Wir zeigen, dass diese Verbesserungen es uns ermöglichen, qualitativ hochwertigere Nachbildungen von Objekten mit vorgeschriebener Verformung zu entwerfen und das haptische Feedback herkömmlicher Zeichenwerkzeuge nachzuahmen. Schließlich untersuchen wir eine Möglichkeit der zukünftigen Arbeit, indem wir wahrnehmungsbasierte Techniken auf die Reproduktion von Erscheinungsbildern anwenden. Wir stellen einen kompletten Arbeitsablauf für die besonders sorgfältige Bearbeitung des Glanzes eines Objekts vor, um die Sichtbarkeit von Dithering-Artefakten, die bei nebeneinander angeordneten Werkstoffen auftreten, zu minimieren.

Wir sind sicher, dass die in dieser These dargestellte wegweisende Arbeit zur Verbindung von Fertigung und Wahrnehmung als Eckpfeiler für zukünftige haptische und die Reproduktion des Erscheinungsbildes betreffende Arbeitsabläufe dienen wird. Ein spannender Weg für zukünftige Arbeiten ist die Integration unseres Compliance-Modells in unsere Arbeit an der Reproduktion von Zeichenwerkzeugen. Ein solches System wäre in der Lage, die drei primären Hinweise zu replizieren, die haptische Wahrnehmung von Objekten regeln, i.e., Reibung, Schwingung und Compliance [Ho00]. Die Anwendungen eines solchen Modells gehen über die Herstellung von Zeichenwerkzeugen hinaus. Zum Beispiel kann die Erstellung von originalgetreuen Nachbildungen der inneren Organe von Menschen Chirurgen dabei unterstützen, komplexe Operationen frühzeitig zu planen und zu proben. Darüber können durch die Fähigkeit, sowohl das Aussehen als auch die Haptik von Körperteilen zu imitieren, realistischere Prothesen hergestellt werden. Die in dieser Arbeit vorgestellten Techniken beruhen auf aktuellen Fertigungsprozessen. Dies erweitert die Auswirkungen unserer Forschungsarbeit und ermöglicht die Einbindung in bestehende Fertigungsanlagen. Durch die Nutzung der Wahrnehmungserkenntnisse können Rapid Prototyping-Geräte reale Artefakte herstellen, die nicht von ihren herkömmlichen Gegenstücken zu unterscheiden sind. Dies erhöht die Freiheit der Designer und ermöglicht die Schaffung von Objekten, die mit anderen Techniken nur schwer oder gar nicht gefertigt werden konnten und deren Kosten sich an die Massenproduktion annähern.

Literaturverzeichnis

[Ch13] Chen, Desai; Levin, David I. W.; Didyk, Piotr; Sitthi-Amorn, Pitchaya; Matusik, Wojciech: Spec2Fab: A reducer-tuner model for translating specifications to 3D prints. ACM Transactions on Graphics, 32(4):135:1–135:10, 2013.

[Fe60] Fechner, Gustav Theodor: Elemente der Psychophysik. Elemente der Psychophysik v. 1. Breitkopf und Härtel, 1860.

[Fe96] Ferwerda, James A; Pattanaik, Sumanta N; Shirley, Peter; Greenberg, Donald P: A model of visual adaptation for realistic image synthesis. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. S. 249–258, 1996.

[Ho00] Hollins, Mark; Bensmaïa, Sliman; Karlof, Kristie; Young, Forrest: Individual differences in perceptual space for tactile textures: Evidence from multidimensional scaling. Perception & Psychophysics, 62(8):1534–1544, 2000.

[ITR06] Israr, Ali; Tan, Hong Z.; Reed, Charlotte M.: Frequency and amplitude discrimination along the kinesthetic-cutaneous continuum in the presence of masking stimuli. The Journal of the Acoustical Society of America, 120(5):2789–2800, 2006.

[Li16] Li, Dingzeyu; Levin, David I.W.; Matusik, Wojciech; Zheng, Changxi: Acoustic Voxels: Computational Optimization of Modular Acoustic Filters. ACM Transactions on Graphics, 35(4), 2016.

[Lu10] Lukac, Rastislav: Computational Photography: Methods and Applications. Harry N Abrams, USA, 1st. Auflage, 2010.

[Ma09] Matusik, Wojciech; Ajdin, Boris; Gu, Jinwei; Lawrence, Jason; Lensch, Hendrik P. A.; Pellacini, Fabio; Rusinkiewicz, Szymon: Printing spatially-varying reflectance. ACM Transactions on Graphics, 28(5):128:1–128:9, 2009.

[Ma13] Masia, Belen; Wetzstein, Gordon; Didyk, Piotr; Gutierrez, Diego: A survey on computational displays: Pushing the boundaries of optics, computation, and perception. Computers & Graphics, 37(8):1012 – 1038, 2013.

[MPC98] Myshkin, Nikolai K.; Petrokovets, M. I.; Chizhik, Sergei A.: Simulation of real contact in tribology. Tribology International, 31(1):79 – 86, 1998.

[Pi20] Piovarci, Michal: Perception-aware computational fabrication : increasing the apparent gamut of digital fabrication. dissertation, Università della Svizzera italiana, 2020.

[SC11] Stoll, Andrea; Cavalcante, Macio: Stick-Slip Characteristics of Leather/Artificial Leather. Automotive Buzz, Squeak and Rattle: Mechanisms, Analysis, Evaluation and Prevention, S. 63, 2011.

[Sc15] Schumacher, Christian; Bickel, Bernd; Rys, Jan; Marschner, Steve; Daraio, Chiara; Gross, Markus: Microstructures to Control Elasticity in 3D Printing. ACM Transactions on Graphics, 34(4), Juli 2015.

[Sk11] Skedung, Lisa; Danerlöv, Katrin; Olofsson, Ulf; Johannesson, Carl Michael; Aikala, Maiju; Kettle, John; Arvidsson, Martin; Berglund, Birgitta; Rutland, Mark W.: Tactile perception: Finger friction, surface roughness and perceived coarseness. Tribology International, 44(5):505 – 512, 2011. Special Issue: ECOTRIB 2009.

[SR13] Schmidt, Ryan; Ratto, Matt: Design-to-Fabricate: Maker Hardware Requires Maker Software. IEEE Computer Graphics and Applications, 33(6):26–34, Nov 2013.

Michal Piovarci geboren in Bratislava, Slowakei, wo er 2014 den Master in Informatik mit Auszeichnung an der Comenius-Universität erwarb. Er begann seine Promotion 2015 mit dem Titel für Informatik and Cluster of Excellence on Multimodal Computing and Interaction am Max-Plank Institut unter der Leitung von Piotr Didyk. 2018 wechselte er zusammen mit seinem Berater zur Università della Svizzera italiana, wo er seine Dissertation 2020 erfolgreich vertrat. Während seines Studiums wurde er zu Forschungsaufenthalten bei MIT CSAIL und Adobe Research eingeladen. Das Ergebnis der Dissertation besteht aus drei ACM SIGGRAPH und einer ACM SIGGRAPH Asia Publikationen, der Erscheinungsort in diesem Bereich sind. Derzeit ist er Postdoktorand am IST Austria unter der Leitung von Bernd Bickel, wo er sich mit rechnergestützter Fertigung, Erscheinungsreproduktion und maschinellem Lernen beschäftigt.

Transparenz öffentlicher Einkaufsdaten in Deutschland

Britta Reuter[1]

Abstract: Die hier vorgestellte Forschung ermöglicht die gesellschaftlich-technologische Auseinandersetzung mit einer größeren Transparenz öffentlicher Einkaufsdaten in Deutschland. Die Annahme ist, dass unter Zuhilfenahme von Technologie Steuermittel effizienter eingesetzt werden könnten und Korruption vorgebeugt werden kann. Ausgangspunkt sind eine umfassende Analyse der (inter)nationalen Forschungsergebnisse rund um die Öffnung öffentlicher Verwaltungs- und Einkaufsdaten, eine empirische Experten-Online-Befragung sowie die Recherche internationaler Best Practices. Auf dieser Grundlage werden die Chancen und Limitationen der Transparenz des öffentlichen Einkaufs für die verschiedenen Akteure aus Verwaltung, Politik, Wirtschaft, Wissenschaft, Medien und Nichtregierungsorganisationen beschrieben. Hieraus werden schließlich 15 Handlungsfelder abgeleitet, mit denen die Mehrwerte einer Öffnung gehoben und ihre Risiken adressiert werden können

1 Relevanz des Themas

Öffentliche Haushalts- und Finanzdiskussionen, nicht zuletzt auch jene um den künftigen Ausgleich der Milliarden-Investitionen zur Bekämpfung der aktuellen Corona-Pandemie, belegen eindrücklich, dass Bürgerinnen und Bürger von Staat und Verwaltung einen transparenten und nachvollziehbaren Einsatz der von ihnen erwirtschafteten Steuermittel verlangen. Während in den letzten zehn Jahren die Offenlegung einer Vielzahl von Verwaltungsdaten erfolgte, ist auffällig, dass insbesondere Daten zu Vergaben und Verträgen, auch auf dem zentralen Open Data Portal Deutschlands , kaum auffindbar sind. Vor dem Hintergrund, dass der öffentliche Einkauf in Europa im Durchschnitt etwa 15% des jährlichen Bruttoinlandsprodukts und damit etwa 400-500 Milliarden Euro für Deutschland beträgt, ist dies nicht nur überraschend, sondern im Jahr 2019 auch unzureichend. Ein erhebliches Sparpotential in diesem Bereich (bei nur 5% entsprächen dies 20-25 Milliarden pro Jahr) bleibt so unangetastet [Re21].

Der Unterschwellenbereich des öffentlichen Einkaufs, auf den diese Arbeit referenziert, betrifft dabei die nationalen Vergabeentscheidungen Deutschlands unterhalb festgelegter Schwellenwerte. Etwa 87% aller Vergaben nach Anzahl und 64% aller Vergaben nach Volumen entfallen auf den Unterschwellenbereich, was die Bedeutung dieses Bereichs nochmal verdeutlicht [Ba14]. Im Vergleich zum Oberschwellenbereich ist dieser jedoch mit

[1] Zeppelin Universität, Am Seemooser Horn 20, 88045 Friedrichshafen, b.reuter@zeppelin-university.net
https://www.govdata.de

CC BY SA

seinen Berichtspflichten sehr intransparent. So gibt es, trotz der im Oktober 2020 eingeführten Vergabestatistik, keine einheitlichen Mechanismen zur Erhebung und Auswertung öffentlicher Einkaufsdaten für die Bundesrepublik Deutschland insgesamt, geschweige denn für die Verteilungen der öffentlichen Vergaben nach Anzahl und Volumina auf den Unterschwellenbereich, getrennt nach Verfahrensarten oder föderalen Ebenen [SE16, OE17].

Der Einsatz von Technologie kann hier einen Beitrag zur Transparenz im öffentlichen Einkauf mit positiven Effekten für die Gesellschaft leisten.

2 Methodisches Vorgehen und wesentliche Ergebnisse

Ausgehend vom Status quo des öffentlichen Einkaufs, will diese Arbeit erstens klären, ob, warum und für wen eine Öffnung der öffentlichen Einkaufsdaten im Unterschwellenbereich sinnvoll sein kann. Zweitens gilt es darzulegen, wie dies praktisch und technisch erfolgen kann, wo mit Limitationen zu rechnen ist und welche Best Practices es gibt. Abschließend wird diskutiert, wie das Leitbild einer offenen Vergabepolitik mit konkreten Handlungsempfehlungen aussehen kann.

Für ein geeignetes Vorgehen ist zu beachten, dass die Öffnung des öffentlichen Einkaufs viele unterschiedliche Fragen politischer, organisatorischer, rechtlicher, technischer und gesellschaftlicher Natur tangiert. Die Breite des Themas darf dabei die nötige Tiefe nicht vernachlässigen. Um dieser Vielfalt gerecht zu werden, wurde als theoretisches Fundament die partizipative Technikfolgenabschätzung eingesetzt [Gr10, Si13]. Entlang der hier verwendeten Wirkungsdimensionen Strategie, Organisation, Recht, Technologie, Transparenz, Partizipation und Kollaboration lässt sie eine interdisziplinäre Analyse zu.

Darüber hinaus wurde mit einem Methodenmix gearbeitet, um neben qualitativen Erkenntnissen auch quantitative Schlussfolgerungen gewinnen zu können.

Herangezogen wurde die deutsche und internationale Forschungsliteratur. Die überwiegend qualitativen Ergebnisse betonten die Notwendigkeit quantitativ-empirischer Erhebungen.

Somit folgte eine Portalanalyse. Diese umfasste bereits veröffentlichte öffentliche Einkaufsdaten von 2018 und 2019 auf dem GovData-Portal des Bundes, vier Länderportalen (Hamburg, Bremen, Nordrhein-Westfalen und Rheinland-Pfalz) sowie weiteren 26 Vergabe- und Bekanntmachungsportalen in Deutschland. Für die Analyse wurde der Gesamtprozess des öffentlichen Einkaufs und ein eigens aggregiertes Datenschema aus 31 Datenpunkten basierend auf Impulsen der Open Contracting Partnership [Op19b] und des Projekts Digiwhist [MF18] gewählt. Ein wesentliches Ergebnis war, dass bereits heute einige öffentliche Einkaufsdaten zur Verfügung stehen, allerdings über diverse Portale verteilt, selten in Form von Verträgen und nie über den Gesamtprozess des öffentlichen Einkaufs.

Um die Forschungsfragen empirisch zu überprüfen, wurde Ende 2018 eine Online-Befragung über insgesamt 40 Fragen an 161 Führungskräfte und Experten aus dem Umfeld des

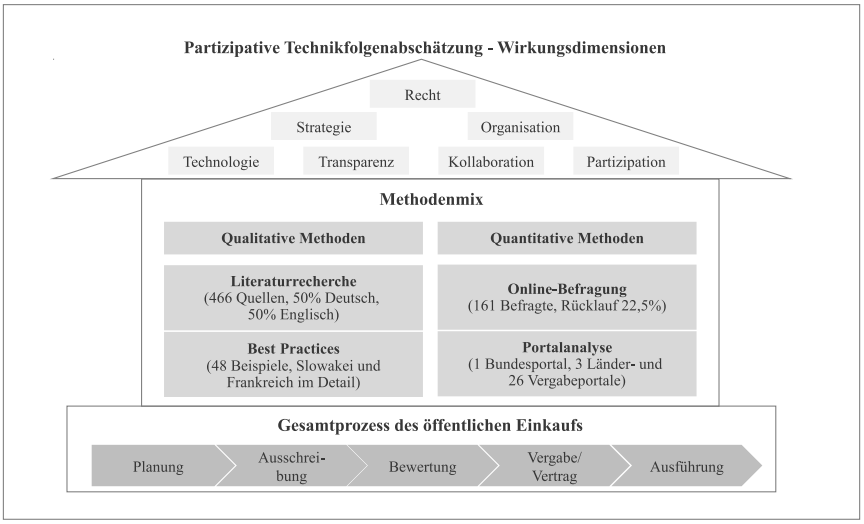


Abb. 1: Wesentliche Strukturelemente der Arbeit [Re21]

öffentlichen Einkaufs verschickt (Beteiligungsquote: 22,5%). Insgesamt befürworteten die Teilnehmer eine Erhöhung der Transparenz öffentlicher Einkaufsinformationen entlang des Gesamtprozesses, wiesen aber auch auf den Differenzierungsbedarf hin (zum Beispiel mit Blick auf bestimmte Warengruppen und Projekte).

Best Practices beinhalten die Zusammenstellung von 48 Erfolgsmethoden und eine ausführliche Erörterung der Länderbeispiele für die Slowakei und Frankreich, die sehr vielschichtige Maßnahmen zur Öffnung des öffentlichen Einkaufs ergriffen haben. Sie bestätigen, dass eine Vielzahl realer Anwendungen auf Deutschland übertragbar wäre. Beispielhafte Resultate einer Öffnung des öffentlichen Einkaufs waren unter anderem:

- **Einführung eines zentralen Vertragsregisters in der Slowakei** in 2011 mit dem Ziel von mehr Transparenz zur Reduktion von Kosten und Korruption: Zwischen 2011 und 2019 wurden fast zwei Millionen Verträge zentral veröffentlicht . Die Anzahl der Bieter erhöhte sich im Schnitt von 1,6 auf 3,7 pro Ausschreibung, Datenjournalismus und Geschäftsentwicklung stiegen deutlich an [Š15]. Die wahrgenommene Korruption konnte sich im Corruption Perceptions Index zwischen 2011 und 2017 um 12 Plätze verbessern, eine der deutlichsten Verbesserungen weltweit [Tr17].
- **Einführung des Einkaufsportals *ProZorro* in der Ukraine** in 2015 mit dem Ziel der Öffnung des öffentlichen Auftragswesens : Bei einem öffentlichen Einkaufsvolumen

https://www.crz.gov.sk/
https://prozorro.gov.ua/en

von ca. 10 Milliarden Euro betrugen die jährlichen Einsparungen ca. 10%, die Anzahl der Gebote konnte um 15% und die der Lieferanten um 45% erhöht werden [FGM17].

Die eingesetzten Methoden zeigen die Stärken und Chancen sowie die Befürwortung der Transparenz im Unterschwellenbereich auf. Gleichwohl dürfen sie nicht über Schwächen und Risiken hinwegtäuschen (zum Beispiel die Berücksichtigung von Betriebs- und Geschäftsgeheimnissen, Datenschutz): Das sinnvolle Maß einer Öffnung muss differenziert betrachtet werden.

Um sich dem anzunähern, wurden die vorliegenden Ergebnisse zunächst in ein Leitbild überführt, bevor Handlungsempfehlungen beschrieben werden. Als Leitbild soll gelten, dass im Jahr 2030 öffentliche Einkaufsdaten in Deutschland zentral über das GovData-Portal als eigener urbaner Datenraum (*Public Procurement Data Space*) in Form von Dokumenten und maschinenlesbar angeboten werden. Zudem sind die Kombination öffentlicher Einkaufsdaten mit anderen Daten als Linked Open Data [GvL12] sowie Echtzeitauswertungen über fehlerhafte Verfahren oder Preise im Verhältnis zur Lieferantenperformance möglich. Ergänzt wird das Angebot öffentlicher Einkaufsdaten in festgelegten Bereichen durch smarte Verträge. Es entstehen neue Geschäftsmodelle und Unternehmen können einfacher an Ausschreibungen teilnehmen. Die erzielten Einsparungen erleichtern nachweislich die Investition von Steuermitteln in anderen Bereichen.

3 Handlungsempfehlungen entlang der Wirkungsdimensionen

3.1 Strategie

Handlungsfeld 1: Verankerung der Öffnung öffentlicher Einkaufsdaten in der Politik
Die Öffnung öffentlicher Einkaufsdaten sollte in den politischen Programmen Deutschlands fest verankert werden, zum Beispiel in der Fortführung der *Digitalen Verwaltung* oder in der *Nationalen E-Government Strategie* (NEGS). Der Forderung nach Öffnung des öffentlichen Einkaufs sollte spätestens mit dem dritten Nationalen Aktionsplan im Jahr 2021 nachgekommen werden – umso mehr, da Deutschland nun auch Teil des Lenkungsausschusses ist und damit eine Vorbildfunktion einnehmen sollte. Damit einhergehend sollte Deutschland ebenfalls Mitglied in der Open Contracting Partnership werden [Op19c].

Handlungsfeld 2: Aufsetzen eines interdisziplinären Projekts
Darüber hinaus sollte ein interdisziplinär besetztes Projekt mit Teilnehmern aus Wirtschaft, Verwaltung, Wissenschaft, Nichtregierungsorganisationen, Politik, Medien und Vertretern der Zivilgesellschaft sowie der föderalen Ebenen (Bund, Länder, Kommunen) aufgesetzt werden, um möglichst frühzeitig eine umfassende Perspektive auf das Thema zu gewinnen. Da die stark fragmentierte kommunale Ebene einen Anteil von bis zu 40–60% am gesamten Beschaffungsvolumen trägt [BS13], sollte diese besonders berücksichtigt werden. Dieses Projekt sollte von einer gemeinsamen Stelle mit hoher Akzeptanz koordiniert werden.

3.2 Organisation

Handlungsfeld 3: Besetzung des Projektes mit verschiedenen Akteuren
Für den Projekterfolg ist es entscheidend, dass unterschiedliche Organisationen des öffentlichen Einkaufs in dem zuvor skizzierten gemeinsamen Projekt vertreten sind. Mit Blick auf die föderale Verankerung könnten dies zum Beispiel jene Bundes- und Landesbehörden sein, die bereits als Einkaufskooperationen organisiert sind, Vergabestellen unterschiedlicher Größen, Vertreter von GovData als zentralem Datenportal, der bestehenden Transparenzportale (zum Beispiel Hamburg, Bremen) oder der Vorreiter-Kommunen (zum Beispiel Köln, Bonn). Weiterhin denkbar sind Teilnehmer statistischer Ämter oder der Rechnungshöfe. Aus der Wirtschaft sollten sowohl Anbieter und Betreiber von Vergabe- und Bekanntmachungsportalen als auch Verbände und Unternehmen in Betracht kommen. Ebenso relevant erscheinen Vertreter querschnittlicher (Nichtregierungs-)Organisationen oder Projekte, die sich bereits konkret mit Einkaufsdaten auseinandergesetzt haben (z.B. *Digiwhist* [MF18]).

Handlungsfeld 4: Definition und Nutzung von Standards
Der Open Contracting Data Standard (Open Contracting Partnership, [Op19b]) ist aktuell der einzige internationale Standard. Er umfasst ca. 360 Datenfelder und sollte bezüglich eines Musterdatenschemas sowie Anwendbarkeit in Deutschland überprüft und an die hiesigen Strukturen (u.a. an Vergaberecht, Vergabeverordnungen, Prozesse, Systeme, Datenfelder) angepasst werden [vL17]. Dies sollte in der Bestandsanalyse erfolgen (Handlungsfeld 10).

Handlungsfeld 5: Ausarbeitung von Geschäftsmodellen und User Stories
Als Teil der gemeinsamen Arbeit sollten Geschäftsmodelle (vorliegend aufgezeigt unter Nutzung des Business Model Canvas in Kombination mit den Geschäftsmodellarchetypen für die Wiederverwendung von öffentlichen Informationen [FO13, OP09], praktische Beispiele und Anwendererzählungen (sogenannte *User Stories*) entwickelt werden, die die quantitativen und qualitativen Mehrwerte benennen. Gleichzeitig werden auf diese Weise die Bedarfe der einzelnen Akteure noch einmal konkret adressiert. Die User Stories sollten von einer Kosten-Nutzenschätzung begleitet werden, die sich im Projektverlauf konkretisiert.

3.3 Recht

Handlungsfeld 6: Ausweitung der Berichtspflichten auf den Unterschwellenbereich
Aktuell gibt es nur wenige Veröffentlichungspflichten, aber auch wenige Veröffentlichungsverbote für den Unterschwellenbereich. Dies lässt also einen Handlungsspielraum zu, den man für die Öffnung des öffentlichen Einkaufs nutzen sollte, wie die Städte Hamburg und Bremen bereits zeigen. Die Einschränkungen orientieren sich im Wesentlichen an anderen Rechtsgebieten wie dem Datenschutz oder den Informationsfreiheitsgesetzen. Für das Informationsfreiheitsgesetz Bund ist es so, dass öffentliche Einkaufsdaten bereits durch den Begriff der *amtlichen Informationen* abgedeckt sind. Demnach müssen öffentliche

Einkaufsdaten gar nicht immer explizit erwähnt sein, um sie anfordern zu können [Op19a]. Die Ausweitung der verpflichtenden eVergabe auf den Baubereich, eine insgesamt breitere Veröffentlichungspflicht für den Unterschwellenbereich mit einer festzulegenden Mindestschwelle in Euro sollten im Vergaberecht, aber auch in der Vergabestatistikverordnung verankert werden.

Handlungsfeld 7: Definition öffentlicher Einkaufsdaten als *High Value Data Set*
Es gilt vor dem Hintergrund der neuen Richtlinie über die Weiterverwendung von Informationen des öffentlichen Sektors (EU 2019/1024) die Definition öffentlicher Einkaufsdaten als *High Value Data Set* unter *Statistik* in der nationalen Umsetzung der neuen Richtlinie zu berücksichtigen. Um den Interpretationsspielraum einzugrenzen, ist es sinnvoll, Begriffe wie *Betriebs- und Geschäftsgeheimnisse* zu präzisieren, so wie dies bereits partiell eingefordert und mit ersten Ansätzen hinterlegt wurde [Op19a].

3.4 Technologie

Handlungsfeld 8: Zentrale Bereitstellung öffentlicher Einkaufsdaten auf *GovData*
Technologisch gibt es mit dem GovData-Portal auf Bundesebene bereits eine vorhandene Infrastruktur. Die Umfrage förderte den klaren Wunsch der Befragten nach einem zentralen Angebot der öffentlichen Einkaufsdaten zutage. Der Prozess der Bereitstellung müsste im Rahmen des Projektes erarbeitet werden. Die Angebote zur Interaktion, also zur Partizipation und Kollaboration, sind beim GovData-Portal insgesamt noch ausbaufähig. Darüber hinaus sollten perspektivisch Technologien wie Blockchain und künstliche Intelligenz berücksichtigt werden. Ein Beispiel hierfür ist Kanada, das mithilfe von Blockchain-Technologie die Registrierung der Lieferanten im öffentlichen Einkauf vorantreiben möchte [Se18].

Handlungsfeld 9: Verschränkung technologie- und anwenderorientierter Perspektiven
Als Mitwirkende sind einerseits die technischen Rollen relevant wie zum Beispiel Architekten, Solution Consultants, Datenanalysten und Entwickler. Andererseits ist aber auch die Integration der Nutzerperspektive durch die Anwender der Daten essentiell, da eine Technologie nur dann ihre Stärke entfaltet, wenn sie tatsächlich genutzt wird. Die Nutzer sollten unterschiedlichen Gruppen entstammen, damit ein möglichst breites Feedback berücksichtigt werden kann.

Handlungsfeld 10: Entwicklung eines Musterdatenschemas
In einer Bestandsaufnahme müssen bestehende Systeme, verfügbare Dokumente und Daten, ihre Unterschiede und die aktuelle Datenqualität im Vordergrund stehen. Die realen Daten müssen mit dem unter Handlungsfeld 4 festgelegten Standard abgeglichen werden [vL17]. Gegebenenfalls machen sie Anpassungen des Standards für Deutschland erforderlich. Die Ergebnisse könnten dann als Musterdatenschema in das Projekt zur Vergabestatistik genauso einfließen wie in die Öffnung der öffentlichen Einkaufsdaten über GovData.

Handlungsfeld 11: Erarbeitung eines Sollkonzepts inklusive Datenqualität
Das Sollkonzept beinhaltet neben dem Musterdatenschema künftige Prozesse der Datenbereitstellung, der -plausibilisierung und -bereinigung sowie Auswertungs- und Visualisierungsmöglichkeiten. Datenqualität herzustellen und aufrechtzuerhalten ist essentiell für eine belastbare Transparenz. Ein klarer Prozess mit Verantwortlichen sowie eine regelmäßige Kontrolle können hier unterstützen. Darüber hinaus sind Innovationsthemen wie Linked Open Data und künstliche Intelligenz für den öffentlichen Einkauf zu bewerten [MF18, vL17]. Gegebenenfalls kann ein Pilot entwickelt werden, der erste Anwendungsmöglichkeiten aufzeigt und den Nutzern hilft, konstruktives Feedback zu geben.

3.5 Transparenz

Handlungsfeld 12: Reflektion von Transparenz zur Fehlerkultur und Problemlösung
Die Ergebnisse der Online-Befragung stützen die Erkenntnisse aus der Literatur , dass eine der größten Stärken der Öffnung öffentlicher Einkaufsdaten die Erhöhung der Transparenz ist (92% Zustimmung), unmittelbar gefolgt von einer Verminderung der Korruption (86% Zustimmung), primär in den Teilprozessen der Ausschreibung/Bekanntmachung, Vergabe und Kontrolle. Als größte Chance wird die Korruptionsprävention mit 83% Zustimmung angesehen. Wichtig sei außerdem, anstelle von "bashing" oder "naming und shaming" in einen kritischen Diskurs zu treten. Es sollten viel stärker die Optimierungsmöglichkeiten und Chancen in den Vordergrund gerückt werden. Dies verdeutlicht, dass der Umgang mit Fehlern eine gezielte Auseinandersetzung benötigt, unter Umständen sogar die bestehende Kultur auf den Prüfstand stellt. Die Online-Befragung unterstützte die Forderung, dass das Gemeinwohl Vorrang vor unbegründeten Geheimhaltungsinteressen haben sollte. Zugleich wurde nochmals betont, dass Transparenz nur Mittel zum Zweck sei: Mit der gewonnenen Transparenz müssten konkrete Probleme gelöst werden.

Handlungsfeld 13: Definition eines Stufenplans
Im Vordergrund sollte die Ausgewogenheit zwischen Umfang der Daten und Umsetzbarkeit des Vorgehens stehen, da zu viele Daten möglicherweise erste Umsetzungsbeispiele verkomplizieren und verlangsamen. Es bietet sich an, zunächst mit Basisdaten zu beginnen und diese kontinuierlich auszubauen. Weitere Möglichkeiten der Abschichtung könnten die schrittweise Öffnung der Daten nach Empfänger sein (zum Beispiel Öffnung gegenüber der Wissenschaft vor Öffnung gegenüber der breiten Öffentlichkeit) oder nach Sektoren (zum Beispiel allgemeine öffentliche Einkaufsdaten vor Verteidigung und Sicherheit) oder Projekten (Projekte mit hoher Bedeutung für die Region vor Projekten unterhalb eines bestimmten Volumens). Das eröffnet die Chance, sich von kleineren zu größeren Empfängergruppen zu steigern oder sich mit zunehmendem Fach- und Projektverständnis von weniger komplexen bis hin zu komplexeren Datenbeständen zu entwickeln.

Zwischen Erhöhung der Transparenz im öffentlichen Einkauf und Korruptionsreduktion wurde vielfach ein positiver Zusammenhang festgestellt, zum Beispiel im systematischen Vergleich nach Hanna et al. [Ha11].

3.6 Partizipation und Kollaboration

Handlungsfeld 14: Erarbeitung einer Partizipations- und Kollaborationsstrategie
Bezüglich Partizipation und Kollaboration empfiehlt es sich, neben einem Projekt-Kernteam einen erweiterten Kreis an Anwendern der verschiedenen Zielgruppen einzuladen. Gemeinsam sollte eine Partizipations- und Kollaborationsstrategie erarbeitet werden, die die verschiedenen Werkzeuge/Mechanismen und ihre Ziele darlegt. Wichtig ist es, über den Gesamtverlauf des Projektes konkrete Mitarbeit und Feedback zu ermöglichen, aber auch die Weiterverwendung der öffentlichen Einkaufsdaten innerhalb und außerhalb der eigenen Organisationen zu erhöhen [MF18]. Dies fußt auf der Erkenntnis, dass die transparente Bereitstellung von Daten nicht zwangsläufig zu einer höheren Nutzung führt.

Handlungsfeld 15: Identifikation von Informationsbedarfen zur Qualifizierung
Im Vordergrund stehen Anforderungen an Informationen und Schulungen, um die öffentlichen Einkaufsdaten und ihre Verwendung im Prozess zu erklären, die Datenstruktur und -formate zu erörtern, Auswertungs- und Visualisierungswerkzeuge sowie Feedbackmöglichkeiten vorzustellen. Sie sollen einer mangelnden Nutzung und falschen Interpretation der Daten entgegenwirken, und zwar in Bezug auf alle Akteure. Dies gewinnt umso mehr an Bedeutung, da immer mehr Daten bereitstehen. Auch hier gilt es, klar abzugrenzen, welche Daten mit welchen Methoden (zum Beispiel technisch-statistisch, kulturell, strategisch, intuitiv, narrativ) ausgewertet werden [Hi14].

4 Diskussion der Ergebnisse

Die Öffnung des öffentlichen Einkaufs ist ein sensibles Thema, welches viele Akteure und Interessen betrifft und eine differenzierte Betrachtung erfordert. Gemeinsamer Nenner sollte stets das Gemeinwohl für unsere Gesellschaft und Demokratie sein. Unstrittig sind die hiermit verbundenen Chancen. Die Arbeit zeigt, dass eine Öffnung der öffentlichen Einkaufsdaten im Unterschwellenbereich für Deutschland sinnvoll sein kann: Eine erhöhte Transparenz unter Nutzung der vorhandenen Technologie kann zu einem optimierten Steuermitteleinsatz führen. Somit können bereits vorhandene Mittel effizienter eingesetzt und an die Stellen gelenkt werden, an denen aktuell Missstände beklagt werden. Profiteure können dabei sowohl die Wirtschaft, die Wissenschaft, die Medien, die Politik, der Bürger wie auch die Verwaltung selbst sein, also alle Akteure. Dies gilt unter dem Vorbehalt einer differenzierten Betrachtung der Erfordernisse (welche Daten sollen für wen und wann bereitgestellt werden) und der geltenden Rechts- und Rahmenbedingungen.

Aktuell scheint es jedoch so, dass vor einer übergreifenden Auseinandersetzung mit diesem Thema zurückgeschreckt wird. Dies wäre jedoch essentiell, um die vorhandenen Herausforderungen und Risiken auszuloten und in Chancen zu verwandeln. Motivierend ist die Einschätzung der Studienteilnehmer, wonach eine deutliche Öffnung innerhalb der nächsten fünf bis zehn Jahre denkbar sei. Die in der Befragung festgestellte Bereitschaft für

eine weitere Öffnung legt nahe, sich mit diesem Thema zukünftig ausführlich zu befassen. Die Politik sollte dieses Momentum aufgreifen, denn nur – und auch das hat die Studie bestätigt – mit einem politischen Willen zur Transparenz kann die Öffnung öffentlicher Einkaufsdaten Realität werden. Die Technologie ist längst vorhanden – das dargelegte Leitbild, die Best Practices und Handlungsfelder liefern konkrete Umsetzungsimpulse.

Literaturverzeichnis

[Ba14] Baumann, Marion; Boggild, Nikolaj; Borkowski, Lukas; Dorschfeldt, Dorian; Gädckens, Charlyn; Michels, Judith; Mutschler-Siebert, Annette; Suchanek, Alexander; Wallau, Frank: 1. Zwischenbericht - Statistik der öffentlichen Beschaffung in Deutschland - Grundlagen und Methodik. Kienbaum Management Consultants GmbH, Düsseldorf, Deutschland, 2014.

[BS13] Beck, Stefanie; Schuster, Ferdinand: Kommunale Beschaffung im Umbruch - Große deutsche Kommunen auf dem Weg zu einem nachhaltigen Einkauf? Institut für den öffentlichen Sektor e. V., KPMG AG, Berlin/Düsseldorf, Deutschland, 2013.

[FGM17] Frauscher, Kathrin; Granickas, Karolis; Manasco, Leigh: Learning Insights: Measuring results from open contracting in Ukraine. https://www.open-contracting.org/2017/04/19/learning-insights-measuring-results-ukraine/, 2017. Stand: 31.08.2019.

[FO13] Ferro, Enrico; Osella, Michele: Eight Business Model Archetypes for PSI Re-Use. https://www.w3.org/2013/04/odw/odw13_submission_27.pdf, 2013. Stand: 26.07.2019.

[Gr10] Grunwald, Armin: Technikfolgenabschätzung - eine Einführung. edition sigma, Berlin, Deutschland, 2010.

[GvL12] Geiger, Christian P.; von Lucke, Jörn: Open Government and Linked Open Government Data. Journal for eDemocracy (JeDEM), 4(2):265–278, 2012.

[Ha11] Hanna, Rema; Bishop, Sarah; Nadel, Sara; Scheffler, Gabe; Durlacher, Katherine: The effectiveness of anti-corruption policy: what has worked, what hasn't, and what we don't know - a systematic review. University of London, London, United Kingdom, 2011.

[Hi14] Hill, Hermann: Aus Daten Sinn machen: Analyse- und Deutungskompetenzen in der Datenflut. Die Öffentliche Verwaltung, 6/2014:213–221, 2014.

[MF18] Mendes, Mara; Fazekas, Mihály: DIGIWHIST Recommendations for the Implementation of Open Public Procurement Data - An Implementer's Guide. Government Transparency Institute (GTI), Budapest, Hungary, 2018.

[OE17] OECD: Government at a Glance 2017. OECD Publishing, Paris, France, 2017.

[OP09] Osterwalder, Alexander; Pigneur, Yves: Business Model Generation: A handbook for visionaries, game changers, and challengers (preview). http://radio.shabanali.com/business-model-generation-osterwalder.pdf, 2009. Stand: 27.07.2019.

[Op19a] Open Contracting Partnership: Mythbusting Confidentiality in Public Contracting. https://www.open-contracting.org/wp-content/uploads/2018/07/OCP18-Mythbusting.pdf, 2019. Stand: 19.06.2019.

[Op19b] Open Contracting Partnership: Open Contracting Data Standard. https://www.open-contracting.org/data-standard/, 2019. Stand: 02.03.2019.

[Op19c] Open Government Partnership Deutschland e. V.: Ergebnisse der Konsultation zum zweiten Nationalen Aktionsplan im Rahmen der Open Government Partnership (OGP). https://bscw.bund.de/pub/bscw.cgi/74087611?op=preview&back_url=71118924%3fclient_size%3d1366x632, 2019. Stand: 10.07.2019.

[Re21] Reuter, Britta: Transparenz öffentlicher Einkaufsdaten in Deutschland: Anforderungen und Handlungsfelder im Kontext von Open Government. Springer Gabler, Wiesbaden, Deutschland, 2021.

[SE16] Schaupp, Markus; Eßig, Michael: Ermittlung des innovationsrelevanten Beschaffungsvolumens des öffentlichen Sektors als Grundlage für eine innovative öffentliche Beschaffung. https://www.koinno-bmwi.de/fileadmin/user_upload/publikationen/Ermittlung_des_innovationsrelevanten_Beschaffungsvolumens_des_oeffentlich..._3_.pdf, 2016. Stand: 29.01.2019.

[Se18] Serghi, Laura: Blockchain-enabled Supplier Registration Information - Pilot Overview. https://ec.europa.eu/docsroom/documents/32211, 2018. Stand: 31.10.2019.

[Si13] Simonis, Georg: Partizipative Technikfolgenabschätzung und -bewertung. Springer VS, Wiesbaden, Deutschland, 2013.

[Tr17] Transparency International e. V.: Corruption Perceptions Index. https://www.transparency.org/news/feature/corruption_perceptions_index_2017, 2017. Stand: 21.01.2021.

[vL17] von Lucke, Jörn: Arbeitskreis Open Government Partnership Deutschland: Zivilgesellschaftliche Empfehlungen für den nationalen Aktionsplan OGP. https://www.researchgate.net/publication/315758978_Arbeitskreis_Open_Government_Partnership_Deutschland_Zivilgesellschaftliche_Empfehlungen_fur_den_nationalen_Aktionsplan_Open_Government_Partnership_-_23_Marz_2017_-_Kompakte_Zusammenstellung_fur_die_B, 2017. Stand: 08.03.2018.

[Š15] Šípoš, Gabriel: Once Riddled with Corruption, Slovakia Sets a New Standard for Transparency. https://www.opensocietyfoundations.org/voices/once-riddled-corruption-slovakia-sets-new-standard-transparency, 2015. Stand: 29.06.2019.

Britta Reuter ist seit mehr als 20 Jahren in der Informations- und Kommunikationstechnologie tätig. Als langjährige Führungskraft und Beraterin - zunächst in der Unternehmensberatung (Accenture), später in der Industrie (Deutsche Telekom AG, ServiceNow Switzerland GmbH) befasst sie sich mit den Schwerpunkten Strategie, Digitalisierung und Systemintegration. Berufsbegleitend erforschte sie von 2013 bis 2019 die Auswirkungen der Offenlegung öffentlicher Einkaufsdaten in Deutschland am The Open Government Institute (TOGI) der Zeppelin Universität, Friedrichshafen, bei Prof. Dr. Jörn von Lucke.

Flexible Mensch-Roboter-Zusammenarbeit durch dynamische Aufgabenteilung unter partieller Arbeitsraumbeobachtbarkeit[1]

Dominik Riedelbauch[2]

Abstract: Leichtbauroboter, die für den sicheren Betrieb ohne Schutzzäune ausgelegt sind, ermöglichen auch kleinen und mittleren Unternehmen den Zugang zu Robotikanwendungen. Auf dieser technischen Grundlage rückt die Frage nach der Aufgabenteilung zwischen Mensch und Roboter in den Fokus der Forschung, sodass sie möglichst wie ein menschliches Team effektiv zusammenarbeiten können. In dieser Arbeit wird dazu ein neuer Ansatz beschrieben, der auf Flexibilität, Kosteneffizienz durch einen reduzierten Satz an Sensorik und Bedienbarkeit durch den Endanwender abzielt. Basierend auf prozeduralem Aufgabenwissen werden Mensch und Roboter als gleichberechtigte Partner auf Augenhöhe betrachtet. Sie entscheiden sich wiederholt für Teilaufgaben und verteilen somit die Arbeit in einem dynamischen Prozess. Dies ermöglicht jederzeit flexible Übergänge zwischen Mensch-Roboter-Koexistenz, parallelem Arbeiten in Kooperation und eng synchronisierter Kollaboration. Die experimentellen Ergebnisse beleuchten insbesondere das Potential zur Beschleunigung von Aufgaben, das dieser Ansatz im Vergleich mit optimalen, fest geplanten Abläufen bietet.

1 Einführung und Stand der Forschung

In kleinen und mittleren Unternehmen (KMU) lässt sich ein wachsendes Interesse an Lösungen zur Teilautomatisierung beobachten. Flexible Einsetzbarkeit, Investitions- und Personalkosten sind hier zentrale Kriterien – starr programmierte Robotersysteme wie in der klassischen Vollautomatisierung sind damit für die Kleinserienfertigung durch hohe initiale Kosten, sperrige Schutzzaunsysteme und die Erfordernis externer Programmierexpertise nicht praktikabel [Pe19]. Die breite Verfügbarkeit von Leichtbaurobotern sowie Fortschritte bei der intuitiven Roboterprogrammierung stellen diesbezüglich einen Wendepunkt dar: Endanwender können mit inzwischen auch kommerziell verfügbaren graphischen Programmiersystemen Arbeitsabläufe teilautomatisieren. Aktuelle Sicherheitskonzepte ermöglichen danach das Arbeiten von Mensch und Roboter am gleichen Arbeitsplatz ohne physische Trennung [LFS17]. Ausgehend von diesen technischen Grundlagen rückt die Frage nach der Verteilung von Aufgaben auf Menschen und Roboter in den Fokus. Die Agenten sollen möglichst wie ein menschliches Team zusammenarbeiten – so können zukünftig durch die Nutzung individueller Stärken von Mensch (kognitive Fähigkeiten, Fingerfertigkeit) und Maschine (Kraft, Präzision, Ausdauer) effizientere und ergonomisch günstigere Arbeitsabläufe entstehen. Mit der Dissertation [Ri20], deren Ergebnisse dieser Beitrag zusammenfasst, wurde dazu ein Konzept vorgeschlagen, das gemäß den Anforderungen in KMU Flexibilität und Bedienbarkeit durch Endanwender als Ziele verfolgt:

[1] Englischer Titel der Dissertation: "Dynamic Task Sharing for Flexible Human-Robot Teaming under Partial Workspace Observability"

[2] Lehrstuhl für Angewandte Informatik III, Universität Bayreuth, dominik.riedelbauch@uni-bayreuth.de

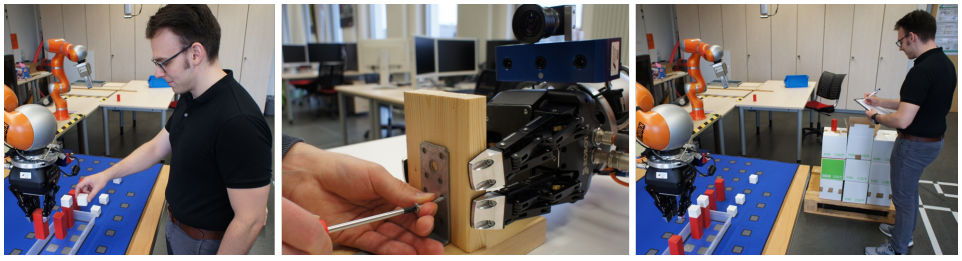


Abb. 1: Ein flexibles hybrides Team zeichnet sich durch dynamische Übergänge zwischen parallelem Arbeiten in Kooperation (links), zeitlich synchroner Kollaboration mit physischem Kontakt (mitte) und der zeitweisen Wahrnehmung unterschiedlicher Aufgaben in Koexistenz (rechts) aus.

Wissen über den Prozess sollte durch den Transfer von Erkenntnissen aus Arbeiten zur intuitiven Roboterprogrammierung vom Nutzer abgefragt und in ein geeignetes Aufgabenmodell überführt werden. In der Ausführungsphase sollten weiterhin flexible Übergänge zwischen entkoppeltem Parallelarbeiten in Kooperation, enger Interaktion mit physischem Kontakt in Kollaboration und der temporären Verfolgung unterschiedlicher Ziele in Koexistenz möglich sein (Abbildung 1). So können z. B. dringende Unterbrechungen wie eine Warenannahme behandelt werden, während das Robotersystem handlungsfähig bleibt und weiter an der ursprünglichen Aufgabe arbeiten kann. Dies ist nur möglich, wenn Mensch und Roboter wie menschliche Teams dynamisch und wiederholt Entscheidungen treffen und koordinieren können, statt einem fest vorgegebenen Plan zu folgen.

Dazu ist ein weitgehend autonomer Roboter mit Fähigkeiten zur Perzeption, Aktion und Kommunikation erforderlich. Der damit verbundene Bedarf an Sensorik sollte so weit wie möglich begrenzt werden und stützt sich deshalb vorrangig auf eine nahe der Roboterhand angebrachte Farb- und Tiefenkamera zur aktiven Inspektion von Bauteilen. Dieser kostengünstige, einfach zu kalibrierende Systemaufbau soll einerseits die Einsetzbarkeit auch für kleine Unternehmen gewährleisten – er bringt andererseits die Problematik mit sich, dass der Roboter den Arbeitsfortschritt zu jedem Zeitpunkt nur teilweise beobachten kann und so mit unvollständigem Wissen planen muss. Die zentrale Fragestellung dieser Arbeit ist deshalb, inwieweit eine produktive, nach obigen Maßstäben flexible Zusammenarbeit möglich ist, wenn dynamische Aufgabenteilung mit eingeschränkter Sensorik und dadurch nur partieller Beobachtbarkeit des Arbeitsraums durchgeführt wird.

Der Stand der Forschung verdeutlicht den Forschungsbedarf bezüglich dieser Fragestellung: Ansätze zur industriellen Mensch-Roboter-Zusammenarbeit stützen sich meist auf a-priori berechnete und damit unflexible, statische Abläufe, die aus fähigkeitenorientierten Optimierungsverfahren resultieren (z. B. [DMD19, JH17, Be05]). Im Gegensatz dazu basieren Verfahren, die mehr Flexibilität ermöglichen, auf der vollen Beobachtbarkeit des Fortschritts zu jedem Zeitpunkt (z. B. [Ju19, Da18, Sh11]) – dies kann nur mit aufwendig in Betrieb zu nehmenden Sensorik-Aufbauten wie Multi-Kamera-Systemen erreicht werden. Die Verfahren berücksichtigen weiterhin keinen intuitiven, graphischen Zugang zur Aufgabenmodellierung für den Anwender. Geeignete, fähigkeiten-basierte Methoden zur graphische Programmierung von Robotern (z. B. [SWW18, An15]) können zwar als

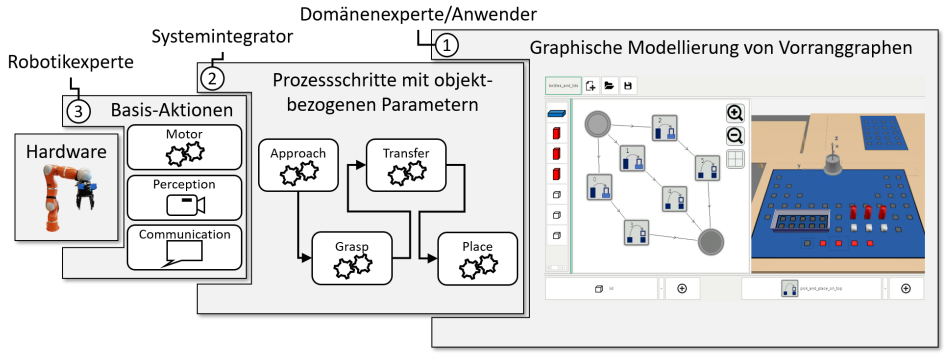


Abb. 2: Das Modell für geteilte Aufgaben ist dreischichtig: Anwender bilden Prozesse ab, indem sie mit einer grafischen Schnittstelle Prozessschritte erzeugen und zu Vorranggraphen gruppieren (Ebene 1). Jeder Prozessschritt ist mit einem vom Systemintegrator gestalteten Kontroll- und Datenfluss zwischen Basisaktionen unterlegt (Ebene 2). Die Ausführung erfolgt durch von Robotikexperten bereitgestellte Implementierung der Basis-Aktionen für ein konkretes Hardware-System (Ebene 3).

Grundlage dienen, bilden aber nur die Arbeitsschritte des Roboters ab. Hier ist die Betrachtung eines Modells, das sich mehrere Agenten teilen können, erforderlich. Zur Untersuchung der wissenschaftlichen Lücke bezüglich dynamischer Aufgabenteilung unter partieller Arbeitsraumbeobachtbarkeit wird deshalb im Folgenden zunächst ein geeignetes Aufgabenmodell und Programmierverfahren eingeführt (Abschnitt 2). Ausgehend von einem Weltmodell, das durch partielle Beobachtbarkeit entstehende Unsicherheiten bezüglich des Weltzustands berücksichtigt (Abschnitt 3), kann der Roboter mit Menschen zur Koordinierung von Aufgaben interagieren (Abschnitt 4). Darauf aufbauend ermöglicht eine prototypische Implementierung die Betrachtung des Potentials flexibler, dynamischer MRK durch Experimente im Labor und in einer Simulationsumgebung (Abschnitt 5).

2 Aufgabenmodellierung

Mentale Modelle beschreiben die kognitiven Prozesse, welche Menschen die Beschreibung und Vorhersage der Funktionsweise von Systemen ermöglichen. Für eine effektive Zusammenarbeit brauchen Menschen ein *geteiltes mentales Modell*, das eine gemeinsame Vorstellung von der geteilten Aufgabe umfasst [Ma00]. Das Problem der Etablierung dieses gemeinsamen Verständnisses reduziert sich für den betrachteten Fall der Teilautomatisierung bestehender Prozesse auf den Transfer prozeduralen Wissens vom Anwender zum Roboter. Als Approximation des mentalen Aufgabenmodells eines Roboters eignen sich Vorranggraphen: Sie stellen mit Knoten die Teilschritte einer Aufgabe dar und definieren mit Kanten „vorher-nachher“-Beziehungen zwischen ihnen. Diese partielle Ordnung kodiert explizit, inwieweit Arbeitsschritte parallel ausführbar sind, und bildet damit die Grundlage effizienter kooperativer Prozesse durch Parallelisierung. Vorranggraphen sind darüber hinaus eine im industriellen Kontext gebräuchliche Aufgabenbeschreibung. Sie lassen sich gut visualisieren und bieten damit einen Anknüpfungspunkt zur Adaptierung von Verfahren für die visuelle Roboterprogrammierung.

In Anlehnung an solche Verfahren führt das vorgeschlagene Modell Abstraktionsschichten ein, um die Komplexität von Roboterprogrammen zu verbergen (Abbildung 2). Auf der höchsten Abstraktionsebene kann der Nutzer Prozessschritte (z. B. Transfer eines Bauteils, Fügen zweier Komponenten) hardwareunabhängig und intuitiv parametrieren, indem er die zu verwendenden Bauteile in einem virtuellen Abbild der Arbeitsumgebung auswählt. Traditionelle Verfahren zur Programmierung mit Fähigkeiten [SWW18, An15] erzeugen auf diese Weise nur für den Roboter eine Sequenz von Arbeitsschritten – die Modellierung zur dynamischen Aufgabenteilung, bei der die Zuordnung zu Agenten erst zur Ausführungszeit erfolgt, erfordert hingegen die Spezifizierung aller Schritte unabhängig von der späteren Zuteilung. Dazu wird jeder Prozessschritt in einer weiteren Komponente des Aufgabeneditors durch ein Symbol dargestellt. Der Nutzer kann diese Symbole dann mit „vorher-nachher"-Beziehungen verbinden und so die Menge der vorher instanziierten Arbeitsschritte mit seinem Domänenwissen zu den gesuchten Vorranggraphen gruppieren.

Auf der nächsten Abstraktionsebene ist jeder Arbeitsschritt aus *Basisaktionen* aufgebaut, die die Rolle des Roboters bei der Ausführung beschreiben. Basisaktionen sind wiederverwendbare Bausteine innerhalb einer Domäne (z. B. Greifen oder Ablegen eines Bauteils), die die Erweiterung des Ansatzes hin zu neuen Prozessschritten für Systemintegratoren erleichtern. Ihre Implementierung in der letzten Abstraktionsschicht bildet die objektbezogenen, anschaulichen Parameter auf Hardwarebefehle ab. In Anlehnung an die Vorgehensweise von Andersen et al. [An15] ist zwischen den Basisaktionen durch Kanten ein Kontroll- und Datenfluss definiert. So lassen sich in einem Graph neben den Kern-Aktionen des Prozessschrittes auch Verzweigungen in Fehlerbehandlungsroutinen integrieren. Für die Mensch-Roboter-Zusammenarbeit werden zusätzlich zu Aktionen für Perzeption und Aktion [An15] auch solche zur expliziten Kommunikation des Roboters mit Menschen unterstützt. Das vorgeschlagene formale Modell bietet weiterhin folgende Garantien für die Koordinierungskomponente (Abschnitt 4): (i) Ein Prozessschritt terminiert entweder erfolgreich, oder im Fehlerfall nach erfolgter Rückabwicklung aller bisher ausgeführten Aktionen, sodass später ein erneuter Versuch geplant werden kann. (ii) Für jedes Bauteil, das in der Eingabe vorkommt, kann sein Zustand nach erfolgreicher Ausführung bestimmt werden. Dies entspricht der automatisierten Ableitung von Vor- und Nachbedingungen für alle beteiligten Objekte. Damit können für beliebig aus den Basisaktionen komponierte Prozessschritte die Effekte während der Aufgabenmodellierung visualisiert werden. Außerdem kann zur Ausführungszeit durch den Abgleich eines Weltmodells (Abschnitt 3) mit Nachbedingungen vom Menschen erzeugter Fortschritt beobachtet werden.

3 Alterndes Weltmodell

Ein symbolisches Roboter-Weltmodell speichert den Zustand einer Menge E von sensorisch erfassten Objekten, z. B. deren Typ und Lage. Ein Fokus dieser Arbeit liegt auf dem Umgang mit partieller Beobachtbarkeit: Mit einer am Roboter-Handgelenk angebrachten Kamera ist zu jedem Zeitpunkt t nur eine Teilmenge E_t^{vis} aller Bauteile im Arbeitsraum für den Roboter sichtbar. Ergebnisse der Objektwiedererkennung zur Zeit t können zwar genutzt werden, um Inhalte des Weltmodells zu bestätigen oder nicht mehr beobachtete Bauteile zu löschen – zeitgleich können jedoch andere Einträge ungültig werden, wenn

Menschen unbeobachtet Bauteile manipulieren. Es gilt zu vermeiden, dass das System Entscheidungen auf Basis veralteter Informationen trifft. Hierfür kommt ein *alterndes Weltmodell* zum Einsatz, mit dem die Verlässlichkeit von Daten heuristisch abgeschätzt wird. Die Theorie des autonomen Verfalls besagt, dass Erinnerungen im menschlichen Gedächtnis über die Zeit verblassen, sofern sie nicht wiederholt werden [Br58]. Diese Theorie liefert einen Ansatzpunkt für die Umsetzung des alternden Weltmodells: Jedem Weltmodelleintrag $e \in E$ wird zum Zeitpunkt t ein Wert $\mathcal{C}_t(e)$ zugeordnet, der die Verlässlichkeit der von e getragenen Information bewertet. Ausgehend vom Maximalwert 1 in dem Moment, wenn das von e beschriebene Objekt beobachtet wird, sinkt $\mathcal{C}_t(e)$ kontinuierlich gegen 0, d. h. größere Werte von $\mathcal{C}_t$ zeigen verlässlichere Daten an. Diese Datenalterung ist nur notwendig, wenn e im Einflussbereich des Menschen liegt und wahrscheinlich manipuliert wird. Dazu bestimmt eine Funktion $\mathcal{HI}$ unter Berücksichtigung ergonomischer Gesichtspunkte und des bisherigen Aufgabenfortschritts einen Wert aus dem Intervall $[0,1]$, der die Wahrscheinlichkeit menschlicher Einflussnahme schätzt. Weiterhin erlaubt eine Konstante $\lambda \in [0;\infty]$ die Steuerung des Generellen Vertrauens, das das System in ältere Informationen hat. Dabei vergisst der Roboter für $\lambda = 0$ nie und verwirft für $\lambda \to \infty$ ältere Daten sofort. Somit kann $\mathcal{C}_t(e)$ inkrementell bestimmt werden:

$$\mathcal{C}_{t+1}(e) = \begin{cases} 1 & \text{if } e \in E_t^{\text{vis}}, \\ \max\left(0,\ \mathcal{C}_t(e) - \lambda \cdot \mathcal{HI}(e)\right) & \text{sonst} \end{cases} \tag{1}$$

Aus diesem Begriff der Datenverlässlichkeit lassen sich Metriken als Grundlage für effektive Roboterentscheidungen ableiten. Sei $E_\tau \subseteq E$ die Teilmenge der Objekte im Weltmodell E, die für einen Prozessschritt τ aus dem Aufgabenmodell gemäß der Vorbedingungen benötigt werden. Sei weiterhin E_τ^{w} eine aus den Nachbedingungen von τ konstruierte Menge von Objektzuständen, die nach erfolgreicher Ausführung von τ zu erwarten sind und auf die die Weltmodellalterung analog angewandt wird. Dann sollte der Roboter bevorzugt versuchen, solche Prozessschritte auszuführen, bei denen hohe $\mathcal{C}_t$-Werte für die Elemente in E_τ die Verfügbarkeit der benötigten Ressourcen vermuten lassen. Alternativ ist die Beobachtung der Nachbedingungen von τ in dem Versuch, Fortschritte des Menschen zu beobachten, ausgehend von E_τ^{e} nur für kleine $\mathcal{C}_t$-Werte zielführend. Damit lassen sich pro Prozessschritt τ die Potentiale $\mathcal{R}$ für eine erfolgreiche Ausführung und $\mathcal{S}$ für einen Wissenszugewinn durch aktive Perzeption der Nachbedingungen bestimmen als

$$\mathcal{R}(\tau) = \prod_{e \in E_\tau} \mathcal{C}_t(e) \qquad\qquad \mathcal{S}(\tau) = 1 - \prod_{e \in E_\tau^{\text{w}}} \mathcal{C}_t(e). \tag{2}$$

Die Metriken $\mathcal{R}$ und $\mathcal{S}$ erlauben damit eine Priorisierung ausstehender Prozessschritte, die die Berücksichtigung partieller Beobachtbarkeit im Koordinierungsverfahren ermöglicht.

4 Mensch-Roboter-Koordinierung

Das Aufgabenmodell bildet alle Prozessschritte T zunächst ohne Annahmen darüber ab, ob Mensch oder Roboter in der Lage sind sie auszuführen. Von Anwendern ohne Robotik-Kenntnisse kann insbesondere Wissen bezüglich der Roboterfähigkeiten nicht erwartet

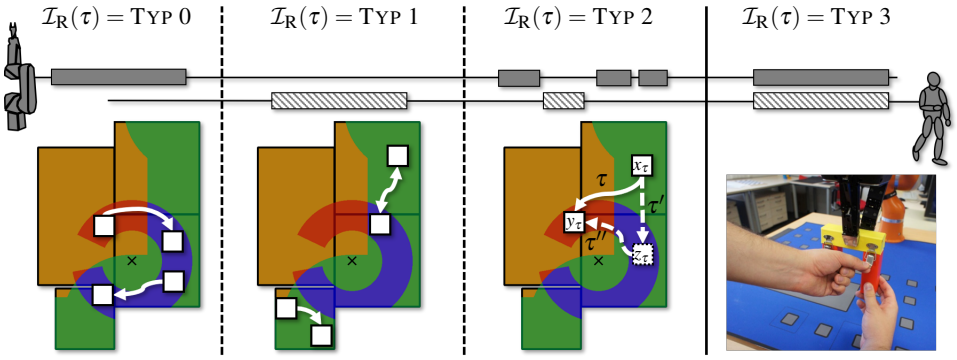


Abb. 3: Prozessschritte aus unterschiedlichen Interaktionskategorien erfordern ein zeitlich unterschiedliche ausgestaltete Teilnahme von Roboter (grau) und Mensch (gestreift). Für „pick-and-place"-Operationen erfolgt die Klassifikation durch Wissen über Arbeitsraumbereiche, die nur der Mensch (grün), nur der Roboter (rot) oder beide Agenten gleichermaßen (blau) erreichen können.

werden – so kann eine Aufgabe z. B. Bauteile umfassen, die für den Roboter unerreichbar sind. Für die Teilnahme des Roboters im dynamischen Arbeitsablauf wird jeder Prozessschritt $\tau \in T$ deshalb zunächst durch eine Klassifikation basierend auf Modellen für die Fähigkeiten von Mensch und Roboter in eine *Interaktionskategorie* $\mathcal{I}_R(\tau)$ eingeordnet (Abbildung 3). Der Roboter kann so für unterschiedliche Kategorien verschiedene Strategien zur Koordinierung verfolgen. Die Kategorien beschreiben also die Semantik der notwendigen Interaktion und Kommunikation abhängig davon, inwieweit die Teilnahme des Roboters an τ möglich bzw. notwendig ist: Sofern τ für den Roboter selbst ausführbar ist (TYP 0), ist keine Kommunikation notwendig – die Koordinierung kann rein über Perzeption erfolgen. Ist τ nicht für den Roboter, aber für den Menschen alleine machbar, so sollte das System seinen Partner aktiv darauf hinweisen (TYP 1). Ein Prozessschritt, der durch sequentielle Kooperation (TYP 2) realisierbar ist (z. B. der Transfer eines Teils aus einem nur für den Menschen in einen nur für den Roboter zugänglichen Teil des Arbeitsbereichs) muss vorher durch explizite Kommunikation abgestimmt werden (z. B. durch Aushandeln eines Übergabeortes). Vor dem Eintritt in zeitlich synchrone Kollaboration (TYP 3) sollte die wechselseitige Bereitschaft geklärt werden, um Wartezeiten zu vermeiden.

Jeder Interaktionskategorie ist ein hierarchischer Zustandsautomat zugeordnet, der das Roboterverhalten beim Auftreten eines entsprechenden Prozessschrittes τ steuert. Die Automaten kodieren auf der höchsten Hierarchiestufe, der sog. *Fortschrittsebene*, den Status der Teilaufgaben. Dazu besteht sie aus drei Zuständen: So ist τ zunächst INAKTIV. Sobald alle Abhängigkeiten gemäß der Vorrangbeziehungen erfüllt sind, erfolgt der Wechsel in den Zustand AKTIV. Konnte durch Beobachtung der Nachbedingungen bestätigt werden, dass τ erledigt wurde, schließt sich der Übergang in den Zustand ERLEDIGT an. Der AKTIV-Zustand untergliedert sich auf einer zweiten *Modus-Ebene* in zwei Unterzustände für AKTION und PERZEPTION. Der PERZEPTION-Zustand beschreibt die Vermutung, dass ein anderer Agent aktuell mit τ beschäftigt ist und folglich zeitnah eine sensorische Prüfung des Fortschritts erfolgen sollte. Analog kodiert der AKTION-Zustand, dass der Roboter zur Koordinierung von τ selbst aktiv beitragen kann. Die Zustände der Modus-Ebene

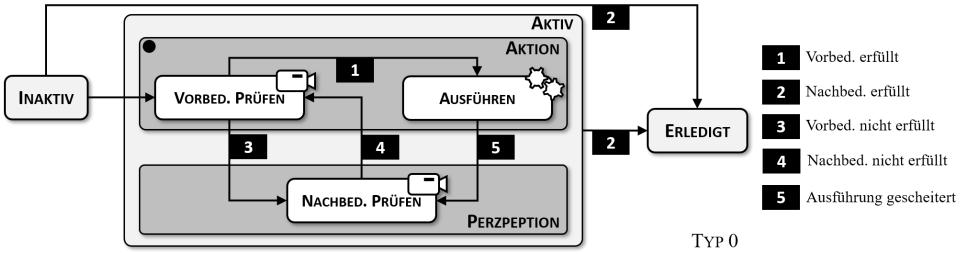


Abb. 4: Hierarchische Zustandsautomaten kodieren das Verhalten für unterschiedliche Prozessschritte auf der Fortschrittsebene (hellgrau), der Modus-Ebene (dunkelgrau) und der Aktivitätsebene (weiß). Die nächste Aktivität hängt vom Ergebnis des Vorgängers ab (schwarz). Für TYP 0 erfolgt die Koordinierung z. B. durch Perzeption der Vor-/Nachbedingungen und die Ausführung des Schrittes.

untergliedern sich weiter in je nach Interaktionskategorie unterschiedliche Abfolgen von *Aktivitäten* zur Perzeption (aktive Auswertung von Vor-/Nachbedingungen durch Bewegung der Kamera), Kommunikation (Benachrichtigung, Aushandeln von Übergabeorten und Eintritt in Kollaboration) und Aktion (Ausführung durch Abarbeitung der Basisaktionen von τ). Für jeden dieser drei Aktivitätstypen bietet je eine Architekturkomponente die Ausführung der ihr zugeordneten Aktivitäten an. Durch den Austausch der Implementierung der Komponenten lässt sich das System leicht auf unterschiedliche Sensorik, Kommunikationskanäle und Robotersysteme erweitern. Als Rückgabe der Aktivitätsausführung werden Ereignisse ausgelöst, die den Transitionen der Automaten zugeordnet sind und so Zustandsübergänge steuern. Durch diesen Aufbau ist der Ansatz leicht anpassbar: Das Koordinierungsverhalten des Roboters lässt sich durch die Bereitstellung neuer Aktivitäten und die Umstrukturierung der Zustandsautomaten ändern. Abbildung 4 zeigt als Beispiel den Automaten für Prozessschritte vom TYP 0. In diesem Fall folgt auf eine erfolgreiche Prüfung der Vorbedingungen durch Bewegung der Kamera der Versuch, den Prozessschritt auszuführen. Falls dieser Versuch scheitert, da z. B. eine notwendige Ressource nicht gegriffen werden konnte, erfolgt der Übergang in den PERZEPTION-Zustand zur Fortschrittsüberwachung usw.

Für die Funktionsweise des Systems ist es zentral, dass die Verfolgung von Transitionen innerhalb des Automaten für einen Prozessschritt in der Regel präemptiv erfolgt und die eingesetzten Kommunikationsmuster nicht-blockierend sind – das Festhalten an einem Schritt, bis er ERLEDIGT ist, könnte das System sonst z. B. bei fehlenden Bauteilen durch die zyklische Prüfung der Vor- und Nachbedingungen oder durch das Warten auf eine Antwort, wenn der Mensch aktuell nicht am Arbeitsplatz ist, blockieren. Durch die Unterbrechung nach je einer Aktivität erfolgt ein regelmäßiges Umplanen und ggf. der Wechsel zur Bearbeitung eines anderen Teils der Aufgabe, um die dauerhafte Handlungsfähigkeit und Produktivität zu gewährleisten. In manchen Situationen ist es erforderlich, dass zwei Aktionen ohne Unterbrechung direkt hintereinander ausgeführt werden. So wäre es beispielsweise für den Menschen wenig intuitiv, wenn der Roboter um Hilfe bei einem kollaborativen TYP 3-Arbeitsschritt bittet, sich nach der damit verbundenen Kommunikation aber zunächst einer anderen Operation widmet. Für diese Fälle sieht das vorgeschlagene Framework die explizite Markierung nicht-unterbrechbarer Blöcke in den Automaten vor.

Algorithmus 1 Koordinierung

repeat
  if $\exists \tau \in T$: `is_in_state`(τ, AKTION) **then**
    $T' \leftarrow \{\tau \in T \mid$ `is_in_state`(τ, AKTION)$\}$
    `trigger_actions`($\arg\max_{\tau \in T'} \mathcal{R}(\tau)$)
  else
    $T' \leftarrow \{\tau \in T \mid$ `is_in_state`(τ, PERZEPTION)$\}$
    `trigger_actions`($\arg\max_{\tau \in T'} \mathcal{S}(\tau)$)
  end if
  `update_progress_estimate`()
until $\forall \tau \in T$: `is_in_state`(τ, ERLEDIGT)

Während der Ausführung wird zu jedem Prozessschritt gespeichert, in welchem Zustand er sich innerhalb des zu seiner Interaktionskategorie passenden Automaten befindet. Damit kann die koordinierte Teilnahme des Roboters an einer Aufgabe vereinfacht mit der folgenden Schleife dargestellt werden (Algorithmus 1): Abhängig davon, ob Prozessschritte im AKTION- bzw. PERZEPTION-Zustand verfügbar sind, wird die Entscheidung für die nächste Operation entweder mit der Metrik $\mathcal{R}$ oder $\mathcal{S}$ des alternden Weltmodells getroffen. Für diejenige Operation τ_{opt} mit dem gemäß der relevanten Metrik größten Gewicht wird die Aktivität ausgelöst, die der Zustand von τ_{opt} im zugeordneten Automaten vorgibt (`trigger_actions(...)`). Dies führt einerseits zur Aktualisierung des Zustands von τ_{opt} durch das Ereignis, das die ausführende Architekturkomponente nach dem Abschluss der Aktivität auslöst. Andererseits kann die Aktivität durch Bewegung des Roboterarms mit der angebrachten Kamera vom Menschen platzierte Bauteile im Weltmodell eintragen – dieses neue Wissen wird anschließend mit den Nachbedingungen der Prozessschritte abgeglichen. Falls für einen Schritt die Bedingungen erfüllt sind, wird dieser in den ERLEDIGT-Zustand versetzt, sodass in jeder Iteration die Schätzung des Aufgabenfortschritts aktualisiert wird. Dieser Zyklus wiederholt sich, bis alle Prozessschritte ERLEDIGT sind.

5 Ergebnisse und Ausblick

Die Methode wurde auf zwei Arten ausgewertet (Abbildung 5): Zunächst wurde der Ansatz prototypisch implementiert. Neben der am Roboter angebrachten Kamera kam für die Bestimmung der vom Menschen erreichbaren Bauteile im Rahmen der Weltmodellalterung ein LIDAR-Sensor zum Einsatz. Die Kommunikation zwischen Mensch und Roboter erfolgt über eine Smartphone-Applikation, die sich an typischen Anwendungen zur asynchronen Übermittlung von Kurznachrichten orientiert. Der Demonstrator unterstützt typische „pick-and-place“-Operationen, einen einfachen Montageschritt und die Anwendung eines Werkzeugs auf Bauteile. Mit einer Nutzerevaluation konnte gezeigt werden, dass der Modellierungsansatz mit Vorranggraphen als intuitiv bewertet wird und eine schnelle Inbetriebnahme ermöglicht. Gleichermaßen wurden eine Sensorkalibrierungsprozedur und die Interaktion bezüglich der Smartphone-Applikation von den Nutzern positiv beurteilt.

Die Betrachtung zahlreicher verschiedener, dynamischer Abläufe unterschiedlicher Aufgaben für eine umfassende Abschätzung des Potentials, das das System zur Beschleu-

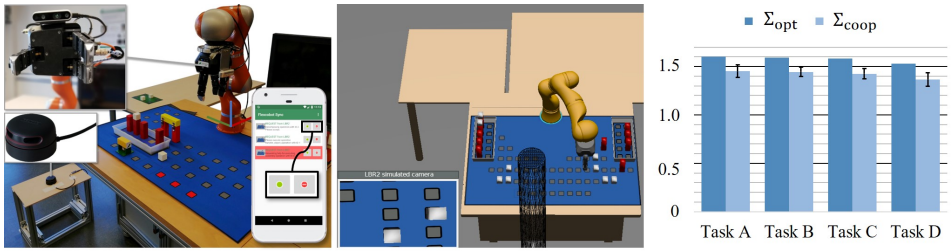


Abb. 5: Zur Auswertung wurde der Ansatz mit einem Leichtbauroboter, einer RGB-D-Kamera, einem LIDAR-Sensor und einer Smartphone-Applikation zur Kommunikation implementiert (links) und in einem Simulationssystem (mitte) hinsichtlich des tatsächlich erreichbaren Speedup (Σ_{coop}) im Vergleich zum maximal möglichen Speedup (Σ_{opt}) verschiedener Aufgaben untersucht (rechts).

nigung von Prozessen bietet, ist im Rahmen einer Nutzerstudie nicht praktikabel umsetzbar. Hierfür wurde stattdessen ein Simulationsansatz gewählt, bei dem ein Satz von Benchmark-Aufgaben wiederholt von einem virtuellen Mensch-Roboter-Team ausgeführt wurde. Menschliches Verhalten mit verschiedenen Strategien (z. B. bevorzugt räumlich lokales Arbeiten/Abschließen von Teilaufgaben/Verarbeitung aller Teile vom gleichen Typ) wurde dabei teilweise randomisiert simuliert. So konnten mehr als 3000 verschiedene Arbeitsabläufe beobachtet werden. Im Durchschnitt konnte das Robotersystem trotz einer angenommenen Diskrepanz zwischen dem höheren Arbeitstempo von Menschen im Vergleich zu als sicher einzuschätzenden Robotergeschwindigkeiten ca. 30% der Aufgaben übernehmen. Im Rahmen der betrachteten Aufgaben und der Modellbildung für die Simulation ist dadurch trotz eingeschränkter Sensorik perspektivisch eine Beschleunigung Σ_{coop} von Aufgaben möglich, die an den durch einen statisch optimierten Ansatz mit optimaler Auslastung beider Agenten möglichen Speedup Σ_{opt} heranreicht.

Zusammenfassend liegt der Beitrag dieser Arbeit in der grundlegenden Konzeption und dem Nachweis des produktiven Potentials eines erweiterbaren Ansatzes zur dynamischen Koordinierung von Menschen mit einem Roboter, der nur über begrenzte Sensorik zur Wahrnehmung des Aufgabenfortschritts verfügt. Die Ergebnisse motivieren umfangreiche zukünftige Untersuchungen: Insbesondere sollen die Skalierbarkeit auf den Einsatz mehrerer Manipulatoren und die Übertragbarkeit dieser Grundlagenarbeit auf konkretere industrielle Anwendungsfälle betrachtet werden. Darüber hinaus soll durch Lernverfahren für menschliche Arbeitsstrategien eine stärkere Anpassung an menschliche Gewohnheiten erreicht und durch komplexere Ansätze zur Fehlerbehandlung die Robustheit flexibler, dynamischer Mensch-Roboter-Zusammenarbeit gesteigert werden.

Literaturverzeichnis

[An15] Andersen, R. H.; Dalgaard, L.; Beck, A. B.; Hallam, J.: An architecture for efficient reuse in flexible production scenarios. In: IEEE International Conference on Automation Science and Engineering (CASE). Gothenburg, S. 151–157, 2015.

[Be05] Beumelburg, K.: Fähigkeitsorientierte Montageablaufplanung in der direkten Mensch-Roboter-Kooperation. Dissertation, Universität Stuttgart, 2005.

[Br58] Brown, J.: Some Tests of the Decay Theory of Immediate Memory. Quarterly Journal of Experimental Psychology, 10(1):12–21, feb 1958.

[Da18] Darvish, K.; Bruno, B.; Simetti, E.; Mastrogiovanni, F.; Casalino, G.: Interleaved Online Task Planning, Simulation, Task Allocation and Motion Control for Flexible Human-Robot Cooperation. In: IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Nanjing, S. 58–65, 2018.

[DMD19] Dalle Mura, M.; Dini, G.: Designing assembly lines with humans and collaborative robots: A genetic approach. CIRP Annals, 68(1):1–4, 2019.

[JH17] Johannsmeier, L.; Haddadin, S.: A Hierarchical Human-Robot Interaction-Planning Framework for Task Allocation in Collaborative Industrial Assembly Processes. IEEE Robotics and Automation Letters, 2(1):41–48, 2017.

[Ju19] Juelg, C.; Hermann, A.; Roennau, A.; Dillmann, R.: Efficient, collaborative screw assembly in a shared workspace. In: Advances in Intelligent Systems and Computing. Jgg. 867. Springer Verlag, S. 837–848, 2019.

[LFS17] Lasota, P. A.; Fong, R.; Shah, J. A.: A Survey of Methods for Safe Human-Robot Interaction. Foundations and Trends in Robotics, 5(3):261–349, 2017.

[Ma00] Mathieu, F. E.; Heffner, T. S.; Goodwin, G. F.; Salas, E.; Cannon-Bowers, J. A.: The influence of shared mental models on team process and performance. Journal of Applied Psychology, 85(2):273–283, 2000.

[Pe19] Perzylo, A.; Rickert, M; Kahl, B.; Somani, H.; Lehmann, C.; Kuss, A.; Profanter, S.; Beck, A. B.; Haage, M.; Rath Hansen, M.; Nibe, M. T.; Roa, M. A.; Sornmo, O.; Gestegard Robertz, S.; Thomas, U.; Veiga, G.; Topp, E. A.; Kesslar, I.; Danzer, M.: SMErobotics: Smart Robots for Flexible Manufacturing. IEEE Robotics & Automation Magazine, 26(1):78–90, mar 2019.

[Ri20] Riedelbauch, D.: Dynamic Task Sharing for Flexible Human-Robot Teaming under Partial Workspace Observability. Dissertation, Universität Bayreuth, 2020.

[Sh11] Shah, J.; Wiken, J.; Williams, B.; Breazeal, C.: Improved human-robot team performance using chaski, a human-inspired plan execution system. In: International Conference on Human-Robot Interaction (HRI). New York, S. 29–36, 2011.

[SWW18] Steinmetz, F.; Wollschlager, A.; Weitschat, R.: RAZER - A HRI for Visual Task-Level Programming and Intuitive Skill Parameterization. IEEE Robotics and Automation Letters, 3(3):1362–1369, 2018.

Dominik Riedelbauch wurde 1990 in Marktredwitz geboren. Er studierte an der Universität Bayreuth angewandte Informatik mit Anwendungsbereich Ingenieurinformatik (B.Sc.) und Computer Science (M.Sc.) mit dem Schwerpunkt Robotik. Aus seiner studienbegleitenden Tätigkeit als studentische Hilfskraft und den Projektarbeiten im Masterstudium ist sein besonderes Interesse an der Robotik-Forschung erwachsen. Seit 2016 forscht und lehrt er als wissenschaftlicher Mitarbeiter von Prof. Dominik Henrich am Lehrstuhl für Angewandte Informatik III (Robotik und eingebettete Systeme) an der Universität Bayreuth. Seit dem Abschluss seiner Promotion im Oktober 2020 führt er seine Arbeit im Bereich der dynamischen und flexiblen Mensch-Roboter-Zusammenarbeit als Postdoc in dieser Gruppe fort.

Bestärkendes Lernen und Metalernen in rückgekoppelten Netzwerken von künstlichen und gepulsten Neuronen[1]

Franz Scherr[2]

Abstract: Das menschliche Gehirn gestattet es uns neue Fähigkeiten scheinbar mühelos zu lernen, oft sogar nur durch Versuch und Irrtum, und das mit nur wenigen Anläufen. Diese Arbeit behandelt die Frage, wie eine solche Lernfähigkeit in gehirnähnlichen Schaltkreisen zustande kommen kann. Durch mathematisch rigoros hergeleitete Algorithmen werden Lernfähigkeiten in biologisch inspirierten Modellen rückgekoppelter neuronaler Netzwerke anhand von Computersimulationen demonstriert. Insbesondere erlaubt der entwickelte Lernalgorithmus solchen Netzwerken selbständig erfolgreiche Strategien durch Versuch und Irrtum zu lernen, etwa um eine hohe Punktezahl in Atari Videospielen zu erzielen. Hierbei erhält das Netzwerk keinerlei Information über die Regeln oder Ziele des Spiels, und es erhält auch keine Demonstration von wirksamen Strategien. Durch eine Optimierung von Lernsignalen, gestattet der Algorithmus erheblich schneller zu lernen. Die vorgestellte Theorie erweitert unser Verständnis davon, wie gehirnähnliche neuronale Netzwerke schwierige Aufgaben lernen können.

1 Einleitung

Das menschliche Gehirn ist ohne Zweifel eines der ausgeklügelsten Wunderwerke der Natur. Dieses Organ ist von essentieller Bedeutung um Überleben zu gewährleisten, da es dem menschlichem Organismus gestattet nach vielfältigen Verhaltensmustern zu agieren. Noch viel beeindruckender ist, dass das Gehirn uns mit der Fähigkeit ausstattet, abstrakte Beziehungen und Konzepte durch Gedanken zu verstehen. Diese Fähigkeiten erlauben uns unsere Umgebung zu deuten und damit langfristige Pläne auszuarbeiten. All das geschieht besonders effizient mit einem Energieverbrauch von etwa 20 Watt.

Diese beeindruckenden Glanzleistungen erweisen sich als erstrebenswert für künftige Entwicklungen von Intelligenz in Computersystemen, wobei das Gehirn oft die Rolle eines Vorbilds einnimmt. Unter anderem sind weit verbreitete Architekturen künstlicher neuronaler Netze von den Schaltkreisen im Gehirn inspiriert.

Tatsächlich ist man aber noch weit davon entfernt, alle Rätsel rund um die Lernfähigkeit des Gehirns zu dekodieren, und damit herauszufinden, welche Eigenschaften und Mechanismen dem menschlichen Gehirn Intellekt und Kognition verleihen. Eines steht jedoch fest: Die experimentellen Methoden, die es erlauben Prozesse in Gehirnen von Säugetieren zu untersuchen, verbessern sich rasant. Mithilfe dieser experimentellen Erkenntisse lassen sich neue Theorien für Lernprozesse in Gehirnen ableiten, deren Implikationen ihrerseits

[1] Englischer Titel der Dissertation: “Learning from rewards and with priors in recurrent networks of artificial and spiking neurons”

[2] Institut für Grundlagen der Informationsverarbeitung, Technische Universität Graz, franz.scherr@tugraz.at

wieder neue Experimente zur Verifikation vorschlagen. Theorien für Lernprozesse im Gehirn werden aber auch stets von einer weiteren Verwendungsmöglichkeit begleitet: In vielen Fällen lassen sich durch solche Erkenntisse auch neue Architekturen und Lernalgorithmen ableiten, die die Lernfähigkeit künstlicher neuronaler Netze verbessern, und damit Computer intelligenter machen.

Das zentrale Thema dieser Arbeit ist die Entwicklung eines Lernalgorithmus, und damit einhergehend eine Theorie für Lernen im Gehirn. Es wird von mathematischen Prinzipien hergeleitet, wie sich neuronale Netzwerke organisieren müssen, um bestimmte Aufgaben und Fähigkeiten zu erlernen. Das Augenmerk liegt dabei auf den folgenden beiden Lernfähigkeiten:

- Lernen mithilfe von unpräzisem Feedback, zum Beispiel nur anhand von gut/schlecht Rückmeldungen (Versuch und Irrtum),
- Schnelles Lernen von Aufgaben durch Optimierung des Lernvorganges.

1.1 Netzwerke von Neuronen und mathematische Modelle

Nervenzellen (Neuronen) sind die elementaren, informationsverarbeiteten Einheiten im Gehirn. Strukturell wird ein Neuron in verschiedene Teile gegliedert, siehe Abbildung 1A. Ein feines Netz aus sogenannten Dendriten sammelt und integriert elektrische Impulse (Aktionspotentiale) von anderen Neuronen. Die Dendriten sind mit dem Zellkern, dem Soma, verbunden. Ist dieser genügend stimuliert, wird seinerseits ein Aktionspotential erzeugt, das über das Axon an andere Neuronen und deren Dendriten gesendet wird. Siehe auch Abbildung 1B[3] und C.

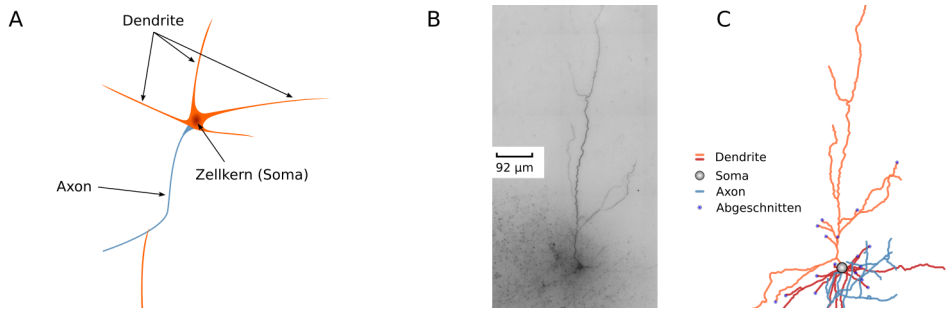


Abb. 1: **Biologische Nervenzellen. A**) Schematische Darstellung. **B**) Eingefärbtes biologisches Neuron im menschlichen Neocortex. **C**) Rekonstruktion des Neurons in B [Al].

Durch die unzähligen Verbindungen (Synapsen) im Gehirn ensteht ein vielfältiges Netzwerk, das durch die wechselseitigen Interaktionen stark rückgekoppelt ist. Die Architektur und die Stärke der Verbindungen in diesem Netzwerk charakterisieren dabei die Netzwerkdynamik und damit die Art und die Natur der Berechnung, die durch das Netzwerk als Ganzes durchgeführt wird, und ihm eine Funktion verleiht.

[3] `https://celltypes.brain-map.org/experiment/morphology/527942865`

Die entscheidende Frage ist, wie die Verbindungen in so einem Netzwerk konfiguriert werden müssen, um neue Fähigkeiten zu installieren, neue Umgebungen zu verstehen, und neue Erinnerungen zu formieren. Die Konfiguration dieser Netzwerkverbindungen, insbesondere deren Stärke, wird im folgenden als "Lernen" verstanden.

Unter Punkt 2 wird ein Algorithmus diskutiert, der im Rahmen dieser Dissertation von mathematischen Prinzipien hergeleitet wurde, und es rückgekoppelten Netzwerken gestattet zu lernen, mitunter auch in dem Szenario wo nur gut/schlecht Rückmeldungen verfügbar sind. Unter dem folgenden Punkt 3 wird auf einen Lernalgorithmus eingegangen, der es rückgekoppelten Netzwerken erlaubt, durch Optimierung des Lernvorganges, einzelne Aufgaben besonders schnell zu lernen.

2 Lernen in rückgekoppelten Netzwerken

In Computersimulationen und simulierten Lernexperimenten lassen sich biologische Neuronen anhand von Modellen studieren, die die charakteristischen Eigenschaften echter biologischer Neuronen approximieren. Speziell wird im folgenden das sogenannte "leaky integrate-and-fire (LIF)" Modell gepulster Neuronen behandelt. In diesem Modell werden eingehende Aktionspotentiale anderer Neuronen in Form einer gewichteten Summe (die Gewichte entsprechen der synaptischen Verbindungsstärke) in ein Membranpotential aufintegriert. Wenn dieses einen bestimmten Schwellwert erreicht, wird ein Aktionspotential an verbundene Neuronen ausgesandt. Ansonsten zerfällt es exponentiell.

Die besondere Schwierigkeit, um erfolgreiches Lernen zu ermöglichen, besteht darin herauszufinden, welche Verbindungen wie stark verändert werden sollen. In Bereichen des maschinellen Lernens wird dieses Problem für nicht rückgekoppelte Netzwerke typischerweise durch die Methode der Fehlerrückpropagierung, besser bekannt als "backpropagation", gelöst. Diese Methode erlaubt es den Gradienten einer Fehlerfunktion E, die suboptimales Verhalten des Netzwerks quantifiziert, zu berechnen. Mithilfe des Gradienten, der das Verhalten der Fehlerfunktion lokal bezüglich der synaptischen Gewichte beschreibt, lässt sich die Fehlerfunktion mittels Gradientenabstieg minimieren. Demgegenüber stehen rückgekoppelte Netzwerke, die üblicherweise Berechnungen effizienter durchführen können, da Neuronen und Synapsen während einer Berechnung öfter zum Ergebnis beitragen können, siehe Abbildung 2A. Doch diese Eigenschaft verkompliziert den Lernvorgang, da sich synaptische Verbindungen indirekter auswirken. Um eine Fehlerfunktion in diesem Fall zu minimieren, werden die zeitlichen Berechnungsschritte virtuell in ein nicht rückgekoppeltes Netzwerk transformiert, siehe Abbildung 2B. Wonach man in dieser aufgerollten Version dann Fehlerrückpropagierung ("backpropagation through time") anwenden kann, siehe Abbildung 2C.

Der große Nachteil dieser Prozedur ist, dass die gesamte Berechungstrajektorie gespeichert werden muss, damit alle Zwischenschritte für den Vorgang der Fehlerrückpropagierung zur Verfügung stehen. Sie würden dann zeitlich umgekehrt abgespielt werden, um Gradienten zu berechnen. Dieser Vorgang erweist sich als besonders unrealistisch, um Lernen in biologischen Gehirnen zu erklären [LS19]. In biologischen Gehirnen muss der Lernvorgang simultan mit der Netzwerkberechnung stattfinden (online).

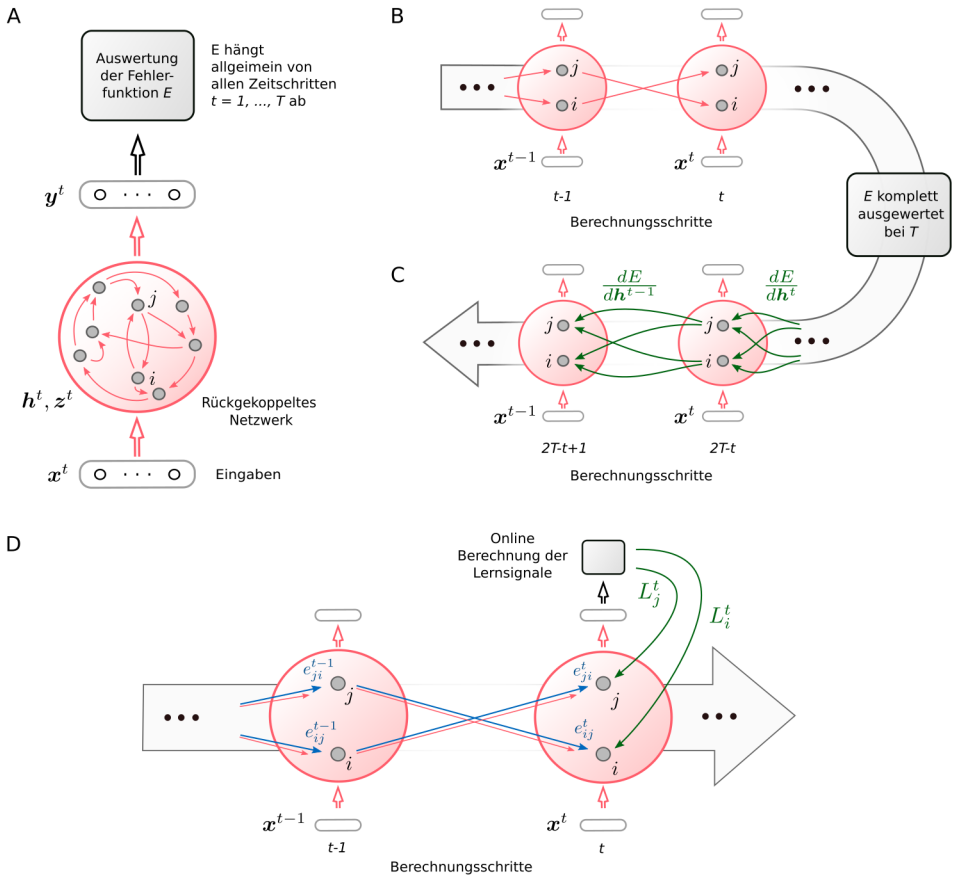


Abb. 2: **Schematische Darstellung von Fehlerrückpropagierung und e-prop. A**) Rückgekoppeltes Netzwerk. **B**) Transformation der Berechnungsvorgänge in ein nicht rückgekoppeltes Netzwerk. **C**) Fehlerrückpropagierung mit umgekehrter Netzwerkdynamik. **D**) Online Lernen mittels e-prop. Eligibility traces werden mit Lernsignalen kombiniert.

Zwei experimentelle Beobachtungen deuten auf alternative Lernalgorithmen mit eben dieser Eigenschaft hin:
Zum einen werden die jüngsten Ereignisse in der lokalen Umgebung einer Synapse auf molekularer Ebene gespeichert [SL13]. Diese kurzfristigen Ereignisspeicher sind mitunter als “eligibility traces” bekannt, und äußern sich experimentell durch synaptische Gewichtsveränderungen, wenn Lernsignale folgen [CL12, Ya14, Ge18].

Zum anderen treten Lernsignale in vielfältiger Weise im Gehirn auf, etwa in der Form von Neuromodulatoren (Dopamin, Acetylcholin) [En19, Ro13], oder auch in der Form von Netzwerkaktivität aus höheren Hirnarealen [SGS19].

Die mathematische Form von Gradienten in einem rückgekoppelten Netzwerk lässt sich im Sinne der vorangegangenen Beobachtungen reorganisieren, sodass den beiden vorherig

genannten Größen konkrete Terme zugeordnet werden können. Es resultiert ein Lernalgorithmus, der, von biologischen Beobachtungen abgeleitet, es rückgekoppelten Netzwerken erlaubt sich online zu konfigurieren, und zu lernen. Bezeichnet wird dieser Algorithmus mit e-prop (kurz für "eligibility propagation").

2.1 Mathematische Basis

Um diese Konzepte in mathematische Begriffe zu fassen, betrachtet man ein rückgekoppeltes Netzwerk, und beschreibt dessen Zeitentwicklung anhand von diskreten Zeitschritten, etwa in Schritten einer Dauer von 1 ms. Im Folgenden wird die Ausgabe jedes Neurons j mit z_j^t bezeichnet, und nimmt den Wert 1 an, falls es zur Zeit t ein Aktionspotential aussendet (und den Wert 0, falls nicht).

Durch Lernen sollen die Gewichte W der synaptischen Verbindungen so konfiguriert werden, dass die Fehlerfunktion E minimiert wird. Hierfür ist es zweckmäßig den Gradienten $\frac{dE}{dW_{ji}}$ herauszufinden, der Auskunft darüber gibt, wie sich der Wert von E lokal ändert, wenn man das Gewicht einer Synapse von Neuron i zu Neuron j ändert. Mit dieser Information wird das Gewicht W_{ji} entsprechend adaptiert ("gradient descent"). Das zentrale Resultat von [Sc20] ist, dass sich $\frac{dE}{dW_{ji}}$ als Produkt zweier Größen darstellen lässt, die den obig genannten experimentellen Beobachtungen entsprechen:

$$\frac{dE}{dW_{ji}} = \sum_t L_j^t e_{ji}^t \,. \tag{1}$$

Hierbei ist e_{ji}^t der Wert der eligibility trace zur Zeit t, die der Synapse von Neuron i zu Neuron j zugeordnet ist und L_j^t das Lernsignal zum selben Zeitpunkt.

Die eligibility trace ist unabhängig von der Form der Fehlerfunktion. Intuitiv gibt sie an, wie stark sich eine Synapse auf die Ausgabe eines Neurons (Aktionspotential) zur entsprechenden Zeit auswirkt. Die genaue Form hängt vom betrachteten Neuronenmodell ab, siehe [Sc20], und ist für das LIF Modell von einfacher Gestalt. Für detailiertere Neuronenmodelle, zum Beispiel LIF Modelle mit adaptiven Schwellwerten zur Erzeugung von Aktionspotentialen [Te18, Be18], ändert sich die konkrete Form von e_{ji}^t. Zusätzlich erlaubt e-prop ganz andere Neuronenmodelle zu betrachten, unter anderem LSTM Einheiten [HS97], die im maschinellen Lernen weit verbreitet sind.

Für das Lernsignal L_j^t, das spezifisch für jedes postsynaptische Neuron j ist, muss $L_j^t = \frac{dE}{dz_j^t}$ gelten, um die Gleichung (1) zu erfüllen. Der Beweis findet sich in [Sc20]. Intuitiv gibt das Lernsignal damit an, wie sich eine Änderung der Ausgabe des Neurons j (Aktionspotential) zum Zeitpunkt t auf den Wert der Fehlerfunktion E auswirkt.

Gleichung (1) suggeriert außerdem, wie eine Veränderung der Gewichte simultan mit der Netzwerkberechnung stattfinden soll: Addiere zum Wert des Gewichts W_{ji} zu jedem Zeitschritt t das Produkt $-\eta L_j^t e_{ji}^t$, wobei η eine kleiner Faktor (Lernrate) ist.

Durch dieses Protokoll ergibt sich die Bedingung, dass das Lernsignal L_j^t zum Zeitschritt t zur Verfügung stehen muss. Allerdings ist $\frac{dE}{dz_j^t}$ zum Zeitpunkt t nicht festgelegt, da dieser Term auch indirekte Einflüsse beinhaltet, die sich durch den Einfluss von z_j^t auf E über künftige Zeitschritte manifestieren. Man bedient sich einer Vereinfachung: Verwerfe die indirekten Einflüsse über künftige Zeitschritte, und verwende als Lernsignal nur instantane, zum aktuellen Zeitschritt auftretende Einflüsse auf die Fehlerfunktion, notiert mit der partiellen Ableitung: $\frac{\partial E}{\partial z_j^t}$. Trotz dieser Vereinfachung muss kaum Einbußen in der Lernfähigkeit hingenommen werden, wie es in [Sc20] anhand von typischen Lernexperimenten gezeigt wird.

2.2 Bestärkendes Lernen durch Versuch und Irrtum

Bei der vorangegangenen Diskussion von e-prop wurde die Fehlerfunktion E allgemein behandelt. In diesem Fall wird auf ein spezielleres Lernproblem eingegangen: Bestärkendes Lernen durch Versuch und Irrtum, bekannt als "deep reinforcement learning (deep RL)". In diesem Szenario interagiert ein Agent, representiert durch ein neuronales Netzwerk, mit einer Umgebung. In dieser Umgebung gibt es wünschenswerte Zustände, die dem Agenten ein skalares Belohnungssignal r^t ("reward") vermitteln. Das Ziel des Agenten ist es, durch die Auswahl geeigneter Aktionen in verschiedenen Zuständen (Strategie), die Summe an erhaltenen Belohnungen zu maximieren. Hierbei stehen zum Lernen einer solchen Strategie keine Hinweise zur Verfügung: Der Agent kann wirksame Strategien nur durch Versuch und Irrtum auffinden und diese nur schrittweise verbessern.

In diesem Sinne wurde in [Sc20] e-prop im Zusammenhang mit einer Fehlerfunktion betrachtet, die der "policy gradient" Methode in deep RL entspricht. Das Ergebnis der Herleitung ist eine Lernmethode, die es rückgekoppelten neuronalen Netzwerken gestattet selbstständig zu lernen, wobei lediglich Belohnungen als Rückmeldungen vorliegen. Konzeptuell ist dieses Szenario von besonderer Bedeutung um Lernvorgänge in biologischen Gehirnen besser zu verstehen.

Im Vergleich zu Gleichung (1) ergeben sich einige Unterschiede. Aktionen des Agenten werden stochastisch erzeugt, und das Lernsignal L_j^t informiert einzelne Neuronen darüber, wie stark die erzeugte Aktion vom Mittelwert abweicht. Insbesondere wird auf das Produkt $L_j^t e_{ji}^t$ ein exponentieller Filter f_γ angewandt, wobei $\gamma < 1$ den Wert zeitlich entfernterer Belohnungen vermindert. Zusätzlich kommt ein weiterer Faktor hinzu: Der "temporal difference error" $\delta^t = r^t + \gamma V^{t+1} - V^t$, wobei r^t das Belohnungungssignal der Umgebung ist, und V^t eine Schätzung des Netzwerks, wie gut die Situation zum Zeitpunkt t ist. Mit diesen Definitionen ergibt sich letztlich die folgende Vorschrift zur Adaptierung der synaptischen Gewichte zum Zeitpunkt t (siehe Herleitung in [Sc20]):

$$\Delta W_{ji} = -\eta \delta^t f_\gamma \left(L_j^t e_{ji}^t \right) \ . \tag{2}$$

In [Sc20] wurde die Wirksamkeit dieser Methode anhand von Lernexperimenten untersucht, in denen ein rückgekoppeltes Netzwerk (der Agent) mit Atari Spielen als Umgebung interagiert, und Strategien lernt.

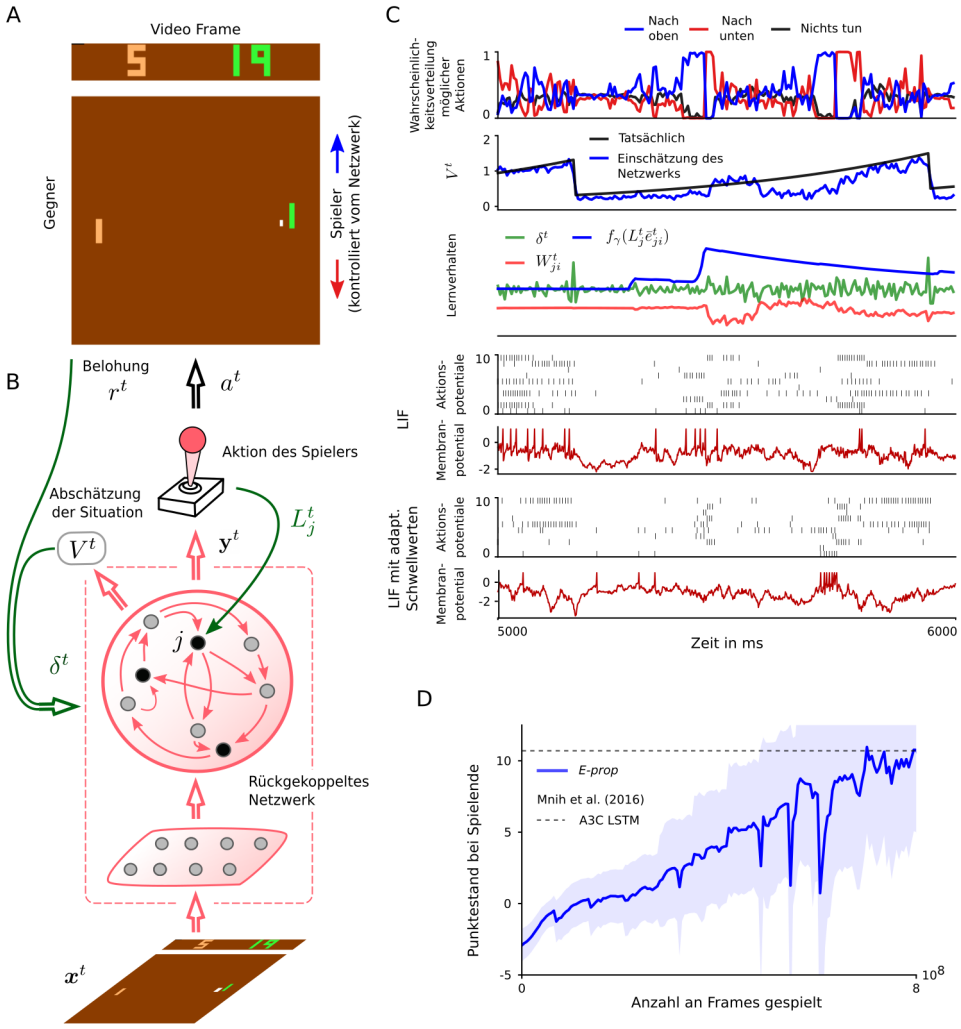


Abb. 3: **Bestärkendes Lernen mit e-prop im Atari Spiel Pong. A**) Der Spieler kontrolliert das grüne Paddel, wird durch Punkte belohnt (+1) und bestraft, falls der Gegner punktet (-1). **B**) Das grüne Paddel wird von einem rückgekoppelten Netzwerk kontrolliert, das den aktuellen Bildschirm als Eingabe wahrnimmt. **C**) Beispielhafter Zeitausschnitt nach erfolgtem Lernen. Von oben nach unten: Wahrscheinlichkeitsverteilung der stochastischen Aktionen, Einschätzung der aktuellen Situation, Lernverhalten einer zufällig ausgewählten Synapse, Aktionspotentiale und Membranpotentiale von 20 aus 400 Neuronen. **D**) Lernfortschritt des Netzwerks in Abhängigkeit der Spieldauer.

Als Beispiel sei hier das bekannte Spiel Pong angeführt, siehe Abb. 3A. In diesem Fall nimmt das Netzwerk die Umgebung über den aktuellen Bildschirminhalt war, und gibt sinnvolle Aktionen (als Wahrscheinlichkeitsverteilung über alle Möglichen) aus, siehe Abb. 3B. Für Pong steht die Auswahl dreier Aktionen zur Verfügung: Das Paddel 1) nach oben bewegen, 2) nach unten bewegen, oder 3) nicht bewegen. Das Belohnungssignal r^t

nimmt den Wert 1 an, falls ein Punkt erzielt wird, und -1, falls der Gegner einen Punkt erzielt. Nach genügend Versuchen und Interaktionen mit dem Spiel findet das Netzwerke erfolgreiche Strategien, siehe Abb. 3C und D.

3 Schnelles Lernen in rückgekoppelten Netzwerken

Obwohl e-prop es rückgekoppelten Netzwerken gestattet nichttriviale Aufgaben zu lernen, dauert der Lernvorgang üblicherweise lange, ungleich der schnellen Lernfähigkeit von Menschen. Im Gegensatz zum Vorhergehenden werden Lernsignale im Gehirn von spezialisierten Hirnarealen erzeugt [En19]. Eine stichfeste Vermutung ist, dass die Erzeugung solcher Lernsignale auch zu einem gewissen Grad durch evolutionäre Prozesse für lebenswichtige Fähigkeiten optimiert wurde.

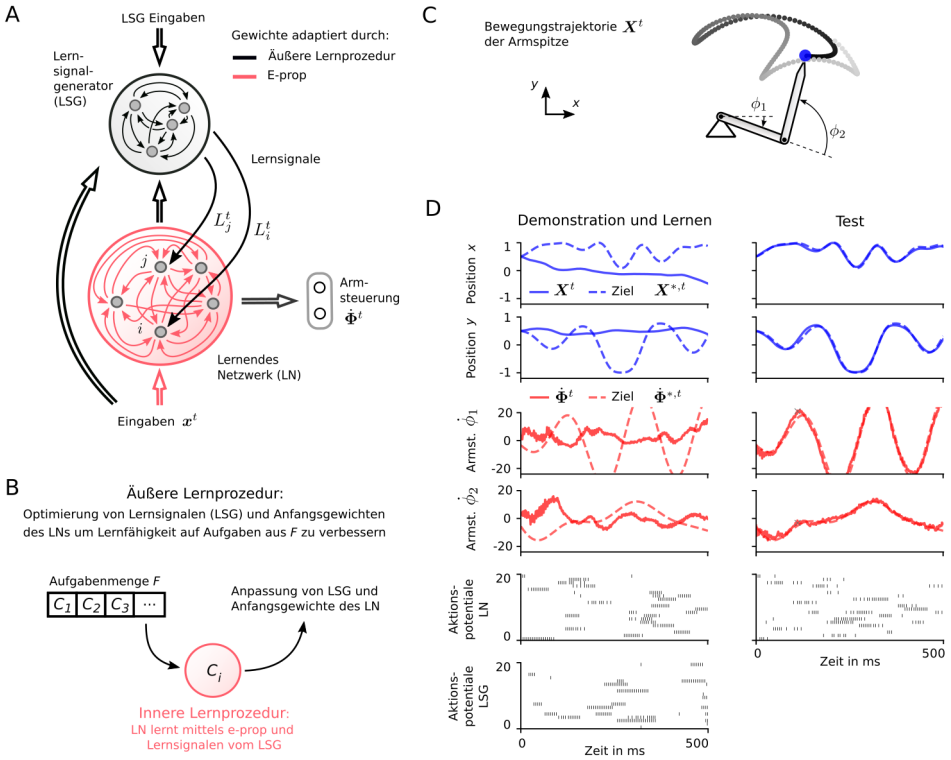


Abb. 4: **Schnelles Lernen von nur einer Demonstration mit e-prop. A**) Lernarchitektur mit zusätzlichem Lernsignalgenerator (LSG). **B**) Schematische Darstellung des Metalernens. **C**) Beispiel einer Bewegungstrajektorie einer Länge von 500 ms. **D** Schnelles Lernen einer Bewegungstrajektorie nach nur einer Demonstration. Die synaptischen Gewichte werden durch e-prop und Lernsignale des LSGs nach der Demonstration adaptiert (links). Danach wird die Bewegungstrajektorie fast perfekt reproduziert (rechts).

In einer Erweiterung von e-prop wird dieses Konzept aufgegriffen, und ein neues Netzwerk eingeführt, dessen Aufgabe es ist besonders hilfreiche Lernsignale zu erzeugen, sie-

he Abb. 4A. Dieses Netzwerk wird hier als Lernsignalgenerator (LSG) bezeichnet, um es vom lernenden Netzwerk (LN) zu unterscheiden.

Wenn nun das LN mittels Gleichung (1) unter zuhilfenahme der Lernsignale vom LSG lernt, stellt sich die Frage: Wie wird der LSG konfiguriert, sodass die erzeugten Lernsignale hilfreich sind? Zu diesem Zweck wird eine weitere Optimierung über einen längeren Zeitrahmen durchgeführt, die intuitiv evolutionären Optimierungen entsprechen würden. Dieses Konzept wird auch als Metalernen oder "learning-to-learn" bezeichnet. Es ergibt sich dabei eine eingebettete Optimierung, siehe Abb. 4B: In der inneren Lernprozedur ("inner loop") lernt das LN eine spezielle Aufgabe C mittels Gleichung (1) und den Lernsignalen gemäß dem LSG. Die resultierende Lernfähigkeit wird an die äußere Lernprozedur ("outer loop") weitergegeben, die ihrerseits versucht die Lernfähigkeit für alle $C \in F$ in einer ganzen Menge F von Aufgaben zu verbessern.

Anhand von Computersimulationen wurde dieses Konzept in [Sc20] untersucht. Unter anderem wurde als F die Menge an Bewegungstrajektorien betrachtet, die von einem Arm mit zwei Gliedern durchgeführt werden können, siehe Abb. 4C. Nach erfolgter Optimierung in der äußeren Lernprozedur über viele verschiedene Bewegungsabläufe, war das LN in der Lage neue, ungesehene Bewegungen mithilfe nur einer Demonstration zu lernen, siehe Abb. 4D. Abgesehen von diesem Experiment wurden in [Sc20] noch andere Aufgabenmengen F betrachtet.

4 Ausblick

Um die Funktion von neuronalen Netzwerken im Gehirn besser zu verstehen ist es unabdinglich zu erfassen, wie diese durch geeignete Lernmechanismen installiert werden. Die Methode der Fehlerrückpropagierung in rückgekoppelten Netzwerken kann nicht zur Beantwortung solcher Fragen herangezogen werden, da dieser Lernalgorithmus aus biologischer Sicht unrealistisch ist. E-prop benötigt keine biologisch unrealistische Komponenten, und erzielt in den betrachteten Fällen eine vergleichbare Lernfähigkeit. Insbesondere gestattet e-prop das Lernen von anspruchsvollen Aufgaben, wie etwa Atari Spiele, nur durch Versuch und Irrtum, was bisher nur biologisch unplausiblen Methoden [Mn16] vorbehalten war.

Weiters lässt sich durch eine erweiterte Optimierung, die versucht evolutionäre Optimierungsaspekte zu integrieren, die Lerngeschwindigkeit von e-prop erheblich steigern. Dadurch kann der Lernalgorithmus an die wichtigsten Lernaufgaben angepasst werden (Metalernen).

Literaturverzeichnis

[Al] Allen Institute: Allen Cell Types Database, 2018.

[Be18] Bellec, Guillaume; Salaj, Darjan; Subramoney, Anand; Legenstein, Robert; Maass, Wolfgang: Long short-term memory and Learning-to-learn in networks of spiking neurons. Advances in Neural Information Processing Systems, 2018.

[CL12] Cassenaer, Stijn; Laurent, Gilles: Conditional modulation of spike-timing-dependent plasticity for olfactory learning. Nature, 2012.

[En19] Engelhard, Ben; Finkelstein, Joel; Cox, Julia; Fleming, Weston; Jang, Hee Jae; Ornelas, Sharon; Koay, Sue Ann; Thiberge, Stephan Y.; Daw, Nathaniel D.; Tank, David W.; Witten, Ilana B.: Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature, 2019.

[Ge18] Gerstner, Wulfram; Lehmann, Marco; Liakoni, Vasiliki; Corneil, Dane; Brea, Johanni: Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Frontiers in Neural Circuits, 2018.

[HS97] Hochreiter, Sepp; Schmidhuber, Jürgen: Long short-term memory. Neural computation, 1997.

[LS19] Lillicrap, Timothy P.; Santoro, Adam: Backpropagation through time and the brain. Current Opinion in Neurobiology, 2019.

[Mn16] Mnih, Volodymyr; Badia, Adria Puigdomenech; Mirza, Mehdi; Graves, Alex; Lillicrap, Timothy; Harley, Tim; Silver, David; Kavukcuoglu, Koray: Asynchronous methods for deep reinforcement learning. ICML, 2016.

[Ro13] Roeper, Jochen: Dissecting the diversity of midbrain dopamine neurons. Trends in neurosciences, 2013.

[Sc20] Scherr, Franz: Learning from rewards and with priors in recurrent networks of artificial and spiking neurons. Dissertation, Graz University of Technology, Dezember 2020.

[SGS19] Sajad, Amirsaman; Godlove, David C.; Schall, Jeffrey D.: Cortical microcircuitry of performance monitoring. Nature Neuroscience, 2019.

[SL13] Sanhueza, Magdalena; Lisman, John: The CaMKII/NMDAR complex as a molecular memory. Molecular brain, 2013.

[Te18] Teeter, Corinne; Iyer, Ramakrishnan; Menon, Vilas; Gouwens, Nathan; Feng, David; Berg, Jim; Szafer, Aaron; Cain, Nicholas; Zeng, Hongkui; Hawrylycz, Michael; Koch, Christof; Mihalas, Stefan: Generalized leaky integrate-and-fire models classify multiple neuron types. Nature communications, 2018.

[Ya14] Yagishita, Sho; Hayashi-Takagi, Akiko; Ellis-Davies, Graham CR; Urakubo, Hidetoshi; Ishii, Shin; Kasai, Haruo: A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science, 2014.

Franz Scherr wurde am 14. Januar 1994 in der Steiermark geboren. Er besuchte die Höhere Technische Bundeslehranstalt Kaindorf. Anschließend wurde das Bachelorstudium in Telematik und Physik, und in weiterer Folge das Masterstudium in Information and Computer Engineering an der Technischen Universität in Graz absolviert. Nachdem bereits in der Masterarbeit im Bereich der neuromorphen Hardware geforscht wurde, kam es zum Doktoratsstudium unter Aufsicht von Prof. Wolfgang Maass.

Verteilte Steuerungen für beweisbar sichere, lebendige und faire autonome Fahrmanöver im Stadtverkehr [1]

Maike Schwammberger [2]

Abstract: Seit einigen Jahren erobern (teil-) autonome Fahrzeuge oder Fahrassistenzsysteme zunehmend die Märkte. Um Fehlfunktionen und Unfällen mit autonomen Fahrzeugen vorzubeugen, ist es von großer Bedeutung die korrekte Funktionsweise solcher Systeme sicherzustellen. In meiner Dissertation untersuche ich daher eine Formalisierung wünschenswerter und gesellschaftlich geforderter Eigenschaften autonomer Fahrmanöver im Stadtverkehr. Eine solche Eigenschaft ist beispielsweise die Sicherheit eines Abbiegemanövers an einer Kreuzung. Um diese Eigenschaften beweisbar zu machen, schlage ich eine Abstraktion realer Straßenverkehrssituationen vor, welche möglichst viele verschiedene Situationen abdecken kann. In diesem abstrakten Modell weise ich durch logische Schlussfolgerungen und mit Hilfe einer Implementierung die wünschenswerten Eigenschaften der von mir entwickelten Kreuzungs-Controller nach.

1 Einleitung

Wir schreiben das Jahr 2040: Kaum jemand fährt noch selber Auto. Wer von A nach B kommen möchte, profitiert stattdessen von einer Flotte autonomer Autos – per Smartphone ist die autonome Mitfahrgelegenheit überall mobil verfügbar. Auch Staus gehören in einer Zeit vernetzter autonomer Systeme mit optimaler Routenplanung der Vergangenheit an. Aufgrund fortgeschrittener Sensorik und intelligenten Planungs-Algorithmen bieten die autonomen Fahrzeuge zudem einen hohen Grad an Sicherheit. Und Maschinen schlafen nun einmal nicht am Steuer ein.

Aber wie verhält es sich wirklich mit den autonomen Fahrzeugen der Zukunft? Insbesondere die *Sicherheit* autonomer Fahrzeuge wird bereits heute, im Jahr 2020, immer wieder in den Medien diskutiert, obwohl sich bisher nur wenige autonome Testfahrzeuge auf den Straßen befinden. Werfen wir daher zum Thema Sicherheit zunächst einen Blick in die jüngere Vergangenheit.

Um allgemein die Sicherheit auf europäischen Straßen zu erhöhen, veröffentlichte die Europäische Kommission 2011 eine „Vision Zero“[3], welche das Ziel hat bis 2050 möglichst keine Straßenverkehrstoten mehr in Europa zu verzeichnen. Ein zweites

[1] Englischer Titel der Dissertation: „Distributed Controllers for Provably Safe, Live and Fair Autonomous Car Manoeuvres in Urban Traffic“

[2] Department für Informatik, Carl von Ossietzky Universität Oldenburg, schwammberger@informatik.uni-oldenburg.de

[3] Europäische Kommission: *White paper: Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system* (2011)

Ziel war die Verkehrstoten bis 2020 zu halbieren. Ein im Jahr 2018 veröffentlichter Bericht des „European Transport Safety Council“ zeigte allerdings bis 2018 lediglich einen Rückgang der Verkehrstoten um 20%. Auch wenn dieses bereits ein nennenswerter Erfolg ist, passte die Europäische Kommission ihr Ziel 2018 so an, dass die Halbierung der Verkehrstoten erst bis 2030 erreicht sein soll.

Da autonome Autos eine Zukunft auf unseren Straßen haben werden, liegt es Nahe, dass zum Erreichen einer solchen „Vision Zero“ auch die Sicherheit autonomer Fahrzeuge untersucht werden muss. Eine solche Sicherheit im Sinne von Kollisionsfreiheit gilt allgemein als nicht-verhandelbare, „harte“ System-Eigenschaft. Neben der Sicherheit gibt es weitere Eigenschaften, die ein autonomes Auto erfüllen sollte und die nötig für seine gesellschaftliche Akzeptanz sind. Beispiele hierfür sind, dass das autonome Fahrzeug ein geplantes Fahrtziel auch wirklich erreicht, oder dass das autonome Fahrzeug anderen VerkehrsteilnehmerInnen nicht die Vorfahrt nimmt.

2 Mein Forschungsbeitrag

In meiner Dissertation[Sc20] habe ich als ersten Schritt die *Sicherheit autonomer Fahrmanöver* untersucht. Hierbei liegt mein Fokus auf *innerstädtischen Kreuzungen*, da bei diesen besonders brisante Verkehrssituationen entstehen können: Fahrzeuge nähern sich der Kreuzung aus verschiedenen Richtungen, innerstädtische Straßennetzwerke sind besonders eng verzweigt und der Verkehr ist oft sehr dicht und unübersichtlich. Weiterhin bieten autonome Fahrzeuge insbesondere im Stadtverkehr spannende Einsatzmöglichkeiten: Autonome Bus-Shuttles und „Mobility on Demand“ Systeme werden bereits in einigen Städten getestet.

Der wesentliche Beitrag meiner Dissertation sind sogenannte *Kreuzungs-Controller* für Fahrmanöver an urbanen Kreuzungen, die mit Hilfe logischer Schlussfolgerungen *sichere*, *lebendige* und *faire* Abbiegemanöver durchführen können. „Lebendig“ bedeutet hier, dass Ziele des Fahrzeuges (in einer bestimmten Zeitschranke) erreicht werden und „fair“ bedeutet in meinem Fall, dass autonome Fahrzeuge sich nicht vordrängeln. Hierzu habe ich die folgenden drei Schritte verfolgt:

- Spezifikation eines **Abstrakten Modells** und einer **Logik** zur Formalisierung realer **Fahrsituationen an innerstädtischen Kreuzungen**, um logische Schlussfolgerungen für die Controller zu ermöglichen (siehe Kapitel 3),
- Semantischer Entwurf und Modellierung der **Kreuzungs-Controller** (siehe Kapitel 4), und
- Nachweis der wünschenswerten Eigenschaften **Sicherheit**, **Lebendigkeit** und **Fairness** für die Controller durch mathematische und modellbasierte Beweise (siehe Kapitel 5 und 6).

In den genannten Kapiteln gehe ich auf die Kernpunkte meiner Arbeit ein. Hierbei motiviere ich in Kapitel 3 auch die Vorteile einer abstrakten Betrachtung von Fahr-

situationen und erläutere in Kapitel 5, warum ich neben der bereits motivierten Sicherheits-Eigenschaft auch die Eigenschaften Lebendigkeit und Fairness betrachtet habe. Um die Vielseitigkeit des Ansatzes hervorzuheben, beschreibt [OS17] eine Case-Study zu einem Unfall-Warn-Protokoll auf Autobahnen, die ich in dieser Kurzfassung allerdings der Kürze wegen nicht betrachte.

3 Eine Abstraktion der realen Welt

Durch Faktoren wie die Anzahl, Positionen und Geschwindigkeiten von Fahrzeugen, aber auch durch viele Möglichkeiten für den Aufbau komplexer Kreuzungen können an einer Kreuzung extrem viele mögliche Fahrsituationen entstehen. Um logische Schlussfolgerungen und mathematische Beweise zu ermöglichen, fasse ich viele ähnliche Fahrsituationen in einem abstrakten Modell zusammen und führe die räumliche Logik UMLSL für Verkehrssituationen ein.

Abstraktion – Warum? Die Komplexität und Vielseitigkeit möglicher Fahrsituationen an innerstädtischen Kreuzungen erschwert es Beweise über Eigenschaften von autonomen Fahrzeugen zu führen. Wie in der Einleitung motiviert, ist es allerdings nötig Eigenschaften wie die Sicherheit autonomer Fahrzeuge zu gewährleisten. Mit Hilfe von Simulationen können sehr viele mögliche Fahrsituationen abgedeckt werden, aber es wird stets weitere mögliche Situationen geben, die nicht untersucht wurden. Eine Eigenschaft wie die Sicherheit kann so nur angenähert, nicht aber gewährleistet werden. Hierfür ist es hilfreich ähnliche Fahrsituationen mit Hilfe einer *Abstraktion*[4] zusammenzufassen um mehr Fälle abdecken zu können. Bei einer Abstraktion werden im Wesentlichen einige Teile eines Modells ausgeblendet.

Forschungsbasis. In [Hi11] wird ein Ansatz vorgestellt, bei dem eine Abstraktion von Fahrsituationen auf Autobahnen genutzt wird um logische Schlussfolgerungen über die Sicherheit von von autonomen Fahrspurwechsel-Manövern zu ermöglichen. Hierzu führen die Autoren die räumliche Logik *Multi-lane Spatial Logic (MLSL)* ein, welche es ermöglicht Fahrsituationen auf Autobahnen zu formalisieren. Die Erweiterung *EMLSL* aus [HLO13] ermöglicht zudem die Schlussfolgerungen über Fahrsituationen auf Landstraßen mit Gegenverkehr. Die wesentliche Abstraktions-Idee der Ansätze aus [Hi11, HLO13] ist **dynamische** und **räumliche Aspekte** voneinander zu trennen. Ein Beispiel für einen dynamischen Aspekt ist die exakte Position eines Fahrzeuges nach einer bestimmten Fahrzeit, welche in der Regel als Integral der Geschwindigkeit berechnet wird. Ein räumlicher Aspekt ist beispielsweise, dass sich ein Fahrzeug A mit ausreichend Abstand hinter einem Fahrzeug B befindet.[5]

Den Ansätzen aus [Hi11, HLO13] folgend, ist mein Anliegen logische Schlussfolgerungen über Verkehrssituationen an innerstädtischen Kreuzungen zu ermöglichen,

[4] siehe auch: Clarke, E., Grumberg, O, Long, D.E.: „*Model Checking and Abstraction*“ (ACM Trans. on Programming Languages and Systems, 1994)

[5] Ein Ansatz dafür, wie diese räumlichen Aspekte wieder mit ihrer Dynamik verbunden werden können, kann hier nachgelesen werden: Olderog, E.-R., Ravn, A.P., Wisniewski, R.: „*Linking spatial and dynamic models, applied to traffic manoeuvres*“ (Provably Correct Systems, 2017).

um formale Beweise für die Eigenschaften Sicherheit, Lebendigkeit und Fairness meiner Kreuzungs-Controller führen zu können. Als ersten Schritt habe ich in meiner Dissertation ein *Abstraktes Modell* für urbane Kreuzungen definiert, bei welchem von dynamischen Aspekten abstrahiert wird und räumliche Aspekte im Fokus stehen.

Das Abstrakte Modell. Das Abstrakte Modell für beliebige Kreuzungen enthält Fahrspuren $0, 1, 2, \ldots$, Kreuzungssegmente $c_0, c_1, c_2, \ldots$ und eindeutige Fahrzeug-Identifikatoren $A, B, C, \ldots$. Informationen wie Positionen, Größe und Geschwindigkeiten der Fahrzeuge werden in einem **Traffic Snapshot** *TS* festgehalten, welcher als eine abstrakte Momentaufnahme der Verkehrssituation verstanden werden kann. Als Beispiel für diese Kurzfassung dient das Abstrakte Modell, welches in Abb. 1 visualisiert ist. Alle folgenden Beispiele werden aus Sicht des *Ego-Fahrzeugs E* motiviert, welches sich einer Kreuzung nähert. Weiterführende Details zum Abstrakten Modell sind veröffentlicht in [**?**, Sc18a].

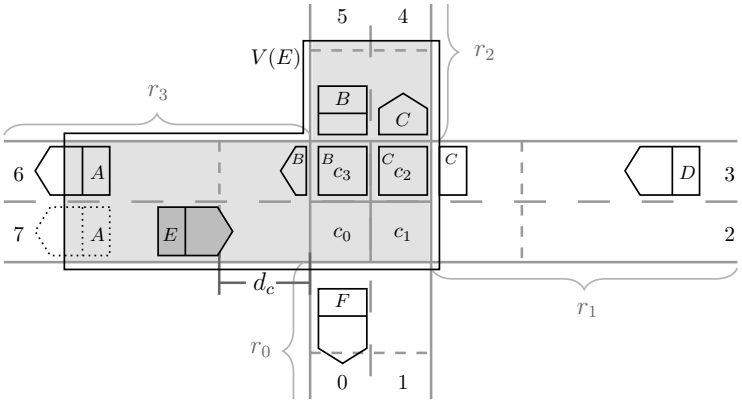


Abb. 1: Beispiel für ein Abstraktes Modell für eine 2*x*2 Kreuzung.

Zwei zentrale Konzepte prägen die Inhalte meiner Arbeit: Zum Einen die **Reservierung** $re(X)$ eines beliebigen Fahrzeugs X, welche den Bereich formalisiert den X gerade physisch einnimmt (siehe durchgezogen gezeichneter Bereich von Fahrzeug A auf Fahrspur 6). Zum Anderen beschreibt der **Anspruch („Claim")** $cl(X)$ von X den Bereich den X zukünftig befahren will. Dieser Anspruch entspricht somit dem Setzen des Blinkers (siehe gepunkteter Bereich von Fahrzeug A auf Fahrspur 7, welcher anzeigt, dass A die Fahrspur wechseln möchte, beispielsweise um ein langsameres Fahrzeug vor sich zu überholen).

In meiner Dissertation wird für ein beliebiges Ego-Fahrzeug E ein **View** $V(E)$ genutzt, mit welchem nur ein bestimmter Bereich um ein Fahrzeug als relevant betrachtet wird. Die Idee hinter der View ist, dass das Fahrmanöver eines beliebigen Fahrzeugs nur von solchen Fahrzeugen beeinflusst werden kann, die diesem nah genug sind. Dieses ist eine durchaus gängige und realistische Annahme und reduziert die Anzahl der zu betrachtenden Fahrzeuge deutlich, welches Berechnungen und Beweisführungen vereinfacht.

Bei Fahrmanövern an Kreuzungen muss häufig um die Ecke geschaut werden, so zum Beispiel auch im Fall von Ego-Fahrzeug E in Abb. 1. Hierfür führe ich in meiner Dissertation einen *Virtuellen Multi-View* $V_M(E)$ ein, welcher im Wesentlichen einen „um die Ecke gebogenen" View $V(E)$ gerade biegt. Das Ego-Fahrzeug wertet in seinem virtuellen View logische UMLSL-Formeln aus, welches in einem gebogenen View nicht möglich gewesen wäre.

Ich habe in meiner Dissertation verschiedene Abstraktionsgrade für den View betrachtet. Einen View $V(E)$, in welchem ein Ego-Fahrzeug E weniger Informationen über andere Fahrzeuge wahrnimmt, beschreibe ich in [Sc17]. Hiermit kann beispielsweise ausgedrückt werden, dass Fahrzeuge wegen eingeschränkter Sensoren weniger wahrnehmen.

Räumliche Logik UMLSL. Über Formeln der **Urban Multi-lane Spatial Logic (UMLSL)** kann die räumliche Anordnung von Fahrzeugen an und auf einer Kreuzung ausgedrückt werden. Auch die Entfernung zu anderen Fahrzeugen oder zu der Kreuzung kann formalisiert werden. Formeln der UMLSL bestehen unter anderen aus Booleschen Operatoren $\neg, \wedge, \vee$, Quantoren $\forall, \exists$ und räumlichen Atomen und Verknüpfungen, welche von der *Interval Temporal Logic* inspiriert sind [Mo85].

Mit der UMLSL-Formel $re(E) \frown free^{>0}$ wird beispielsweise über die räumliche Verknüpfung $\frown$ formalisiert, dass sich „rechts von" dem Bereich, den die Reservierung von Ego-Fahrzeug E in Abb. 1 einnimmt, freier Platz von einer Länge größer als 0 befindet.

Für den Kreuzungs-Controller (siehe Kapitel 4) wird unter anderem die folgende UMLSL-Formel $ca(E)$ (*„crossing ahead of E"*) genutzt:

$$ca(E) \equiv \langle re(E) \frown (free^{<d_c} \wedge \neg\langle cs\rangle) \frown cs\rangle, \tag{1}$$

Formel (1) ist visuell in Abb. 2 erklärt, welche eine Nahaufnahme der Situation aus Abb. 1 zeigt. Mit Formel (1) wird ausgedrückt, dass der freie Platz zwischen der Reservierung $re(E)$ von Ego-Fahrzeug E und der Kreuzung cs kleiner als d_c ist. Wenn die Formel $ca(E)$ wahr wird, wie beispielsweise im gezeigten Beispiel, dann bedeutet dieses, dass E nah genug an einer Kreuzung ist und dass der Kreuzungs-Controller beginnen muss, das Kreuzungsmanöver zu koordinieren.

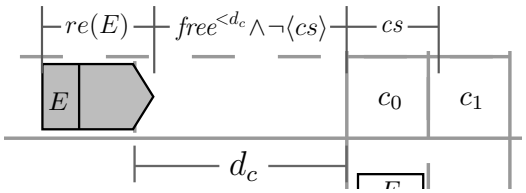


Abb. 2: Ausschnitt aus Abb. 1 zur Illustration von Formel (1), „crossing ahead of E" ($ca(E)$).

4 Kreuzungs-Controller für Abbiege-Manöver

Mit Hilfe von UMLSL-Formeln klassifizieren die Kreuzungs-Controller Fahrsituationen und führen Abbiegemanöver erfolgreich durch.

Als semantische Basis der Kreuzungs-Controller habe ich in meiner Dissertation **Automotive-Controlling Timed Automata (ACTA)** eingeführt, welche eine Erweiterung der etablierten *Timed Automata*[AD94] darstellen. ACTA nutzen die Zeitaspekte von Timed Automata und erlauben zudem UMLSL-Formeln als *Guards* und *Invarianten*. Weiterhin können mit Hilfe von speziellen *Controller-Aktionen* Fahrmanöver formalisiert werden und Informationen über *Broadcast-Kanäle mit Daten-Attributen* ausgetauscht werden. ACTA sind nicht auf den Kreuzungs-Controller begrenzt, sondern können zur Formalisierung diverser Controller für Fahrsituationen genutzt werden. Beispielsweise können sie auch als semantische Basis für die Fahrspurwechsel-Controller aus [Hi11, HLO13] dienen.

Das **Protokoll des Kreuzungs-Controllers** besteht im Wesentlichen aus drei Phasen:

1. Keine Kreuzung in der Nähe („Away“),
2. Kreuzung befindet sich voraus („Crossing ahead“), und
3. Fahrzeug ist auf der Kreuzung („On crossing“).

Abb. 3 zeigt einen vereinfachten Ausschnitt des Kreuzungs-Controllers, welcher den Wechsel des Controllers von Phase 1 in Phase 2 veranschaulicht. Solange der Controller sich in Zustand q_0 in der initialen Phase „Away“ befindet, ist für ihn nichts zu tun, da keine Kreuzung in der Nähe liegt. Sobald für das Ego-Fahrzeug E der UMLSL-Guard (1) $ca(E)$ („crossing ahead“, p. 5) wahr wird, wechselt der Controller in Zustand q_1, in welchem die Invariante $ca(ego)$ gilt, da die Kreuzung solange voraus bleibt, bis E auf die Kreuzung fährt. Weiterhin setzt der Controller beim Zustandsübergang mit der Controller-Aktion $\texttt{cc}(E)$ seinen Anspruch („claim“) auf seinen Pfad durch die Kreuzung und eine Uhrenvariable x auf 0. Die zusätzliche Invariante $x \leq t$ in Zustand q_1 stellt sicher, dass der Controller innerhalb von t Zeiteinheiten sein Kreuzungsmanöver vorbereitet. Hierzu überprüft der Controller, ob sich der von Fahrzeug E benötigte *Pfad durch die Kreuzung* überschneidet mit den Beanspruchungen $cl(X)$ oder Reservierungen $re(X)$ anderer Fahrzeuge X. Falls der Controller dieses aufgrund eingeschränkter Informationen nicht einschätzen kann, sendet er eine Broadcast-Anfrage an die anderen Fahrzeuge (siehe [Sc17]). Nur falls der Pfad durch die Kreuzung frei ist, startet der Controller das Kreuzungs-Manöver und wechselt in Phase 3., „On crossing“.

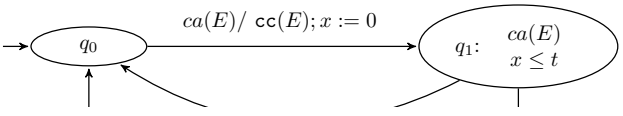


Abb. 3: Vereinfachter Ausschnitt des Kreuzungs-Controllers.

5 Eigenschaften des Kreuzungs-Controllers

Was bedeuten die Eigenschaften Sicherheit, Lebendigkeit und Fairness und warum spielen sie eine Rolle für autonome Fahrmanöver an Kreuzungen?

In der Einleitung habe ich motiviert, warum ich in meiner Arbeit die **Sicherheit** autonomer Fahrmanöver untersuche. Eine solche Sicherheitseigenschaft kann allerdings nicht isoliert betrachtet werden. Wenn mein neues autonomes Fahrzeug sicher in meiner Garage steht und nie bewegt wird, dann wird es auf jeden Fall sicher bleiben, da es keine Kollisionen auslösen kann. Die Sicherheits-Eigenschaft ist damit trivialerweise erfüllt. Aber ein Fahrzeug, welches sich nie bewegt ist auch kein sinnvolles System. Daher ist die zweite Eigenschaft, die ich betrachte, die **Lebendigkeit** eines autonomen Fahrzeuges. In diesem Fall bedeutet Lebendigkeit, dass ein wünschenswertes Ziel irgendwann, oder im Laufe der System-Lebendigkeit immer wieder, erreicht wird. Es ist allerdings wiederum nicht wünschenswert, dass das Fahrzeug irgendwann, beispielsweise nach zehn Stunden erst, die Kreuzung überquert. Daher untersuche ich als Verfeinerung von Lebendigkeit eine **zeitlich beschränkte Lebendigkeit**. Hierbei untersuche ich die Wahrscheinlichkeit dafür, dass ein Fahrzeug innerhalb eines Zeitintervalls $[0,k]$ eine Aktion erfolgreich durchführt.

Ein solches Zeitintervall sollte allerdings differenziert betrachtet werden: Im Best Case führt das Fahrzeug die Aktion möglichst schnell durch und im Worst Case benötigt das Fahrzeug immer k Zeiteinheiten für das Manöver. Hiermit können *unfaire* Szenarien auftreten: Beispielsweise kann ein Fahrzeug stets als erstes eine Kreuzung erreichen, aber sie immer als letztes überqueren. Daher habe ich ein **Fairness-Protokoll** entwickelt, bei welchem Fahrzeugen Prioritäten zugewiesen werden, abhängig davon wann sie an einer Kreuzung ankommen. Wenn ein Fahrzeug vor der Kreuzung warten muss, dann erhöht sich seine Priorität und damit auch sein Recht die Kreuzung vor einem Fahrzeug mit niedrigerer Priorität zu befahren. Ein praktischer Vorteil dieses Ansatzes ist, dass man einem autonomen Rettungsfahrzeug automatisch eine besonders hohe Priorität zuweisen könnte. Damit geben andere autonome Fahrzeuge dem Rettungsfahrzeug stets Vorfahrt.

Bei (zeitlich beschränkter) Lebendigkeit und Fairness handelt es sich um **temporale Eigenschaften**:

- **Lebendigkeit** bedeutet, dass ein Ziel (z.B. Kreuzung überqueren) *irgendwann in der Zukunft* erreicht wird,
- **Zeitlich beschränkte Lebendigkeit** bedeutet, dass ein Ziel innerhalb einer Zeitschranke k erreicht wird, und
- **Fairness** bedeutet, dass ein eigenes Ziel nicht viel später erreicht wird als das Ziel eines konkurrierenden Fahrzeugs (welches zur gleichen Zeit/ später die Kreuzung erreicht).

Diese Temporalität nutze ich für die Nachweise der Eigenschaften aus, welche ich im folgenden Kapitel skizziere.

6 Beweisskizzen

Der Kreuzungs-Controller wird auf die Probe gestellt: Mit mathematischen und modellbasierten Beweisen zeige ich, dass die Eigenschaften Sicherheit, Lebendigkeit und Fairness zu keinem Zeitpunkt verletzt werden.

Die Sicherheit des Kreuzungs-Controllers kann für ein beliebiges betrachtetes Fahrzeug E als UMLSL-Formel ausgedrückt werden:

$$Safe_E \equiv \neg\exists c \neq E \wedge \langle re(E) \wedge re(c) \rangle \tag{2}$$

Formel (2) sagt aus, dass es kein anderes Fahrzeug c als E gibt, welches eine Überschneidung des eigenen reservierten Bereichs $re(c)$ mit dem reservierten Bereich $re(E)$ von E hat. Damit sagt die Formel im Wesentlichen aus: „Es gibt keine Kollisionen mit Fahrzeug E“. Die Sicherheit des Kreuzungs-Controllers zeige ich **induktiv über dessen Semantik**: Man starte bei einer beliebigen initialen, sicheren, Verkehrssituation TS_0. Nun gilt es zu zeigen, dass für alle beliebigen, von TS_0 erreichbaren Verkehrssituationen TS_k die Sicherheitseigenschaft (2), $Safe_E$, erhalten bleibt. Konkret zeige ich, dass weder durch das Vergehen von Zeit, noch durch eine ausgeführte Controller-Aktion eine Verkehrssituation TS_k entstehen kann, bei welcher $Safe_E$ verletzt wird. Da E ein beliebiges Fahrzeug ist, lässt sich die Gültigkeit von $Safe_E$ auf jedes beliebige Fahrzeug übertragen.

Wie im vorherigen Kapitel beschrieben, handelt es sich bei (zeitlich beschränkter) Lebendigkeit und Fairness um temporale Eigenschaften. Da die Logik UMLSL keine temporalen Operatoren enthält, lassen sich diese Eigenschaften nicht als UMLSL-Formeln ausdrücken. Für das **Model-Checking von Timed Automata** [ACD90] werden daher Anfragen in *Timed Computation Tree Logic (TCTL)* verwendet. Beim Model-Checking wird zu einem gegebenen *Modell* automatisiert überprüft, ob es eine gegebene *Spezifikation* erfüllt.

Um (zeitlich begrenzte) Lebendigkeit und Fairness nachzuweisen, habe ich meinen Ansatz in dem Model-Checking Tool **UPPAAL SMC** [Da15] implementiert. Das Modell ist in meinem Fall eine in UPPAAL kodierte Kreuzung zusammen mit je einer Instanz des Kreuzungs-Controllers für jedes betrachtete Fahrzeug an der Kreuzung. Die untersuchten Spezifikationen („verification queries“) sind die untersuchten Eigenschaften, die ich für die Verifikation mit UPPAAL als TCTL-Formeln formuliert habe. Lebendigkeit lässt sich beispielsweise als TCTL-Formel

$$\texttt{A} <> \texttt{Observer(E).success} \tag{3}$$

formulieren. Hierbei ist `Observer(E)` ein *Beobachter-Automat*, welcher den Kreuzungs-Controller von Fahrzeug E während der Verifikation beobachtet und genau dann in den Zustand `Observer(E).success` wechselt, wenn der Kreuzungs-Controller von Fahrzeug E in die Phase „On crossing“ wechselt. Die Query (3) wird von UPPAAL genau dann als gültig verifiziert, wenn in allen möglichen System-Abläufen (`A`) irgendwann (<>) der Zustand `Observer(E).success` erreicht wird.

Die zeitlich begrenzte Variante dieser Query ist

$$\texttt{Pr}[<= \texttt{k}](<> \texttt{Observer(E).success}). \tag{4}$$

Zu Query (4) gibt UPPAAL aus, mit welcher Wahrscheinlichkeit der Zustand `Observer(E).success` innerhalb einer Zeitschranke k erreicht wird. Hiermit habe ich somit untersuchen können, wie schnell Fahrzeuge mit meinen Kreuzungs-Controllern die Kreuzung überqueren können. Dieses habe ich in abgewandelter Form auch für die Fairness-Queries ausgenutzt. Details zu den Liveness- und Fairness-Verifikationen mit UPPAAL SMC sind in [Sc18b, BS19] veröffentlicht.

7 Fazit

Die wesentlichen Beiträge meiner Dissertation sind

- ein universelles Abstraktes Modell für Fahrsituationen an Kreuzungen, inklusive räumlicher Verkehrslogik UMLSL,
- Automotive-Controlling Timed Automata zur Konstruktion eines Kreuzungs-Controllers, und
- mathematische und modellbasierte Beweise über die Eigenschaften Safety, Liveness und Fairness des Controllers.

Durch seine Allgemeinheit ist mein Ansatz nicht auf Abbiegemanöver an Kreuzungen limitiert: Das Abstrakte Modell, die Logik sowie das ACTA-Modell sind auf andere Verkehrsszenarien übertragbar. Diese Wiederverwendbarkeit lässt Spielraum für diverse Erweiterungen (z.B. autonome Manöver auf Parkplätzen, im Kreisverkehr, formale Modellierung von Abbiegespuren, ...).

Bisher schwer oder nur ungenau nachweisbare und komplexe Ansätze zum autonomen Fahren können zudem mit meiner Methode abstrahiert und beweisbar werden. Der nächste Schritt ist dann von dem beweisbaren abstrakten Konstrukt in Richtung Realität zu gehen und Teile der Abstraktion wieder zu konkretisieren. Einen Ansatz hierzu bespreche ich in [Sc17], indem ich Sensorungenauigkeit und damit nicht perfektes Wissen zulasse.

Literaturverzeichnis

[ACD90] Alur, Rajeev; Courcoubetis, Costas; Dill, David L.: Model-checking for real-time systems. In: 5th IEEE Symposium on Logic in Computer Science, Proc. S. 414–425, 1990.

[AD94] Alur, Rajeev; Dill, David L.: A Theory of Timed Automata. Theoretical Computer Science, 126(2):183–235, 1994.

[BS19] Bischopink, Christopher; Schwammberger, Maike: Verification of Fair Controllers for Urban Traffic Manoeuvres at Intersections. In: Formal Methods. FM 2019 International Workshops - Porto, Portugal, October 7-11, 2019, Revised Selected Papers, Part I. Jgg. 12232 in Lecture Notes in Computer Science. Springer, S. 249–264, 2019.

[Da15] David, Alexandre; Larsen, Kim G.; Legay, Axel; Mikučionis, Marius; Poulsen, Danny Bøgsted: Uppaal SMC tutorial. STTT, 17(4):397–415, 2015.

[Hi11] Hilscher, Martin; Linker, Sven; Olderog, Ernst-Rüdiger; Ravn, Anders P.: An Abstract Model for Proving Safety of Multi-lane Traffic Manoeuvres. In: 13th Int. Conference on Formal Engineering Methods, ICFEM, Proc. Springer, S. 404–419, 2011.

[HLO13] Hilscher, Martin; Linker, Sven; Olderog, Ernst-Rüdiger: Proving Safety of Traffic Manoeuvres on Country Roads. In: Theories of Programming and Formal Methods – Essays Dedicated to Jifeng He on the Occasion of His 70th Birthday. Jgg. 8051 in LNCS. Springer, 2013.

[HS16] Hilscher, Martin; Schwammberger, Maike: An Abstract Model for Proving Safety of Autonomous Urban Traffic. In: 13th Int. Colloquium on Theor. Aspects of Computing ICTAC, Proc. Jgg. 9965 in LNCS. Springer, S. 274–292, 2016.

[Mo85] Moszkowski, Ben: A Temporal Logic for Multilevel Reasoning About Hardware. Computer, 18(2):10–19, 1985.

[OS17] Olderog, Ernst-Rüdiger; Schwammberger, Maike: Formalising a Hazard Warning Communication Protocol with Timed Automata. In: Models, Algorithms, Logics and Tools – Essays Dedicated to Kim G. Larsen on the Occasion of His 60th Birthday. Jgg. 10460 in LNCS. Springer, S. 640–660, 2017.

[Sc17] Schwammberger, Maike: Imperfect Knowledge in Autonomous Urban Traffic Manoeuvres. Electr. Proc. in Theor. Comp. Sci., 257:59–74, 2017.

[Sc18a] Schwammberger, Maike: An abstract model for proving safety of autonomous urban traffic. Theoretical Computer Science, 744:143–169, 2018.

[Sc18b] Schwammberger, Maike: Introducing Liveness into Multi-lane Spatial Logic lane change controllers using UPPAAL. Electronic Proceedings in Theoretical Computer Science, 269:17–31, 2018.

[Sc20] Schwammberger, Maike: Distributed Controllers for Provably Safe, Live and Fair Autonomous Car Manoeuvres in Urban Traffic. Dissertation, University of Oldenburg, 2020.

Maike Schwammberger ist eine Wissenschaftliche Mitarbeiterin an der Carl von Ossietzky Universität Oldenburg in der Arbeitsgruppe „Hybride Systeme“ von Prof. Dr. Martin G. Fränzle. Nachdem sie Mathematik, Kunst und Medien und Informatik an der Universität Oldenburg studierte, trat sie 2014 als Doktorandin der Arbeitsgruppe „Entwicklung korrekter Systeme“ von Prof. Dr. Ernst-Rüdiger Olderog bei. Mit ihren Forschungsergebnissen zur abstrakten Modellierung und Beweisführung im Bereich autonomer Fahrmanöver hat sie im Jahr 2016 einen Best Paper Award für [HS16] erhalten und wurde 2019 zum GI-Dagstuhl Seminar 19023 zum Thema *Explainable Software for Cyber-Physical Systems* eingeladen. Während ihrer Zeit als Doktorandin hat sie sich auch regelmäßig in Lehrveranstaltungen engagiert.

Ein Ansatz zur Softwaretechnischen Unterstützung des Qualifikationsbasierten Lernens (QBL) an Hochschulen[1]

Matthias Then[2]

Abstract: Der Ansatz des *Qualifikationbasierten Lernens (QBL)* knüpft an vorhandene Forschungsergebnisse und Softwarelösungen im Bereich des kompetenzbasierten Lernens an, der Fokus liegt dabei auf der Entwicklung leistungsfähiger Softwarekomponenten. Das vor diesem Hintergrund entwickelte QBL Domänenmodell ermöglicht sowohl die Verwendung standardisierter Kompetenz- bzw. Qualifikationsrahmenwerke, als auch den Entwurf eigener Strukturen, wie sie zur flexiblen Gestaltung institutionsspezifischer Bildungsangebote benötigt werden. Elemente unterschiedlicher Rahmenwerke können zueinander in Beziehung gesetzt werden, was die Entstehung eines umfangreiches Qualifikationsnetzes begünstigt. Neben dem Domänenmodell beinhaltet QBL ein Architekturmodell zur Integration qualifikationsbasierter Softwarekomponenten in Hochschul-IT-Landschaften und ein Servicemodell. Im praktischen Teil der Arbeit werden konkrete Anwendungsszenarien ausgearbeitet, realisiert und evaluiert. Ein in diesem Kontext entstandener Forschungsprototyp ist QBL4Moodle, ein Plugin für das Learning Management System Moodle.

1 Einführung

Der Ansatz des *Qualifikationsbasierten Lernens* (engl. *Qualifications-Based Learning, QBL*) entstand vor dem Hintergrund, dass die Idee des *Kompetenzbasierten Lernens* (engl. *Competence-Based Learning, CBL*) zunehmend Eingang in die Lehr-/Lernprozesse an Hochschulen und anderen Bildungseinrichtungen findet. Demzufolge wird auch deren softwaretechnische Unterstützung verstärkt nachgefragt, insbesondere bei der Konzeption und Durchführung von Modulen und Kursen in den zentralen Lehr-/Lernplattformen wie dem *Learning Management System* (*LMS*) *Moodle* [Mo21]. Bestehende Ansätze, Modelle und Softwaresysteme werden den steigenden Anforderungen jedoch nur bedingt gerecht, Verbesserungsbedarf besteht noch in vielen Bereichen. Das generelle Ziel von QBL besteht darin, Softwarelösungen für CBL-Visionen zu realisieren wie: Erstellung von Kursen und Studienplänen, Entwurf und Durchführung von Lehr-/Lernszenarien, Speichern von Nutzerprofilen und hochschulübergreifender Vergleich kompetenzbezogener Informationen wie beispielsweise *Lernziele* (engl. *Learning Goals*) und Zugangsvoraussetzungen zu Kursen, Modulen und Studiengängen. QBL greift dabei auf vorhandene Forschungsergebnisse und Softwaresysteme zurück und schlägt Ergänzungen und Erweiterungen vor.

Hinweis zur Terminologie: Der Begriff QBL wurde eingeführt, da in CBL der Begriff Kompetenz nicht immer konsequent von anderen Qualifikationstypen wie Fertigkeit oder Kenntnis abgegrenzt wird. Im Folgenden wird daher analog zu [Th20] als Oberbegriff

[1] Englischer Titel der Dissertation: „Supporting Qualifications-Based Learning (QBL) in a Higher Education Institution's IT-Infrastructure"

[2] FernUniversität in Hagen, matthias.then@fernuni-hagen.de

für Qualifikationen jeglicher Art die Bezeichnung *Kompetenz/Qualifikation* (engl. *Competence/Qualification, CQ*) verwendet.

Im Vorfeld dieser Arbeit wurde beobachtet, dass zur Definition CQ-basierter Lernziele von Studiengängen und Lehrveranstaltungen häufig noch Freitext verwendet wird, Gleiches gilt für Zugangsvoraussetzungen. Das ist insofern ein Problem, dass dadurch die Vergleichbarkeit erschwert wird. Der Einsatz standardisierter *CQ Rahmenwerke* (engl. *CQ Frameworks, CQF*) hat sich noch nicht flächendeckend durchgesetzt, zudem werden sie softwaretechnisch noch nicht in ausreichendem Maß unterstützt. Dazu kommt, dass auch auf Ebene der Hochschulen nur in seltenen Fällen institutionsweite CQ Kataloge vorhanden sind. Außerdem ist unklar, wie sich solche Konzepte in die IT-Landschaften von *Hochschuleinrichtungen* integrieren lassen und wie der Datenfluss zwischen den beteiligten Systemen aussieht.

Anhand dieser Beobachtungen wurden Forschungsfragen formuliert, aus denen die folgenden *Forschungsziele* (engl. *Research Goals, RG*) abgeleitet wurden:

- **RG1:** Erkundung und Analyse des Themenfelds *Kompetenzbasiertes Lernen* aus der Perspektive des Softwareentwicklers. Dabei sind folgende Aspekte zu untersuchen und zu bewerten: Existierende Ansätze, verfügbare Softwareunterstützung, Anwendbarkeit in verteilten Architekturen, Datenflüsse und Kombinierbarkeit mit gängigen Austauschformaten und Interoperabilitätsstandards.
- **RG2:** Entwurf eines allgemeinen *QBL Modells (QBLM)*, das in der Lage ist, CQFs umfassend zu unterstützen und CQ-bezogene Daten innerhalb einer Hochschul-IT-Landschaft auszutauschen. Das QBLM beinhaltet ein Domänenmodell, ein Architekturmodell und ein Servicemodell.
- **RG3:** Entwicklung und initiale Evaluation einer prototypischen Softwarelösung. Dabei ist ein gängiges open-source LMS dahingehend zu erweitern, dass es grundlegende QBL Funktionalitäten auf Basis des QBLM zur Verfügung stellt.
- **RG4:** Entwurf, Realisierung und Evaluation verteilter QBL Anwendungsszenarien innerhalb einer geeigneten Hochschul-IT-Landschaft. Der im Rahmen von RG3 zu entwickelnde Prototyp und andere involvierte Softwarekomponenten sind entsprechend zu erweitern.

2 Stand von Wissenschaft und Technik

Im Folgenden werden einige Ansätze, Technologien und Softwaresysteme kurz beschrieben, die für die Entwicklung von QBL von entscheidender Bedeutung waren.

Im Hinblick auf die CQ-basierte Vergleichbarkeit von Kursen, Modulen, Studiengängen und Lehr-/Lerninhalten wurden neben diversen standardisierten CQFs auch institutionsspezifische CQ Kataloge und Domänentaxonomien aus dem Bereich der *Informations- und Kommunikationstechnologie* untersucht. Von besonderer Bedeutung für die Entwicklung

von QBL war dabei der *Europäische Qualifikationsrahmen für lebenslanges Lernen (EQF)* [Eu08], der als Vorlage zur Entwicklung konkreter, domänenspezifischer CQFs verwendet wird und dessen EQF-Levels eine EU-weite Vergleichsbasis für Kompetenzniveaus bieten. Ebenso hervorzuheben ist das *European e-Competence Framework (e-CF)* [Eu14], eine sektorspezifische EQF-Implementierung für den Bereich Informations- und Kommunikationstechnologie.

Ein CQ-basierter Ansatz, der ähnliche Ziele wie QBL verfolgt, wurde im Rahmen des *TENCompetence Projekts* (siehe [Ko09], Kap. 1, 18 und 19) erarbeitet. TENCompetence steht dabei für „Building the European Network for Lifelong Competence Development". Wie auch QBL propagiert TENCompetence den Einsatz standardisierter Formate, was u.a. daran zu erkennen ist, dass sich die Modellierung von Lehr-/Lernprogrammen und -inhalten stark an *IMS Learning Design* [KT06] orientiert. Diese Strukturen werden um kompetenzbezogene Informationen ergänzt. Einen Überblick über das resultierende Domänenmodell findet man in [Ko08]. Die im Rahmen von TENCompetence erarbeiteten Konzepte, insbesondere das erwähnte Domänenmodell, werden als geeignete Grundlage für QBL angesehen, QBL kann somit als eine Weiterentwicklung von TENCompetence betrachtet werden.

Ein nützliches Werkzeug bei der Planung und Durchführung verteilter Lehr-/Lernszenarien ist die Interoperabilitätsspezifikation *IMS Learning Tools Interoperability* (*LTI*) [Im21]. LTI kann beispielsweise eingesetzt werden, um in einen LMS-Kurs externe (d.h. von anderen Softwaresystemen gehostete), zugangsbeschränkte Ressourcen einzubetten. Dabei kann es sich beispielsweise um Aufgaben, Tests, Videos, Lernprogramme oder Spiele handeln. Dies geschieht auf der Basis von definierten Services; die LTI Basisservices sind: Single Sign On, Tool Launch und Return of Outcomes.

Bei der Realisierung innovativer Lehr-/Lernszenarien spielen LMSe heutzutage eine zentrale Rolle, daher wird im Rahmen von RG3 die Erweiterung eines gängigen LMSs angestrebt. Bei der Entscheidung für ein geeignetes Basissystem fiel die Entscheidung auf *Moodle* [Mo21] (Moodle: „Modular Object-Oriented Dynamic Learning Environment"), ein weit verbreitetes, frei verfügbares open-source LMS. Für Moodle sprechen Argumente wie die leicht erweiterbare Plugin-Architektur, leistungsfähige APIs und die Vielzahl definierter Erweiterungspunkte. Zudem bietet Moodle bereits umfassende Unterstützung für kompetenzbasierte Szenarien an, auch wenn hier noch Verbesserungsbedarf besteht. Auch aus „praktischen Gründen" ist Moodle eine sinnvolle Wahl, weil es an der *FernUniversität in Hagen* (*FUH*) [Fe21] als standarmäßiges LMS eingesetzt wird. Diese Arbeit und die meisten der damit verbundenen Anwendungsszenarien sind an der FUH verortet.

Zur Planung und Durchführung von QBL Anwendungsszenarien wird, insbesondere im Kontext von RG4, eine geeignete IT-Infrastruktur benötigt. Eine solche bietet die auf Fernlehre spezialisierte FUH, eine staatlich anerkannte, akkreditierte, aus öffentlichen Mitteln finanzierte Universität, die mit ca. 75.000 Studierenden die größte Universität im deutschsprachigen Raum ist. Einen Überblick über die Systemlandschaft der FUH vermittelt [Th20] in den Abschnitten 2.5.3, 2.5.4, 3.3.2 und 3.3.3.

3 Konzeption des QBL Modells

Das QBLM beinhaltet drei Komponenten, die im Folgenden kurz erläutert werden. Dabei handelt es sich um ein *Domänenmodell* (engl. *Domain Model, QDM*), ein *Architekturmodell* (engl. *Architectural Model*) und diverse *Servicemodelle* (engl. *Service Distribution Models*), die sich auf konkrete Anwendungsszenarien beziehen.

3.1 QBLM Domänenmodell

Wie in Kap. 2 bereits erwähnt, kann QBL als Weiterentwicklung der aus dem TENCompetence Projekt hervorgegangenen Ansätze verstanden werden. Das gilt auch für das QDM, das in Abb. 1 (entspricht Figure 54 in [Th20]) in Form eines UML Klassendiagramms dargestellt ist. Die blau, rot, orange und grau eingefärbten Elemente wurden direkt aus dem TENCompetence Domänenmodell übernommen, die von QBL beigesteuerten Erweiterungen sind in grüner Farbe eingezeichnet. Die blauen Elemente repräsentieren Lehr-/Lerninhalte, die orangefarbenen studentische Aktionen, Ziele und erhaltene Bewertungen. Die grauen Klassen stehen für das Kompetenzmodell, auf das sich die meisten QBL-spezifischen Erweiterungen beziehen. Auf Grundlage des QBLM sollen die nachfolgend beschriebenen Konzepte realisiert werden.

Lehrveranstaltungen wie Online-Kurse werden als *Lerneinheiten* verstanden, die sich aus *Lernaktivitäten* und *Wissensressourcen* zusammensetzen. Jedes dieser Elemente trägt seinen Teil zum CQ-basierten Lernziel des Kurses bei. Lernaktivitäten und Wissensressourcen werden als unabhängige Elemente mit eigenen Lernzielen und Zugangsvoraussetzungen realisiert, was sowohl die modulare Erstellung CQ-basierter Kurse, als auch die Definition von Bearbeitungssequenzen, d.h. CQ-basierter *Lernpfade*, begünstigt. Um zu verdeutlichen, dass für alle der im QDM blau dargestellten Elemente die Definition CQ-basierter Lernziele und Zugangsvoraussetzungen ermöglicht werden soll, führt QBL den Begriff des *Qualifikationsrelevanten Lernelements* (*QRLE*) ein, das im QDM von der Klasse *Lernelement QRLE* repräsentiert wird. Diese steht in einer Generalisierungsbeziehung zu den blauen Klassen, eine QRLE-Instanz kann also de facto für einen Studiengang, ein Modul einen Kurs, eine Lerneinheit, eine Lernaktivität oder eine Wissensressource stehen.

CQ-basierte Lernziele und Zugangsvoraussetzungen werden in QBL in Form von *CQ Profilen* abgebildet. Um ein Maximum an Vergleichbarkeit von CQ-bezogenen Daten zu erreichen, propagiert QBL die konsequente Verwendung standardisierter CQFs und institutionsspezifischer CQ Kataloge. Um dabei allerdings die Möglichkeiten beim Entwurf innovativer Lehr-/Lernzenarien nicht unnötig einzuschränken, muss auch der Entwurf eigener kurs-, modul- oder studiengangsspezifischer CQ Kataloge unterstützt werden. Technisch gesehen handelt es sich bei solchen lokalen CQ Katalogen ebenfalls um CQFs, im QDM werden diese daher, wie auch die standardisierteb CQFs, über die Klasse *CQ Rahmenwerk* abgebildet. Um auch Beziehungen zwischen CQs verschiedener CQFs und Verbindungen zu Elementen beliebiger Domänentaxonomien zu ermöglichen, wurde über die Klassen *Schlagwort* und *Semantisches Schlagwort* ein Tagging-Mechanismus eingeführt.

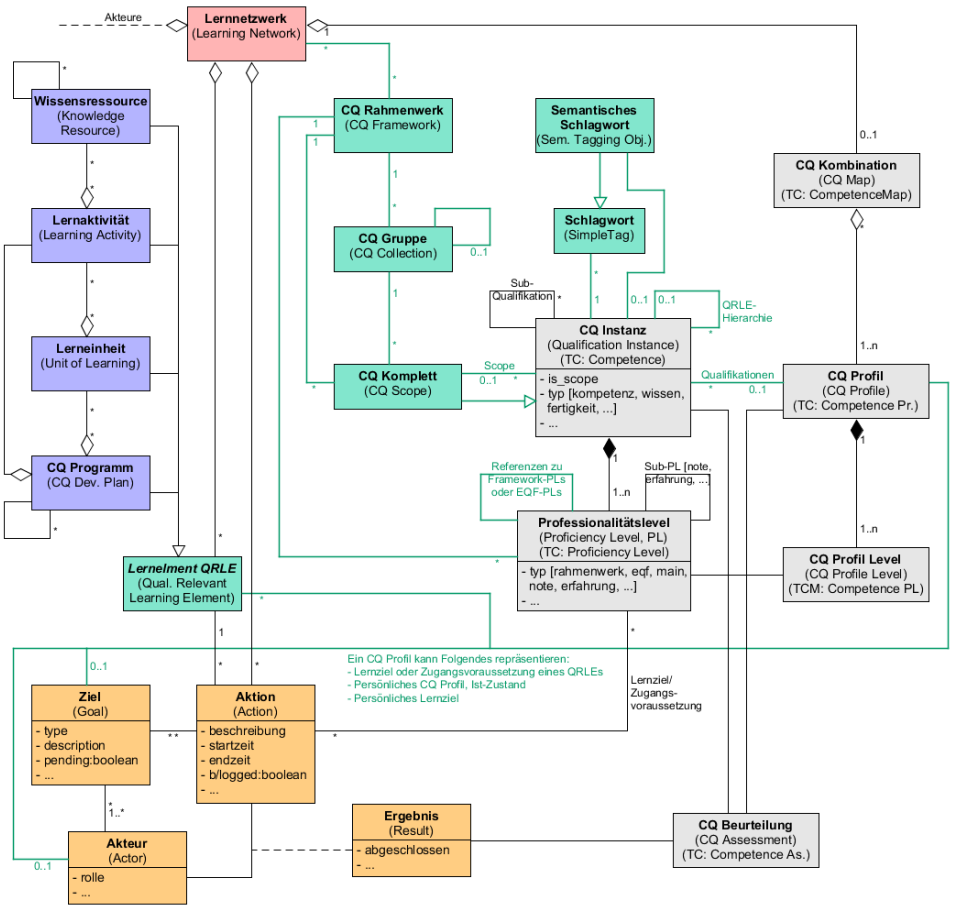


Abb. 1: QBLM Domänenmodell (QDM)

Die CQ-basierten Lernziele/Zugangsvoraussetzungen eines jeden QRLEs werden über CQ Profile spezifiziert. Gleiches gilt für das persönliche Lernziel (Soll-Profil) und den aktuellen Stand nachgewiesener CQs (Ist-Profil) eines/einer jeden Nutzer*in. Nach erfolgreicher Bearbeitung eines QRLEs wird das Ist-Profil der/des Nutzer*in aktualisiert. Die Differenz zwischen dem Ist- und dem Soll-Profil gilt es, durch geeignete *CQ Programme* zu überbrücken, die geeignete QRLEs wie beispielsweise Kurse, Lernaktivitäten und Wissensressourcen beinhalten. Das persönliches Lernziel ist erreicht, sobald das Ist-Profil dem Soll-Profil entspricht.

Eine detailliertere Beschreibung der Hintergründe und des Mehrwerts der einzelnen QDM-Elemente findet man in den Kapiteln 3.2, 4.1 und 5.2 in [Th20].

3.2 QBLM Architekturmodell und Servicemodelle

Das QBLM Architekturmodell konzentriert sich auf *Anwendungsbereiche* und *Softwarekomponenten*, die von der QBL-Einführung betroffen sind. Es lässt sich unterteilen in eine FUH-spezifische und eine allgemeine Variante. In beiden Fällen wurde zunächst der Ist-Zustand modelliert und später um zusätzliche Komponenten und Beziehungen ergänzt, die zur Einführung von QBL benötigt werden. Die im Rahmen dieser Dissertation behandelten Anwendungsszenarien beziehen sich auf die FUH-spezifische Variante im Sollzustand, die in Fig. 64 in [Th20] zu sehen ist. Der in Abb. 2 dargestellte Ausschnitt zeigt den Anwendungsbereich *Lernanwendungen*, der für die nachfolgend beschriebenen Anwendungsszenarien von zentraler Bedeutung ist. Bei den grünen Elementen handelt es sich um Softwarekomponenten, die in die Anwendungsszenarien involvierten Komponenten und Interaktionen sind farblich hervorgehoben (helles grün, rote Umrandung). Umfassende Beschreibungen aller Komponenten und Anwendungsbereiche findet man in [Th20], Kap. 2.5 und 3.3.

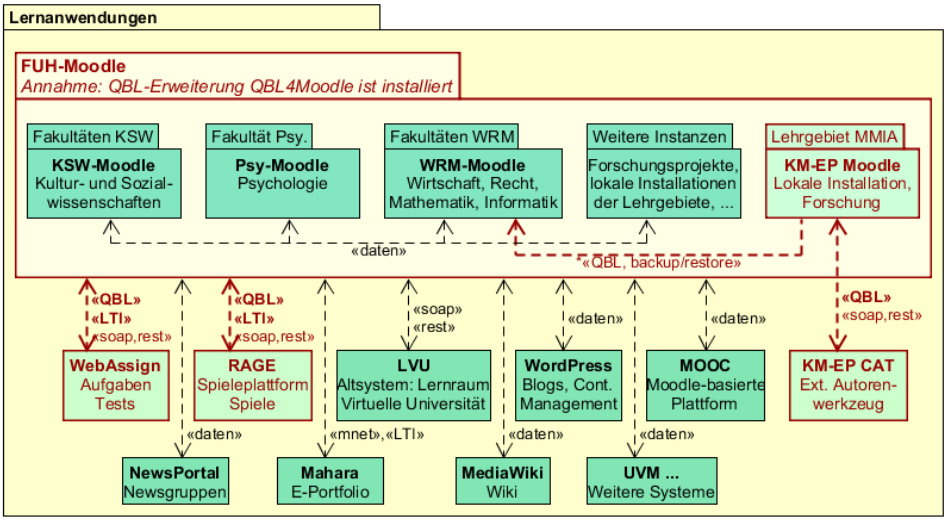


Abb. 2: Ausschnitt aus dem FUH-spezifischen Architekturmodell

Servicebasierte Interaktionen zwischen den einzelnen Komponenten werden im QBLM Architekturmodell (beispielsweise Abb. 2 und Fig. 64 in [Th20]) durch mit ⟨daten⟩ beschriftete Beziehungen angedeutet. Diese wurden in Form von QBLM Servicemodellen weiter spezifiziert, jedes davon bezieht sich auf ein konkretes Anwendungsszenario an der FUH. Die Anwendungsszenarien werden im Folgenden kurz skizziert, die zugehörigen Servicemodelle findet man in [Th20], Kap. 3.3.5, 3.3.6 und 3.3.7.

- **AS1:** *QBL-Plugin für Moodle*. Das erste Anwendungsszenario konzentriert sich auf das LMS Moodle, zu realisieren ist die im Rahmen von RG3 geforderte QDM-basierte LMS-Erweiterung. Dabei gilt es, möglichst viele der in Kap. 3.1 beschriebenen Konzepte umzusetzen.
- **AS2:** *LTI-basierte Moodle-WebAssign-Integration*. Das an der FUH entwickelte Aufgaben- und Testsystem *WebAssign* [Sc21a] ist fester Bestandteil der Systemlandschaft und wird inzwischen auch für Online-Klausuren eingesetzt. Zu entwickeln sind LTI-kompatible, servicebasierte Interaktionsmechanismen, die auch den Austausch CQ-bezogener Daten ermöglichen.
- **AS3:** *LTI-basierte Moodle-Lernspiel-Integration*. Der im Rahmen von AS2 entworfene Interaktionsmechanismus soll auf kompetenzbasierte Lernspiele übertragen werden. Die Rolle des an Moodle angebundenen Aufgabensystems übernehmen dabei Lernspiele oder Spieleplattformen.
- **AS4:** *Verwendung externer Autorenwerkzeuge*. Kurse, Module und Studienpläne, die mit einem externen Autorenwerkzeug (in Abb. 2: *KM-EP-CAT* [Wa18]) erzeugt wurden, sollen nach Moodle exportiert werden. Dieses Szenario war Bestandteil der Dissertation von Benjamin Wallenborn [Wa18]. Hier in AS4 soll der Funktionsumfang dieses Transfers erweitert werden.

Für AS2 und AS3 sind die in Abb. 2 mit ⟨LTI⟩ beschrifteten Verbindungen von besonderer Bedeutung, weil hier in hohem Maß auf einen LTI-basierten Datenaustausch zwischen dem LMS (Moodle) und den externen Systemen (WebAssign bzw. Spielekomponenten) gesetzt wird. Abb. 3 bezieht sich auf AS2 und zeigt eine schematische Darstellung der servicebasierten Interaktionen zwischen Moodle und WebAssign.

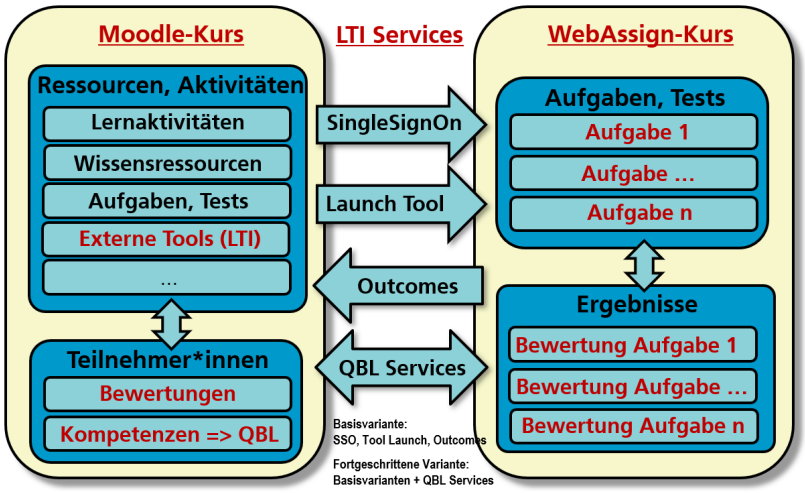


Abb. 3: LTI-basierte Moodle-WebAssign-Integration

Unter Verwendung des in Moodle standarmäßig vorhandenen Aktivitätstyps *Externes Tool* werden in einen Moodle-Kurs zugriffsgeschützte Aufgaben aus WebAssign nahtlos eingebunden. Nahtlos bedeutet, dass der/die in Moodle eingeloggte Nutzer*in beim Aufruf einer Aufgabe zunächst automatisch in WebAssign authentifziert wird (LTI-Service *Single Sign On*), im Erfolgsfall wird dann die angeforderte Ressource geladen (LTI-Service *Tool Launch*). Ob diese dann in einem separaten Browserfenster angezeigt oder im Moodle-Kurs direkt eingebettet werden soll, kann Moodle-seitig konfiguriert werden. Die bearbeiteten Aufgaben und die erhaltenen *Bewertungen* (Punktzahlen, Korrekturanmerkungen, etc.) werden in WebAssign gespeichert. Zusätzlich wird der LTI-Service *Return of Outcomes* verwendet, um die erreichten Punktzahlen an den Moodle-Kurs zu übergeben. Die Moodle-Aktivität Externes Tool stellt dann sicher, dass diese den *Bewertungen* (engl. *Gradebooks*) der Teilnehmer*innen hinzugefügt werden.

Bereits diese *Basisvariante* der LTI-basierten Moodle-WebAssign-Integration bietet die Möglichkeit, Studierenden in Abhängigkeit von den erhaltenen Bewertungen CQs zu attestieren. Man muss hierfür bei der Einrichtung der jeweiligen Externes-Tool-Aktivität die Abschnitte Bewertungen, Aktivitätsabschluss und Kompetenzen in geeigneter Weise konfigurieren und miteinander verknüpfen. Zusätzliche Möglichkeiten bietet die *fortgeschrittene Variante*, die bei der Rückgabe von Ergebnissen nicht auf den Outcomes Service begrenzt ist, der sich auf numerische Werte beschränkt. Über das in Abb. 3 mit *QBL Services* bezeichnete Servicepaket wird ein umfassenderer Datenaustausch ermöglicht.

Bei der Basisvariante werden Moodle-seitige Erweiterungen nur dann benötigt, wenn man sich bei der Kompetenzvergabe nicht nur im Rahmen der von Moodle standardmäßig angebotenen Tools bewegen und stattdessen QBL-Szenarien realisieren möchte. Letzteres setzt voraus, dass auf der verwendeten Moodle-Instanz das im Rahmen von AS1 entwickelte Moodle-Plugin installiert ist. In WebAssign war zum Zeitpunkt der Planung dieses Anwendungsszenarios noch keinerlei LTI-Unterstützung vorhanden, diese wurde erst im Rahmen dieser Dissertation implementiert. Bei der fortgeschrittenen Variante werden in beiden Systemen Erweiterungen benötigt.

4 Prototypische Implementierungen und deren Evaluation

Im Rahmen von AS1 bzw. RG3 wurde das Moodle-Plugin *QBL4Moodle* entwickelt, eine detaillierte Dokumentation ist in [Th20], Kap. 4.1 und in [THH19] zu finden. Beim Entwurf des Datenbankschemas und bei der Implementierung wurde konsequent über die angebotenen APIs und Erweiterungspunkte gearbeitet. Die funktionalen Anforderungen bezogen sich auf QBL-Basisfunktionalitäten wie Unterstützung standardisierter und nichtstandardisierter CQFs, Spezifikation CQ-basierter Lernziele/Zugangsvoraussetzungen von QRLEs, Zuweisung erreichter CQs an die Studierenden, sowie Realisierung persönlicher CQ Profile. Dazu kommen noch die im Kontext von AS2-AS4 benötigten Erweiterungen. QBL4Moodle wurde in verschiedenen Evaluationsszenarien unter Beteiligung von Wissenschaftlern, Entwicklern und Kursautoren mit positivem Ergebnis validiert. Kritisiert wurde, dass von den Nutzer*innen eine umfassende Einarbeitung erwartet wird. Dies ist nicht nur der Software geschuldet, sondern auch der Komplexität des QDM. Es gilt da-

her, neben der Implementierung zusätzlicher QBL-Funktionalitäten auch an der Benutzerfreundlichkeit zu arbeiten (Benutzerführung, Wizards, weniger textlastige Einarbeitungsdokumente). Die positive Bewertung von QBL4Moodle lässt sich auch daran ablesen, dass es derzeit im Rahmen von Nachfolgeprojekten verwendet und weiterentwickelt wird.

Die im Rahmen dieser Dissertation entstandene Basisvariante der LTI-basierten Moodle-WebAssign-Integration (AS2) ist seit dem Wintersemester 2017/18 fester Bestandteil der IT-Landschaft der FUH und wird inzwischen auch für Online-Klausuren eingesetzt. Eine Anwenderdokumentation (Titel: „Einbinden von Übungssystem-Aufgaben als LTI-Tool in Moodle“) findet man in [Sc21a] im Abschnitt „Handbücher für Kursbetreuer“. Eine technische Dokumentation zur Implementierung ist in [Sc21b] verfügbar, zudem wird das Thema in [Th20], Kap. 4.2 ausführlich behandelt.

Die Umsetzung von AS3 wurde im Rahmen dieser Dissertation in die Wege geleitet, erste prototypische Umsetzungen werden in [Th20], Kap. 4.3 beschrieben. Das Anwendungsszenario wird im Rahmen von nachfolgenden Forschungsarbeiten weiterverfolgt, erste Veröffentlichungen der damit befassten Doktoranden liegen bereits vor. Gleiches gilt für die Realisierung und Weiterentwicklung von AS4.

5 Fazit und Ausblick

Abschließend lässt sich sagen, dass der Themenkomplex QBL auf breites Interesse gestoßen ist. Das zeigt sich auch daran, dass derzeit mehrere nachfolgende Forschungs- und Entwicklungstätigkeiten im Gange sind, zu denen bereits Publikationen vorliegen. Bei den im Rahmen dieser Dissertation entstandenen Softwarelösungen ist hervorzuheben, dass die Basisvariante der LTI-basierten Moodle-WebAssign-Integration in den Produktionsbetrieb der FUH übernommen wurde und zunehmend Eingang in die Lehr-, Lern- und Prüfungsszenarien der Lehrgebiete findet. Die anderen hier erwähnten Forschungsprototypen sind auf einem guten Weg und werden kontinuierlich weiterentwickelt, mehr dazu siehe [Th20], Kap. 6.3. Zusammenfassend lässt sich feststellen, dass die Forschungsziele dieser Dissertation erreicht wurden.

Literaturverzeichnis

[Eu08] Europäische Union: Recommendation of the European Parliament and of the Council of 23 April 2008 on the establishment of the European Qualifications Framework for lifelong learning. In: Official Journal of the European Union, S. 1-7, 06.05.2008.

[Eu14] European Committee for Standardization CEN: European e-Competence Framework (e-CF) version 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. 2014.

[Fe21] FernUniversität in Hagen: Startseite der FernUniversität in Hagen - die größte Hochschule in Deutschland. https://www.fernuni-hagen.de, 01.02.2021.

[Im21] IMS Global Learning Consortium: Learning Tools Interoperability. https://www.imsglobal.org/activity/learning-tools-interoperability, 01.02.2021.

[Ko08] Koper, Rob: The TENCompetence Domain Model - Version 1.1. 2008. https://core.ac.uk/download/pdf/55533686.pdf.

[Ko09] Koper, Rob: Learning Network Services for Professional Development. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00977-8. DOI 10.1007/978-3-642-00978-5.

[KT06] Koper, Rob; Tattersall, Colin: Learning Design - A Handbook on Modelling and Delivering Networked Education and Training. Springer Berlin Heidelberg, ISBN 978-3-540-27360-8, 2006.

[Mo21] Moodle.org: Moodle open-source learning platform. https://moodle.org, 01.02.2021.

[Sc21a] Schulz-Gerlach, Immo: Hilfe/Handbücher - Online-Übungssystem. https://online-uebungssystem.fernuni-hagen.de/hilfe/hilfe.html, 01.02.2021.

[Sc21b] Schulz-Gerlach, Immo: Technische Doku: LTI-Integration - Das Online-Übungssystem als LTI-Tool-Provider. https://online-uebungssystem.fernuni-hagen.de/download/LTI_Moodle/LTI_TechnicalDoc.html, 01.02.2021.

[Th20] Then, Matthias: Supporting Qualifications-Based Learning (QBL) in a Higher Education Institution's IT-Infrastructure. Dissertation, FernUniversität in Hagen, März 2020. DOI 10.18445/20200309-141118-0. https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001608.

[THH19] Then, Matthias; Hoang, Minh Duc; Hemmje, Matthias: A Moodle-based software solution for Qualifications-Based Learning (QBL). FernUniversität in Hagen, 25 Feb. 2019. DOI: 10.18445/20190225-103757-0. https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001501.

[Wa18] Wallenborn, Benjamin: Entwicklung einer innovativen Autorenumgebung für die universitäre Fernlehre. Dissertation, FernUniversität in Hagen, September 2018. DOI: 10.18445/20180911-091907-0. https://ub-deposit.fernuni-hagen.de/receive/mir_ mods_00001428.

Matthias Then wurde am 27. März 1969 in Würzburg geboren und begann seine berufliche Laufbahn zunächst als Bauingenieur. Zwischen 2000 und 2003 absolvierte er an der FUH ein berufsbegleitendes Zusatzstudium in praktischer Informatik. Im Anschluss daran war er bis Juli 2009 als wissenschaftlicher Mitarbeiter am Lehrgebiet Software Engineering der FUH beschäftigt. Mit dem Themengebiet digitale Fernlehre ist er daher sowohl aus der Studenten-, als auch aus der Betreuer- und Autorenperspektive vertraut. Nach Ablauf der zulässigen Beschäftigungsdauer wechselte er ins Zentrum für Medien und IT an der FUH, wo er bis heute im Status eines wissenschaftlichen Mitarbeiters als Softwareentwickler und -architekt arbeitet. Zu seinen Aufgaben gehört die Weiterentwicklung der Lehr-/Lernplattformen, insbesondere der FUH-eigenen Moodle-Version, die für die speziellen Anforderungen der FUH optimiert ist. Als langjähriges FUH-Mitglied hat er sowohl ein persönliches, als auch ein berufliches Interesse an digitaler Lehre, innovativer Lehr-/Lernsoftware, dem Themenkomplex Lebenslanges Lernen und ganz allgemein an Digitalisierungsstrategien im Hochschulbereich. Es war daher naheliegend, sich für ein Promotionsprojekt aus diesem Bereich zu entscheiden. Am 04. März 2020 schloss er sein im Dezember 2014 begonnenes Promotionsvorhaben erfolgreich ab.

Algorithmik der Identifikation von Kausalen Effekten in Graphischen Modellen[1]

Benito van der Zander[2]

Abstract: Graphische, kausale Modelle repräsentieren Zufallsvariablen mitsamt ihren gegenseitigen Einflüssen als Graphen, und können die Ergebnisse von Experimenten aus rein beobachteten Daten vorhersagen. Diese Modelle haben große Bedeutung in Forschungsbereichen wie Epidemiologie, der Wirtschaftswissenschaft und der Sozialwissenschaft, in denen randomisierte kontrollierte Studien unmöglich sind oder unethisch wären, jedoch große Datenmengen zur Verfügung stehen. Obwohl graphische, kausale Modelle schon intensiv erforscht wurden, sind die meisten Ergebnisse theoretischer Natur und es fehlen effiziente Algorithmen, um die Modelle in künstlicher Intelligenz oder zur Analyse von Big Data anzuwenden.

In meiner Dissertation habe ich zwei Methoden untersucht, um die kausalen Effekte von Experimenten aus gegebenen beobachteten Daten und dem dazugehörigen graphischen kausalen Model zu berechnen: das Adjustieren für Störfaktoren in nicht-parametrisierten Systemen und die Instrumentvariablenmethode in linearen Systemen. Für beide Ansätze habe ich innovative Polynomialzeitalgorithmen entwickelt; abgesehen von einigen Situationen, für die ich gezeigt habe, dass das Problem der Berechnung NP-vollständig ist. Die vorgeschlagenen algorithmischen Methoden haben die bisher bekannten Verfahren wesentlich verbessert. Sie wurden in der weitverbreiteten Open-Source-Software DAGitty implementieren.

1 Einführung

Die Standardmethode, um die Auswirkungen von etwas wie einer Intervention oder einem Medikament zu untersuchen, ist es ein Experiment durchzuführen, bei dem Subjekte zufällig in eine Behandlungsgruppe und eine Kontrollgruppe aufgeteilt werden. Die Intervention wird nur in der ersten Gruppe durchgeführt, so dass die Effekte in beiden Gruppen miteinander verglichen werden können. Jedoch lassen sich viele wichtige Fragen nicht mit solchen randomisierten kontrollierten Studien beantworten, zum Beispiel „Verursacht Rauchen Lungenkrebs?“, „Verlangsamen nächtliche Ausgangssperren die Ausbreitung von Coronaviren?“, „Stammt die Klimaerwärmung von menschgemachten CO2-Emissionen?“, oder „Führen niedrigere Steuern zu mehr Wirtschaftswachstum?“. Für solche Fragen wären die nötigen experimentellen Studien unethisch, zu teuer, oder geradezu unmöglich. So wäre es ethisch nicht akzeptable Nichtraucher einer Rauchergruppe zuzuweisen; während einer Pandemie kann man nicht abwarten, bis die Effektivität der Maßnahmen ermittelt wurde; und die meisten Wissenschaftler haben nicht mehrere Länder und Planeten zur Verfügung, die sie für ihre Studien kontrollieren könnten.

Daher müssen solche Fragen aus rein beobachteten Daten beantwortet werden. Graphische, kausale Modelle sind eine wichtige Methode um beobachtete Daten entsprechend

[1] Englischer Titel der Dissertation: ”Algorithmics of Identifying Causal Effects in Graphical Models” [vdZ]

[2] Institut für Theoretische Informatik, Universität zu Lübeck, benito@tcs.uni-luebeck.de

zu analysieren, in Gebieten wie der Wirtschaftswissenschaft [AP08, Im14], Sozialwissenschaft [El13] und Epidemiologie [RGL08]. Solche Modelle repräsentieren Zufallsvariablen als Knoten in einem Graph und die Abhängigkeiten zwischen Variablen als Kanten.

Das wichtigste Modell sind azyklische gerichtete Graphen (DAGs), bei denen die Variablen von ihren Eltern abhängen [Pe09]. Ändert sich der Wert einer Variable, so beeinflusst sie direkt die Werte ihrer Kinder, welche wiederum ihre Kinder beeinflussen und so fort. Wir betrachten DAGs mit n Knoten (Zufallsvariablen) $\mathbf{V} = \{V_1, \ldots, V_n\}$. Für jede Variable V_i gibt es die bedingte Wahrscheinlichkeit $P_i(V_i = v_i \mid pa_i)$, dass Variable V_i den Wert v_i in Abhängigkeit der Werte pa_i aller Elternknoten $Pa(V_i)$ annimmt. Diese Wahrscheinlichkeitsverteilung $P_i(V_i = v_i \mid pa_i)$ kann beliebig gewählt werden, und auch der Wertebereich der Zufallsvariablen spielt keine Rolle. Zu beachten ist nur, dass P_i lokal ist, also von anderen Variablen als $V_i \cup Pa(V_i)$ nicht beeinflusst wird. Wir wollen Eigenschaften ermitteln, die für alle solche Wahrscheinlichkeitsverteilungen gelten, und nur von der Struktur des Graphens abhängen[3].

Aus diesen lokalen Verteilungen einzelner Variablen folgt eine Wahrscheinlichkeitsverteilung aller Variablen $P(V_1 = v_1, \ldots, V_n = v_n)$, kurz $P(\mathbf{v})$, als Produkt aller Verteilungen: $P(\mathbf{v}) = \prod_{i=1}^{n} P_i(v_i \mid pa_i)$. Üblicherweise wird der Index i der Faktoren weggelassen und $P_i(v_i \mid pa_i)$ als $P(v_i \mid pa_i)$ geschrieben, denn man kann P_i als Einschränkung der Verteilung P auf die Variable V_i und ihre Eltern betrachten.

Als ein Beispiel kann man mit dem Graph in Abbildung 1(a) untersuchen, ob das Risiko an Diabetes (Variable D) zu erkranken durch verbesserte Schulbildung (Variable *LE* „low education“) gesenkt werden kann [TL11, vdZLT19, RGL08, Kapitel 12]. In diesem Modell gibt es genau drei Ursachen des Diabetesrisikos: Das genetische Diabetesrisikos der Mutter (*MR*), ob sie Diabetes entwickelt hat (*MD*), und die Bildungsstufe (*LE*). Weder *MR* noch *MD* beeinflussen *LE*. Eine vierte Variable, Familieneinkommen (*FI*), beeinflusst *MD* und *LE*, hat jedoch keinen direkten Effekt auf *MR* oder *D*. Sie beeinflusst jedoch *D* über *MD* indirekt. Die zum Graph gehörende Wahrscheinlichkeitsverteilung ist dann $P(fi, mr, md, le, d) = P(fi)P(mr)P(md \mid fi, mr)P(le \mid fi)P(d \mid md, mr)$.

Eine bekannte Anwendung dieser graphischen Modelle, ohne jeglichen Kausalitätsbezug, ist es, dass man bedingte Unabhängigkeiten aus dem Graphen ablesen kann, welche für alle solche faktorisierte Wahrscheinlichkeitsverteilungen gelten. Mengen von Variablen $\mathbf{X}$ und $\mathbf{Y}$ sind nämlich unabhängig, also $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$, wenn es keinen nicht-blockierten Pfad von einem Knoten $X \in \mathbf{X}$ zu einem Knoten $Y \in \mathbf{Y}$ gibt. Ein *Pfad* ist eine Sequenz von Knoten $V_{i_1}, V_{i_2}, \ldots, V_{i_k}$, so dass V_{i_j} und $V_{i_{j+1}}$ durch eine Kante verbunden sind. Der Pfad ist blockiert (von der leeren Menge), wenn er $V_{i_{j-1}} \to V_{i_j} \leftarrow V_{i_{j+1}}$ enthält[4]. In dem Fall wird der Knoten V_{i_j} *Kollider* (auf diesem Pfad) genannt. Im Beispiel 1 (a) sind *FI* und *MR* sowie *LR* und *MR* unabhängig, jedoch keine anderen Paare von Variablen[5].

[3] präziser gesagt, betrachtet der erste Teil der Dissertation allgemeine Verteilungen und der zweite Teil einen Spezialfall der Verteilungen in linearen Modellen.

[4] Die in einem nicht-blockierten Pfad erlaubten Kantenfolgen sind also $V_{i_{j-1}} \to V_{i_j} \to V_{i_{j+1}}$, $V_{i_{j-1}} \leftarrow V_{i_j} \leftarrow V_{i_{j+1}}$ und $V_{i_{j-1}} \leftarrow V_{i_j} \to V_{i_{j+1}}$.

[5] Der Zusammenhang zwischen Pfaden und Unabhängigkeit gilt nun erstmal nur in eine Richtung. Für alle entsprechend faktorisierbaren Wahrscheinlichkeitsverteilung sind *FI,MR* sowie *LE,MR* unabhängig. Für alle an-

Für die *bedingte Unabhängigkeit* ist zusätzlich eine Menge von Variablen $\mathbf{Z}$ gegeben. Die Variablenmengen $\mathbf{X}$ und $\mathbf{Y}$ sind bedingt unabhängig gegeben $\mathbf{Z}$, also $P(\mathbf{X},\mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Y} \mid \mathbf{Z})$, wenn $\mathbf{Z}$ alle Pfade zwischen $X \in \mathbf{X}$ und $Y \in \mathbf{Y}$ blockiert. Die Menge $\mathbf{Z}$ invertiert nun die Definition der Blockiertheit. Der Pfad ist blockiert, wenn er für ein $V_{i_j} \in \mathbf{Z}$ die Kanten $V_{i_{j-1}} \to V_{i_j} \to V_{i_{j+1}}$, $V_{i_{j-1}} \leftarrow V_{i_j} \leftarrow V_{i_{j+1}}$ oder $V_{i_{j-1}} \leftarrow V_{i_j} \to V_{i_{j+1}}$ enthält. Ebenso ist er blockiert, wenn er $V_{i_{j-1}} \to V_{i_j} \leftarrow V_{i_{j+1}}$ enthält und *kein* Nachfahre von V_{i_j} in $\mathbf{Z}$ ist.

Im Beispiel 1 (e) kann also mit $\mathbf{Z} = \{MD\}$ der Kollider *MD geöffnet* werden, so dass der Pfad $FI \to MD \leftarrow MR$ von *MD* nicht blockiert ist. Somit sind *FI* und *MR* nicht unabhängig gegeben *MD*. Es ergibt auch intuitiv Sinn, am Anfang können sich *FI* und *MR* über *MD* nicht beeinflussen, da *MD* nur eine gemeinsame Folge von beiden ist. Ist der Wert von *MD* jedoch fest, so kann die Kenntnis von *FI* zu Wissen über *MR* führen, oder umgekehrt. Zum Beispiel ist das Einkommen hoch und die Mutter hat Diabetes, so ist es wahrscheinlicher, dass sie ein hohes genetisches Risiko besitzt. Ist das Einkommen dagegen niedrig und die Mutter hat nicht Diabetes, hat sie vermutlich auch kein besonderes hohes genetisches Risiko[6].

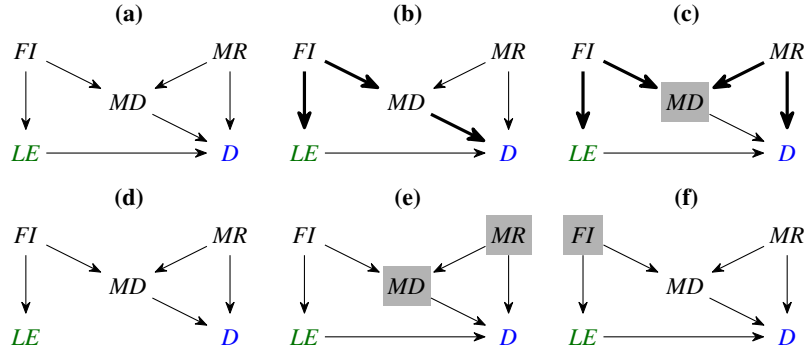


Abb. 1: Graphisches Modell, welches den Einfluss von Bildung (*LE*) auf das Diabetesrisiko (*D*) modelliert [TL11, vdZLT19, vdZ, RGL08, Kapitel 12]. Als Kovariablen gibt es das Familieneinkommen (*FI*), das genetische Diabetesrisikos der Mutter (*MR*), und ob die Mutter Diabetes hat (*MD*). Fall (a) zeigt den Graphen selbst. Fälle (b) und (c) zeigen jeweils einen nicht blockierten Pfad zwischen *LE* und *D*. Fall (d) zeigt den Graphen ohne gerichtete Pfade von *LE* zu *D*. Fälle (e) und (f) zeigen, wie man alle nicht-kausalen Pfade zwischen (*LE*) und (*D*) blockieren kann.

deren Paare gibt es Wahrscheinlichkeitsverteilungen, so dass diese nicht unabhängig sind. Die anderen Paaren sind also nicht *für alle* Wahrscheinlichkeitsverteilungen unabhängig. Es kann jedoch auch Wahrscheinlichkeitsverteilungen geben, für die alle Variablen unabhängig sind. Da die einzelnen Faktoren P_i von $P(\mathbf{v})$ beliebig gewählt werden können, können sie auch so gewählt werden, dass die Verteilung P_i nicht von den Eltern pa_i abhängt. Solche Verteilungen werden „non faithful“ genannt, spielen aber im Folgenden keine Rolle.

[6] Eine solche Korrelation von Ursachen nach Festhalten einer Folge kennen Statistiker auch als Berkson-Paradox.

2 Identifizieren des kausalen Effektes mittels Adjustierung

2.1 Der kausale Effekt

Die grundlegende Idee der kausalen Modellierung ist nun, dass aus der Faktorisierung der Verteilung $P(\mathbf{v})$ der Effekt eines Experimentes berechnet werden kann, ohne das Experiment tatsächlich durchzuführen[7]. Würde man ein Experiment durchführen, welches Variablen $\mathbf{X} \subseteq \mathbf{V}$ auf Werte $\mathbf{x}$ setzt, wären die Werte von $\mathbf{X}$ konstant $\mathbf{x}$, somit würden sie nicht mehr von ihren Eltern abhängen. Dies entspricht einfach dem Entfernen der Faktoren $P_{X_i}(x_i \mid pa_{X_i})$ für alle $X_i \in \mathbf{X}$ aus der Gesamtverteilung. Der Gesamteffekt der Variablen $\mathbf{X}$ ist also eine neue Wahrscheinlichkeitsverteilung $P(\mathbf{v} \mid do(\mathbf{x}))$, die Wahrscheinlichkeit, dass alle Variablen Werte $\mathbf{v}$ haben, nachdem ein Experiment Variablen $\mathbf{X}$ auf Werte $\mathbf{x}$ gesetzt hat:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \begin{cases} \prod\limits_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i \mid pa_i) & \text{für } \mathbf{v} \text{ konsistent zu } \mathbf{x}, \\ 0 & \text{sonst,} \end{cases} \tag{1}$$

wobei konsistent bedeutet, dass die Werte $\mathbf{v}$ dieselben Werte für die Variablen $\mathbf{X}$ enthalten wie $\mathbf{x}$. Man kann auch den kausalen Effekt $P(\mathbf{y} \mid do(\mathbf{x}))$ auf eine Teilmenge $\mathbf{Y} \subseteq \mathbf{V}$ betrachten, der sich aus der Verteilung aller Variablen durch Marginalisieren berechnen lässt: $P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{v} \setminus \mathbf{y}} P(\mathbf{v} \mid do(\mathbf{x}))$. Dabei summiert $\sum_{\mathbf{v} \setminus \mathbf{y}}$ über die möglichen Werte aller Variablen $\mathbf{V} \setminus \mathbf{Y}$. Diese Verteilung entspricht ungefähr der Verteilung eines DAGs in dem alle in $\mathbf{X}$ eingehenden Kanten entfernt wurden.

Sind alle Faktoren $P(v_i \mid pa_i)$ bekannt, so lässt sich der kausale Effekt $P(\mathbf{v} \mid do(\mathbf{x}))$ nach Formel (1) leicht berechnen. In der Praxis, kennt man nicht alle Faktoren, und die Variablen sind partitioniert in eine Menge von beobachteten, „observed“ Variablen $\mathbf{O}$ und eine Menge von unbeobachteten, latenten Variablen $\mathbf{V} \setminus \mathbf{O}$. Die Faktoren der unbeobachteten Variablen dürfen bei der Berechnung nicht verwendet werden. Im Beispiel würde man auf jeden Fall *LE* und *D* kennen, da diese Variablen untersuchten werden, und vermutlich wurde auch *MD* abgefragt. *FI* könnte Datenschutzrechtlich problematisch zu ermitteln sein, und daher zu den unbeobachteten Variablen zählen. *MR* ist schwer zu bestimmen, selbst wenn die Gene der Mutter sequenziert wurden, ist es erstmal unbekannt, welche Gene für Diabetes relevant sind.

2.2 Adjustierung

Das Ziel der Dissertation ist es nun den kausalen Effekt $P(\mathbf{y} \mid do(\mathbf{x}))$ zu identifizieren. Dabei ist der Graph und Variablenmengen $\mathbf{X}, \mathbf{Y}, \mathbf{O}$ gegeben, und wir suchen eine Formel, die äquivalent zu $P(\mathbf{y} \mid do(\mathbf{x}))$ ist und nur Wahrscheinlichkeitsverteilungen über den beobachteten Variablen und keinen *do*-Operator enthält. Wir betrachten nicht beliebige Formeln,

[7] Woraus Judea Pearl seine Kausalitätstheorie entwickelt hat, für die er einen Turing-Award gewonnen hat.

sondern im ersten Teil ausschließlich Formeln, die den Effekt mittels Adjustierung identifizieren. Hierbei wird eine Menge von (Stör)variablen $\mathbf{Z} \subseteq \mathbf{O} \subseteq \mathbf{V}$ gesucht, ein sogenanntes *Adjustment Set* $\mathbf{Z}$, für die gilt:

$$P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z}). \tag{2}$$

Diese Formel berechnet den Erwartungswert von der bedingten Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ über alle möglichen Werte der Variablen $\mathbf{Z}$. Adjustment Sets sind die am häufigsten verwendete Methode zur Bestimmung des kausalen Effekts, weil sie gut erforschte statistische Eigenschaften haben [Va09, GK17]. Im Wesentlichen werden die möglichen Werte in Teilpopulationen aufgeteilt, so dass innerhalb jeder Teilpopulationen die rein beobachtete Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ der durch ein Experiment entstehenden Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x}), \mathbf{Z} = \mathbf{z})$ entspricht.

Das Standardverfahren zum Wählen der Menge $\mathbf{Z}$ ist Pearls *Back-Door-Kriterium* [Pe93, Pe09]. Das Back-Door Kriterium sagt, $\mathbf{Z}$ darf keinen von Nachfahren $\mathbf{X}$ enthalten und muss alle Pfade, deren erste Kante auf einen Knoten in $\mathbf{X}$ zeigt, blockieren[8]. In den Graphen von Abbildung 1 und Abbildung 2 (a) können mit dem Back-Door Kriterium alle Adjustment Sets gefunden werden. Im Graphen 2 (b) funktioniert das Kriterium jedoch nicht, denn es gibt nur ein Adjustment Set bestehend aus Nachfahren von X_1.

Ein vollständiges Adjustment-Kriterium, welches von einer Menge $\mathbf{Z}$ genau dann erfüllt ist, wenn $\mathbf{Z}$ ein Adjustment Set ist, wurde erst von [SVR10] entdeckt. Dieses Kriterium sagt, $\mathbf{Z}$ darf keinen Nachfahren eines Knotens auf einem echten, kausalen Pfad von $\mathbf{X}$ zu $\mathbf{Y}$ enthalten[9], und muss alle echten, nicht-kausalen Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$ blockieren. Ein *echter* Pfad ist ein Pfad, der nur einen Knoten von $\mathbf{X}$ enthält (d.h. als Startknoten); ein *kausaler* Pfad ist ein gerichteter Pfad, der nur von $\mathbf{X}$ weg gerichtete Kanten enthält; und ein *nicht-kausaler* Pfad ist ein Pfad, der nicht kausal ist.

Die zu untersuchende, offene Frage ist, wie man effizient überprüft, ob das Kriterium von einer Menge erfüllt ist, und wie man eine solche Menge findet. Viele Implementierungen von Algorithmen für kausalen Graphen [Ky98, Ka12] enumerieren alle möglichen Mengen oder Pfade, und testen dann auf triviale Weise, ob ein Kriterium erfüllt ist. Dies ergibt jedoch eine exponentielle Laufzeit und ist für große Graphen nicht praktikabel.

Dafür haben wir ein neues vollständiges Kriterium eingeführt, welches wir *Konstruktives Back-Door Kriterium* nennen und welches das Adjustment-Kriterium ändert, damit der Begriff des nicht-kausalen Pfades nicht mehr vorkommt. Die erste Bedingung verbietet weiterhin Nachfahren von echt kausalen Pfaden. Die zweite Bedingung ist, dass $\mathbf{Z}$ alle Pfade im Echten Back-Door-Graphen blockieren muss. Der *Echte Back-Door-Graphen* wird konstruiert, indem jeweils die erste Kante aller kausalen Pfade von $\mathbf{X}$ zu $\mathbf{Y}$ gelöscht wird. Eine Menge $\mathbf{Z}$ blockiert dann alle nicht-kausalen Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$, genau dann, wenn sie alle Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$ im echten Back-Door-Graphen blockiert.

[8] Es gibt auch eine äquivalente Formulierung mittels einem Back-Door-Graphen, in dem alle aus $\mathbf{X}$ ausgehenden Kanten gelöscht werden.

[9] Nachfahren des Startknotens sind erlaubt, Nachfahren des Endknoten eines Pfades nicht.

Im Beispiel von Abbildung 1 betrachten wir $\mathbf{X} = \{LE\}$ und $\mathbf{Y} = \{D\}$. Dann liegt nur $\{D\}$ auf kausalen Pfaden und der (echte) Back-Door-Graph ist der Graph von Fall (d). Blockiert werden muss der in (b) markierte Pfad. Dies kann entweder mit dem Adjustment Set $\{FI\}$ wie in Fall (f) erfolgen, oder mittels Knoten *MD*. Da *MD* wiederum den Pfad aus Fall (c) öffnet, ist ein zweites mögliche Adjustment Set nur $\{MD, MR\}$. Im Beispiel von Abbildung 2 (b), ist der (echte) Back-Door-Graph in Fall (b') gezeigt. Hier müssen alle Pfade von Y_2 zu $\{X_1, X_2\}$ blockiert werden, so dass das einzige Adjustment Set $\{Z_1, Z_2\}$ ist.

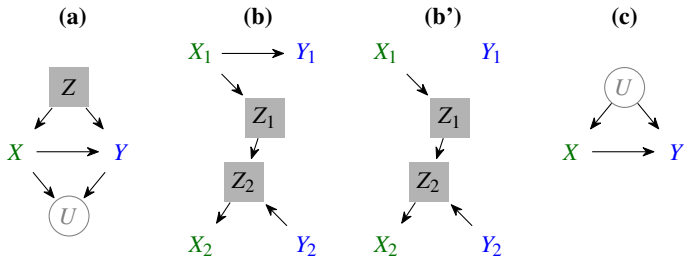


Abb. 2: Im Fall (a) kann der kausale Effekt von $\mathbf{X} = \{X\}$ auf $\mathbf{Y} = \{Y\}$ leicht mit den Eltern $\{Z\}$ von X als Adjustment Set identifiziert werden. Im Fall (b) kann der kausale Effekt von $\mathbf{X} = \{X_1, X_2\}$ auf $\mathbf{Y} = \{Y_1, Y_2\}$ mit dem Adjustment Set $\mathbf{Z} = \{Z_1, Z_2\}$ identifiziert werden. Graph (b') zeigt den dazugehörigen echten Back-Door-Graphen. Z_2 muss im Adjustment Set sein, um den Pfad zwischen X_2, Y_2 zu blockieren. Z_1 muss dann hinzugefügt werden, weil der Kollider Z_2 auf dem Pfad zu X_1 geöffnet wurde. Im Fall (c) ist es unmöglich den kausalen Effekt zu identifizieren. Das Adjustment Set müsste den Pfad über U blockieren, U ist jedoch als nicht beobachtet markiert und darf daher nicht im Adjustment Set sein.

Übrig bleibt das Problem eine Menge zu finden, welche alle Pfade in einem kausalen (Back-Door)-Graphen blockiert. Standardgraphtheoretische Algorithmen lassen sich hierfür nicht verwenden, weil die Definition mit Kollidern, welche erst blockiert sind, und dann geöffnet werden können, ungewöhnlich ist. Auch für kausale Graphen gab es zwar schon bekannte Algorithmen, welche überprüfen, ob eine Menge alle Pfade blockiert [Sh98], eine solche blockierende Menge finden, eine blockierende, minimale Menge finden [TPP98], eine Minimum-Menge finden [ADC96] oder alle diese Mengen aufzählen [TL11]. Wir mussten jedoch die Algorithmen erweitern, so dass sie Mengen $\mathbf{Z}$ unter der Einschränkung $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ für beliebige Mengen $\mathbf{I} \subseteq \mathbf{R}$ finden können. Zudem ist es uns gelungen, die Laufzeit für das Finden einer minimalen Menge von $\mathcal{O}(n^2)$ zu $\mathcal{O}(n+m)$ bezüglich der Zahl der Knoten n und Kanten m zu verbessern.

2.3 Adjustierung in generalisierten kausalen Modellen

DAGs sind die grundlegenden kausalen Modelle, aber sie setzen voraus, dass man alle relevanten Variablen und die Richtung aller Kanten kennt. In vielen Situationen fehlen diese Informationen jedoch. Daher haben wir alle unsere Ergebnisse und Algorithmen zu zwei weiteren Klassen von Graphen verallgemeinert: RCGs und MAGs.

Chain Graphen (CGs) enthalten zusätzlich ungerichtete Kanten, welche besagen, dass zwei Variablen einander direkt beeinflussen, jedoch unbekannt ist, welche der Variablen der Elternknoten und welche der Kindknoten ist [LW89, Fr90]. Ein Chain Graph repräsentiert eine Klasse von DAGs, und zwar alle DAGs, die durch das Ersetzen aller ungerichteten Kanten im CG durch gerichtete Kanten entstehen. Zudem müssen alle von einem CG repräsentierten DAGs dieselben (bedingten) Unabhängigkeiten haben, so dass sie statistisch ununterscheidbar sind. Nicht jede Ersetzung von ungerichteten Kanten ist erlaubt, zum Beispiel darf aus einem ungerichteten Zyklus kein gerichteter Zyklus entstehen, da sonst kein DAG entstünde. Auch darf bei der Ersetzung auf keinem Pfad zwischen zwei Variablen ein Kollider entstehen, außer es gibt es anderen Pfad der diese Variablen verbindet. Restricted Chain Graphs (RCGs) nennen wir diejenigen Chain Graphen, die mindestens einen DAG repräsentieren. Eine Unterklasse von RCGs sind CPDAGs, welche oft verwendet werden, da sie gut aus beobachteten Daten gelernt werden können.

Modelle mit unbekannten Variablen können durch einen *Ancestral Graph* dargestellt werden [RS02], der alle DAGs repräsentiert, welche dieselben Vorfahrrelationen wie der Ancestral Graph haben. Zum Beispiel, besagen die Vorfahrrelationen in Abbildung 1, dass *FI* ein Vorfahr von *D* und kein Vorfahr von *MR* ist. Wird der Graph als Ancestralgraph betrachtet, so repräsentiert er auch jeden DAG, der beispielsweise den Graph mit einer Variable erweitert, die zwei der bisherigen Variablen als Kinder hat. Es darf jedoch keine Variable eingefügt werden, die ein Kind von *FI* und ein Elternteil von *MR* ist, da diese Variable die Vorfahrrelationen ändern würde, indem *FI* ein Vorfahr von *MR* würde.

Maximale Ancestral Graphen (MAGs) sind Ancestral Graphen, bei denen alle Paare von Variablen, zwischen denen sich nicht alle Pfade blockieren lassen, direkt durch eine Kante verbunden sind. Wird der Graph aus Abbildung 1 als MAG betrachtet, dürfte man eine Variable mit Kindern *MR* und *D* einfügen. Eine Variable *U* mit Kindern *LE* und *D* darf jedoch nicht eingefügt werden. Dann gäbe es nämlich Pfade $FI \rightarrow LE \leftarrow U \rightarrow D$ und $FI \rightarrow LE \rightarrow D$. Der zweite Pfad kann nur an Knoten *LE* blockiert werden, wodurch sich der erste Pfad öffnen würde, der dann nicht mehr mit den bisherigen Variablen blockiert werden kann. Diese Variable *U* ließe sich jedoch einfügen, gäbe es eine Kante $FI \rightarrow D$. Anders als DAGs können MAGs auch Marginalisierung von Verteilung repräsentieren.

3 Identifizierung in linearen Modellen (SEMs)

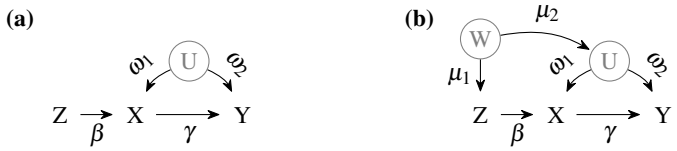


Abb. 3: (a): Das klassische Instrumentvariablenmodell. (b): *Z* ist eine bedingte Instrumentvariable. *U* and *W* sind unbeobachtete Variablen.

Sehr häufig werden kausale Modelle unter der Annahme verwendet, dass alle Variablen reelle Zahlen sind und alle Abhängigkeiten zwischen den Variablen linear sind [Bo89, Du75], so genannte Structural Equation Models (SEMs).

Zum Beispiel, repräsentiert der Graph in Abbildung 3 (a), vier Zufallszahlen mit je einer linearen Gleichung: $Z = \varepsilon_Z$; $U = \varepsilon_U$; $X = \beta Z + \omega_1 U + \varepsilon_X$; $Y = \gamma X + \omega_2 U + \varepsilon_Y$.
Jede Variable hängt von ihren Eltern und unabhängig gleichverteilten, zufälligen Errortermen $\varepsilon_\cdot$ ab[10]. Dies bedeutet beispielsweise, ändert man den Wert von Z um 1, dann ändert sich der (Mittel-)Wert von X um β. Das Modell ist kausal, ändert man den Wert von X, dann ändert sich der Wert von Z nicht. Die Abhängigkeiten der Variablen sind deterministisch, durch Addition der Zufallszahlen $\varepsilon_\cdot$ entstehen jedoch Zufallsvariablen. Geht man davon aus, alle Errorterms sind Gaußverteilt, dann sind die Faktoren $P_i(V_i \mid pa_i)$ in der Wahrscheinlichkeitsverteilung ebenfalls Gaußsche Normalverteilungen.

Aus diesem Modell lassen sich die Korrelation/Kovarianzen zwischen den Variablen quantitativ bestimmen, indem man die Parameter entlang der Pfade multipliziert und alle Pfade addiert. So ist die Kovarianz $\mathrm{Cov}(Z,X) = \beta$; $\mathrm{Cov}(Z,Y) = \beta\gamma$; und $\mathrm{Cov}(X,Y) = \gamma + \omega_1\omega_2$.

Die Kovarianzen zwischen allen Paaren von Variablen, sind dabei die beobachteten Daten. Das Ziel der Identifizierung in SEMs ist es die Parameter der Gleichungen aus den Kovarianzen zu ermitteln. Zum Beispiel folgt hier: $\gamma = \frac{\beta\gamma}{\beta} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)}$. Man sagt, Z ist eine *Instrumentvariable* zur Identifizierung des direkten kausalen Effekts von X auf Y.

Auf dieselbe Weise kann man den Effekt in Abbildung 3 (b) als $\gamma = \frac{\beta\gamma}{\beta} = \frac{\mathrm{Cov}(Y,Z|W)}{\mathrm{Cov}(X,Z|W)}$ bestimmen. Es müssen jedoch die mit W bedingten Kovarianzen verwendet werden, um den Pfad über W zu blockieren, damit das Ergebnis nicht von $\mu_1\mu_2$ verfälscht wird.

Es gibt ein (nicht vollständiges) Kriterium, um zu entscheiden, ob eine Variable Z eine (bedingte) Instrumentvariable zur Identifizierung des Effekts von X auf Y ist [Pe01]: und zwar muss es eine Menge $\mathbf{W}$ geben, die keine Nachfahren von Y enthält, so dass es einen nicht von $\mathbf{W}$ blockierten Pfad von Z zu X gibt und alle Pfade von Z zu Y ohne die Kante $X \to Y$ von $\mathbf{W}$ blockiert sind.

Wir haben dieses Kriterium untersucht, und gezeigt, dass es ein NP-vollständiges Problem ist $\mathbf{W}$ zu finden. Daraus folgt, gegeben X, Y und Z, so ist es auch NP-vollständig zu entscheiden, ob Z eine bedingte Instrumentvariable für den Effekt von X auf Y ist.

Trotzdem ist es, gegeben X und Y, in $\mathcal{O}(n(n+m))$ Polynomialzeit möglich, eine bedingte Instrumentvariable Z für den Effekt von X auf Y und eine Menge $\mathbf{W}$ zu finden. Hierzu haben wir gezeigt: (1) Existiert eine bedingte Instrumentvariable Y für X auf Y, so existiert auch eine (möglicherweise andere) bedingte Instrumentvariable Y, deren Menge $\mathbf{W}$ ausschließlich aus Vorfahren von Y und Z besteht. (2) Besteht $\mathbf{W}$ aus Vorfahren von Y und Z, so kann $\mathbf{W}$ in Zeit $\mathcal{O}(n+m)$ gefunden werden. Obwohl $\mathbf{W}$ keine minimale Menge sein muss, kann unser Algorithmus zum Finden von minimalen Mengen aus der vorherigen Sektion dazu wiederverwendet werden.

Schließlich haben wir weitere Arten von Instrumentvariablen untersucht, einzelne Instrumentvariablen, die sich ähnlich verhalten, und Instrumental Sets [Br10], bei denen mehrere Instrumentvariablen kombiniert werden. Instrumental Sets mit leerer Menge $\mathbf{W} = \emptyset$ lassen

[10] Alternativ wird häufig die unbeobachtete Variable U in den Gleichungen weggelassen. D.h. $Z = \varepsilon_Z$; $X = \beta Z + \varepsilon_X$; $Y = \gamma X + \varepsilon_Y$, mit der zusätzlichen Bedingung, dass die Kovarianz zwischen ε_X und ε_Y gleich $\omega_1\omega_2$ ist.

sich in Polynomialzeit finden, aber im Allgemeinen ist das Finden von Instrumental Sets NP-vollständig.

4 Fazit

Die Forschung der Dissertation ergab die ersten effizienten Algorithmen, die ein Adjustment Set oder ein minimales Adjustment Set finden können, genau dann wenn ein solches existiert. Es ergab auch das erste vollständige Kriterium zur Charakterisierung von Adjustment Sets in MAGs und RCGs. Zudem war es die erste genauere Untersuchung der Komplexität von Instrumentvariablen mit einem Algorithmus zum effizienten Finden von bedingten Instrumentvariablen. Die Algorithmen ermöglichen die automatische Analyse kausaler Zusammenhänge. Sie sind praktisch implementierbar, und wurden in der weitverbreiteten Open-Source-Software DAGitty veröffentlicht.

Literaturverzeichnis

[ADC96] Acid, Silvia; De Campos, Luis M: An algorithm for finding minimum d-separating sets in belief networks. In: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., S. 3–10, 1996.

[AP08] Angrist, Joshua D.; Pischke, Jörn-Steffen: Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2008.

[Bo89] Bollen, Kenneth A: Structural equations with latent variables. John Wiley & Sons, 1989.

[Br10] Brito, Carlos: Instrumental sets. Heuristics, Probability and Causality. A Tribute to Judea Pearl, S. 295–307, 2010.

[Du75] Duncan, Otis Dudley: Introduction to structural equation models. Academic Press, 1975.

[El13] Elwert, Felix: Graphical Causal Models. In: Handbook of Causal Analysis for Social Research, Handbooks of Sociology and Social Research, S. 245–273. Springer, 2013.

[Fr90] Frydenberg, Morten: The chain graph Markov property. Scandinavian Journal of Statistics, 17:333–353, 1990.

[GK17] Glynn, Adam; Kashin, Konstantin: Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program. Journal of the American Statistical Association, 2017.

[Im14] Imbens, Guido: Instrumental Variables: An Econometrician's Perspective. Statistical Science, 29(3):323–358, 2014.

[Ka12] Kalisch, Markus; Mächler, Martin; Colombo, Diego; Maathuis, Marloes; Bühlmann, Peter: Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software, 47(11):, 2012.

[Ky98] Kyono, Trent Mamoru: Commentator: A Front-End User-Interface Module for Graphical and Structural Equation Modeling. Bericht R-364, University of California, Los Angeles, 1998.

[LW89] Lauritzen, Steffen; Wermuth, Nanny: Graphical models for associations between variables, some of which are qualitative and some quantitative. The Annals of Statistics, 17:31–57, 1989.

[Pe93] Pearl, Judea: Comment: Graphical models, causality and intervention. Statistical Science, 8:266–269, 1993.

[Pe01] Pearl, Judea: Parameter identification: A new perspective. Bericht R-276, UCLA, 2001.

[Pe09] Pearl, Judea: Causality. Cambridge University Press, 2009.

[RGL08] Rothman, Kenneth J.; Greenland, Sander; Lash, Timothy L.: Modern Epidemiology. Wolters Kluwer, 2008.

[RS02] Richardson, Thomas; Spirtes, Peter: Ancestral Graph Markov Models. Annals of Statistics, 30:927–1223, 2002.

[Sh98] Shachter, Ross D.: Bayes-Ball: The Rational Pastime. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, S. 480–487, 1998.

[SVR10] Shpitser, Ilya; VanderWeele, Tyler; Robins, James: On the Validity of Covariate Adjustment for Estimating Causal Effects. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press, S. 527–536, 2010.

[TL11] Textor, Johannes; Liśkiewicz, Maciej: Adjustment Criteria in Causal Diagrams: An Algorithmic Perspective. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. AUAI Press, S. 681–688, 2011.

[TPP98] Tian, Jin; Paz, Azaria; Pearl, Judea: Finding Minimal D-separators. Bericht R-254, University of California, Los Angeles, 1998.

[Va09] VanderWeele, Tyler J.: On the relative nature of overadjustment and unnecessary adjustment. Epidemiology, 20(4):496–499, 7 2009.

[vdZ] van der Zander, Benito: Algorithmics of Identifying Causal Effects in Graphical Models.

[vdZLT19] van der Zander, Benito; Liśkiewicz, Maciej; Textor, Johannes: Separators and Adjustment Sets in Causal Graphs: Complete Criteria and an Algorithmic Framework. Artificial Intelligence, 270:1–40, 2019.

Benito van der Zander hat seinen Bachelor-Abschluss über Ereignisprotokolle in Düsseldorf gemacht, seinen Master-Abschluss über Computer Vision in Aachen, und zuletzt in Lübeck über Kausalität promoviert. Während der Zeit der Promotion forschte er für ein DFG-Projekt und betreute Übungsgruppen am Institut für Theoretische Informatik. Bereits als Schüler hatte er Vorlesungen an der Universität besucht, eigene Software verkauft, wurde Bundessieger beim Bundeswettbewerb Informatik und errang ein paar Medaillen bei europäischen Informatikolympiaden. Er programmiert aktiv an Open-Source-Projekten: DAGitty mit den Algorithmen der Dissertation, Bibliothek-App VideLibri zur Literaturverwaltung von ausgeliehenen Büchern, LaTeX-Editor TeXstudio, sowie das Webseiten-Automatisierungstool und XQuery-Interpreter Xidel.

Energiebeschränkte Echtzeitsysteme und ihre Worst-Case-Analysen[1]

Peter Wägemann[2]

Abstract: Der zuverlässige Betrieb von Anwendungen, welche sowohl Zeit- als auch Energiebeschränkungen aufweisen, ist eine zentrale Herausforderung im Bereich der eingebetteten Systeme. Um die Ausführung von Aufgaben innerhalb von Ressourcenbudgets zu garantieren, benötigen energiebeschränkten Echtzeitsysteme Schranken der Worst-Case-Ausführungszeit sowie des Worst-Case-Energieverbrauchs. Neben dem Problem der Bestimmung von oberen Schranken haben Worst-Case-Analysewerkzeuge das grundlegende Problem, dass die Genauigkeit der ausgegebenen Schranken unbekannt ist, in Hinblick auf den tatsächlichen schlimmsten anzunehmenden Fall.

Die Dissertation adressiert die genannten Probleme, indem zunächst ein Analyseansatz für Schranken des Energieverbrauchs aufgezeigt wird, welcher vollständige Echtzeitsysteme mit allen Leistungsverbrauchern verarbeitet. Hinsichtlich des Analysepessimismus der Schranken stellt die Dissertation einen Ansatz zur Bestimmung der Genauigkeit von Analysen basierend auf automatisch generierten Benchmark-Programmen vor. Um die notwendigen Bestandteile für den sicheren Betrieb zu komplettieren, präsentiert die Dissertation einen Betriebssystemkern, welcher auf Szenarien dynamisch reagiert, bei denen eine Ressource stärkere Beschränkungen aufweist als die andere.

1 Einleitung

Eine zentrale Herausforderung dieses Jahrzehnts im Bereich der Computersysteme ist der zuverlässige Betrieb von eingebetteten Systemen mit starken Energiebeschränkungen, zum Beispiel durch die unkontinuierliche Energiezufuhr durch Energie-Harvesting-Mechanismen [Co18]. Diese Klasse von Rechensystemen ist von praktischer Relevanz in Form von implantierbaren medizinischen Geräten, wie beispielsweise Herzschrittmacher oder Defibrillatoren [Ou19]. Diese Anwendungen haben einerseits harte Echtzeitanforderungen, da sowohl die Sensordatenerfassung der Herzaktivität als auch die Aktorik von möglichen Schockimpulsen innerhalb von bestimmten Zeitschranken ausgeführt werden müssen. Zusätzlich existieren – aufgrund des Batteriebetriebs – Anforderungen an den Energiebedarf aller Aufgaben des Systems. Um eine verlässliche Planung unter Berücksichtigung der verfügbaren Zeit-/Energieressourcen zu garantieren, sind die Werte der *schlimmsten anzunehmenden Ausführungszeit (engl. worst-case execution time, WCET)* sowie des *schlimmsten anzunehmenden Energieverbrauchs (engl. worst-case energy consumption, WCEC)* für jede Aufgabe im System notwendig.

Statische Programmcodeanalysen sind ein grundlegendes Mittel im Bereich der Echtzeitsysteme für die Bestimmung von verlässlichen Schranken des Ressourcenbedarfs. Zwar

[1] Englischer Titel der Dissertation: „Energy-Constrained Real-Time Systems and Their Worst-Case Analyses"

[2] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), waegemann@cs.fau.de

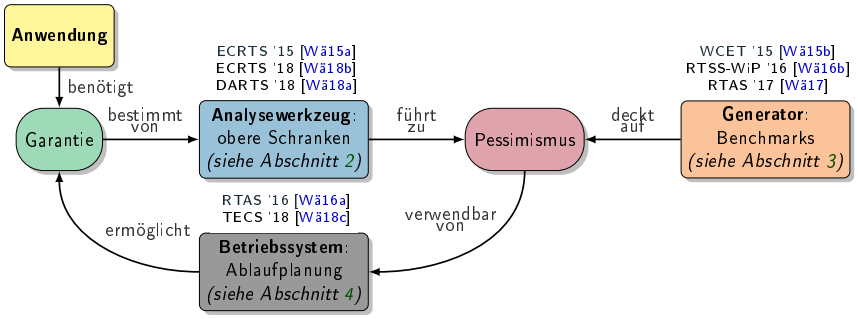


Abb. 1: Konzeptionelle Struktur der Dissertation

existieren praxistaugliche Analysen für die WCET [Wi08], jedoch sind diese nicht direkt verwendbar für das fundamental andere Verhalten des Energieverbrauchs und somit für die Bestimmung von WCEC-Werten. Im Gegensatz zur Laufzeitanalyse ist beispielsweise eine Betrachtung von Echtzeitprioritäten für den Energieverbrauch unzureichend: So können auch niederpriore Aufgaben verschiedene Geräte (z. B. Transceiver) temporär aktivieren, welche dadurch den Leistungsbedarf des Gesamtsystems ändern. Dieser erhöhte Leistungsbedarf beeinflusst wiederum den Energieverbrauch aller Aufgaben, unabhängig von deren Echtzeitpriorität. Diese *wechselseitige Beeinflussung* des Ressourcenbedarfs zwischen allen Aufgaben im System muss bei der WCEC-Analyse berücksichtigt werden.

Ein fundamentales Problem bei der Verwendung von statischen Worst-Case-Analysen ist die Evaluation und die Validierung der oberen Schranke des Ressourcenbedarfs, die das Analysewerkzeug ermittelt. Durch sichere Abstraktionen bestimmen Analysewerkzeuge wiederum sichere Schranken. Allerdings ist sowohl der tatsächliche WCET- als auch der tatsächliche WCEC-Wert von beliebigen Programmen unbekannt, da allgemein solchen nicht-trivialen Eigenschaften aus Programmen nicht automatisiert abgeleitet werden können [KKZ13]. Dadurch ist die Genauigkeit der Abstraktionen der Analysewerkzeuge nicht bewertbar, wodurch der Analysepessimismus unbekannt bleibt. Das fehlende Wissen über diese Vergleichslinien (d. h. tatsächliche WCET-/WCEC-Werte) verhindert auch die Beantwortung der Frage, ob die Implementierung der Analyse Fehler aufweist und möglicherweise Unterschätzungen des Ressourcenbedarfs ausgibt.

Verlässliche Schranken des Ressourcenbedarfs sind notwendige Voraussetzungen für den Entwurf von energiebeschränkten Echtzeitsystemen. Zur Laufzeit benötigen diese Systeme einen Betriebssystemkern zusammen mit einer ressourcengewahren Ablaufplanung, um einen *sicheren Betrieb* zu gewährleisten. Für einen *effizienten Betrieb* ist Wissen über den erwarteten Analysepessimismus notwendig, da die Ablaufplanung diesen nutzen kann.

Die Dissertation [Wä20] adressiert die genannten Probleme der Ermittlung von WCEC-Schranken, der Bestimmung von Analysepessimismus und des Betriebs der energiebeschränkten Echtzeitsysteme. Die konzeptionelle Struktur ist in Abbildung 1 illustriert. Folgende drei Lösungsansätze stellen die Hauptbeiträge der Dissertation dar: Der WCEC-Analysator *SysWCEC* bestimmt sichere Schranke des Energiebedarfs für eingebettete Echtzeitsysteme mit festen Prioritäten (siehe Abschnitt 2). Der Benchmark-Generator *GenE*

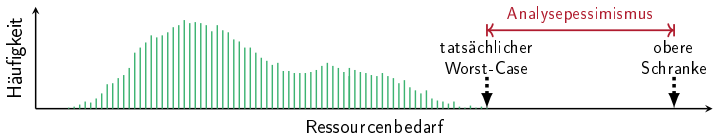


Abb. 2: Exemplarisches Histogramm des Ressourcenbedarfs einer Aufgabe

ermöglicht erstmals eine Evaluation und Validierung von statischen Worst-Case-Analysatoren, basierend auf dem Wissen des tatsächlichen Worst-Case-Ressourcenbedarfs (siehe Abschnitt 3). Der Betriebssystemkern *EnOS* unterstützt den sicheren Betrieb von energiebeschränkten Echtzeitsystemen durch a priori Wissen über WCET- und WCEC-Werte (siehe Abschnitt 4). Die Struktur dieses Artikels ergibt sich aus diesen drei Hauptbeiträgen, wie in Abbildung 1 dargestellt. Abschnitt 5 fasst die Ergebnisse zusammen.

2 Worst-Case-Analysen

Die Probleme bei der Bestimmung von WCEC-Werten ist Inhalt von Abschnitt 2.1. Abschnitt 2.2 beschreibt die Lösungen durch *SysWCEC*. Zunächst werden jedoch Hintergrundinformationen zu Worst-Case-Analysen und das Systemmodell präsentiert.

Analysepessimismus Abbildung 2 zeigt ein exemplarisches Histogramm von Ressourcenverbräuchen (entweder Zeit oder Energie) einer Aufgabe. Dabei variiert der Ressourcenbedarf zum Beispiel durch verschiedene Eingabedaten oder durch unterschiedliche initiale Hardwarezustände zu Beginn der Ausführung der Aufgabe. Der tatsächlich Worst-Case-Wert ist im Allgemeinen nicht genau bestimmbar, weshalb Analysatoren pessimistische Annahmen treffen, die zu Überschätzungen dieses Wertes führen. Der Abstand zwischen Worst-Case-Wert und oberer Schranke kennzeichnet den Pessimismus des Analysators für die sichere Ausführung der Aufgabe. Die Bestimmung vom erwarteten Grad des Pessimismus (siehe Abschnitt 3) ist eine notwendige Voraussetzung für den effizienten Betrieb von energiebeschränkten Echtzeitsystemen (siehe Abschnitt 4).

Systemmodell Die adressierte Klasse von energiebeschränkten Echtzeitsystemen hat einen Rechenkern. Die Aufgaben der Anwendung werden mit festen Prioritäten abgearbeitet. Weiterhin können die Aufgaben Betriebsmittel zu Synchronisationszwecken anfordern. Das Auftreten von asynchronen Interrupts ist in diesen Systemen oftmals gegeben, wenn beispielsweise Zeitgeber-Interrupts für die Laufzeitüberwachung eingesetzt werden. Für die statische Worst-Case-Analyse stellen diese Interrupts eine große Herausforderung dar, da das mögliche dynamische Verhalten statisch abgebildet werden muss. Das Auftreten dieser Unterbrechungen ist über ihre minimale Zwischenankunftszeit begrenzt. Für den Leistungsbedarf des Systems (und somit auch den Energiebedarf über die Zeit gesehen) spielen die softwaregesteuerten Geräte eine entscheidende Rolle. Dabei sind Geräte nicht notwendigerweise als extern anzusehen, wie beispielsweise Transceiver zur Kommunikation. Interne Geräte (wie Zeitgeber-Subsysteme oder Analog-Digital-Konverter) haben das gleiche Verhalten aus Sicht der statischen Analyse: Das Aktivieren des Geräts führt zu einer erhöhten Leistung des Gesamtsystems, und das Deaktivieren reduziert wiederum den Verbrauch, wie im Folgenden näher erläutert.

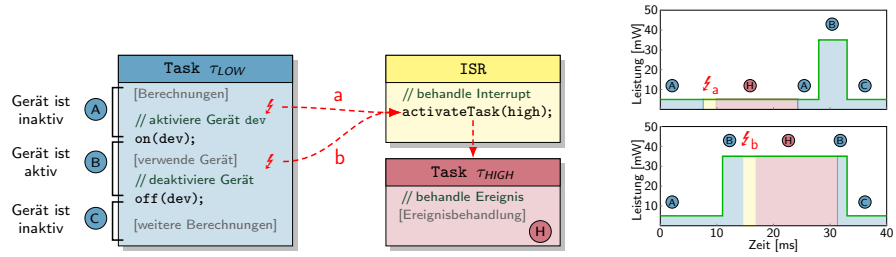


Abb. 3: Der Leistungsbedarf des Gesamtsystems beeinflusst die Energie jeder Aufgabe.

2.1 Problemstellungen der WCEC-Analyse

Zur Energieeinsparung werden Geräte nur dann aktiviert, wenn die Anwendung den Dienst benötigt. Abbildung 3 zeigt eine niederpriore Aufgabe (engl. task) τ_{LOW}, eine Interrupt-Service-Routine (ISR) und eine höherpriore Aufgabe τ_{HIGH}, die von der ISR aktiviert wird. Die niederpriore Aufgabe führt eine temporäre Aktivierung eines Geräts aus, welche den Leistungsbedarf des Gesamtsystems von 5 mW auf 35 mW anhebt. Wie auf der rechten Seite in Abbildung 3 illustriert ist, ergeben sich zwei Szenarien beim Energiebedarf zur Fertigstellung der niederprioren Aufgabe τ_{LOW}: In Szenario a (siehe $\lightning_a$ in Abbildung 3) tritt der Interrupt im Zustand der geringen Leistungsaufnahme auf. Im Gegensatz dazu wird der Interrupt in Szenario b ($\lightning_b$) im hohen Leistungszustand abgearbeitet, wodurch sich hier der Worst-Case hinsichtlich des Energiebedarfs ergibt (schlimmste anzunehmende Fläche unter der Kurve). Dieser Worst-Case-Energiebedarf gilt sowohl für τ_{LOW} (gesamte Fläche) als auch für τ_{HIGH} (rote Fläche unter (H)), weil τ_{HIGH} durch die Aktivierung von τ_{LOW} mit der hohen Leistung von 35 mW startet. Für die Analyse der Laufzeit sind die kontextsensitiven Leistungszustände irrelevant, da lediglich die Beeinflussung der niederprioren Aufgaben durch höherpriore Aufgaben betrachtet werden muss. Jedoch muss die WCEC-Analyse die systemweite Aktivität der Geräte sowie deren Leistungszustände berücksichtigen. Bei der Modellierung des Energieverhaltens ist eine Betrachtung der *wechselseitigen Beeinflussung* von Aufgaben erforderlich.

2.2 *SysWCEC* – systemweite WCEC-Analyse

Zur Lösung des Problems der WCEC-Analyse stellt die Dissertation den *SysWCEC*-Ansatz vor [Wä18b]. Die Grundidee von *SysWCEC* besteht aus vier Teilen: *(1.) Dekomposition*: Der Code des Zielsystems wird untergliedert in Abschnitte, welche eine gemeinsame Menge an aktiven Geräten besitzen. *(2.) Pfadanalyse*: Basierend auf dem Ergebnis der Dekomposition des Systems wird eine explizite Aufzählung aller möglichen Programmpfade ausgeführt. *(3.) Problemformulierung*: *SysWCEC* verwendet das gewonnene Wissen aus der Pfadanalyse für die Formulierung eines ganzzahlig linearen Programms (engl. integer linear program, ILP). *(4.) WCEC-Bestimmung*: Das Lösen des ILPs ergibt schlussendlich eine WCEC-Schranke der zu analysierenden Aufgabe.

Der folgende Abschnitt erläutert das Vorgehen von *SysWCEC* näher, ausgehend vom Beispiel in Abbildung 3. Die linke Seite von Abbildung 4 zeigt nun das untergliederte Sys-

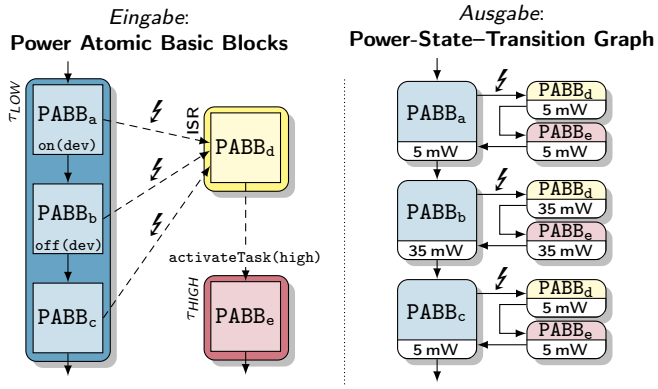


Abb. 4: Der *SysWCEC*-Ansatz verwendet die Untergliederung des Systems in Blöcke mit gleicher Leistungsaufnahme für eine Aufzählung aller möglichen Systemzustände.

tem: Die ursprüngliche niederpriore Aufgabe τ_{LOW} besteht jetzt aus einzelnen *Power Atomic Basic Blocks (PABBs)*, welche sich unter anderem durch eine gemeinsame Menge an aktiven Geräten kennzeichnen und eine Erweiterung der Atomic Basic Blocks darstellen [SSP11]. Die Aufgabe τ_{LOW} besteht somit aus den drei Teilen $\texttt{PABB}_\texttt{a}$ (Code vor Aktivierung), $\texttt{PABB}_\texttt{b}$ (Code während Aktivierung) und $\texttt{PABB}_\texttt{c}$ (Code nach Aktivierung).

Die rechte Seite von Abbildung 4 zeigt das Ergebnis der Pfadanalyse, den *Leistungszustandsgraphen (engl. Power-State–Transition Graph, PSTG)*. Der Algorithmus der Pfadanalyse beginnt mit dem initialen Zustand der geringen Leistungsaufnahme (5 mW). Unter Beachtung der Betriebssystemsemantik und der potenziell auftretenden Interrupts werden Transitionen im PSTG eingefügt. Dabei wird bei einem gerätebezogenen Aufruf (z. B. Aktivierung des Geräts) der neue Leistungszustand propagiert. Ansonsten gibt die Analyse den vorherrschenden Kontext der Geräte und deren Leistung über die möglichen Systempfade weiter, solange bis alle Pfade exploriert wurden. Das Wissen über alle Systempfade löst das Analyseproblem der wechselseitigen Beeinflussung der Aufgaben. *SysWCEC* verwendet das Wissen über Pfade des PSTG für die Nebenbedingungen des ILPs. Die Zielfunktion des ILPs ist die Maximierung des Energiebedarfs über alle möglichen Systemzustände, in denen die zu analysierende Aufgabe ausgeführt werden kann. Das Lösen des ILPs mittels eines mathematischen Optimierers liefert die obere WCEC-Schranke.

Die Formulierung der Kosten für die Zielfunktion des ILPs verwendet den maximalen (kontextsensitiven) Leistungszustand P_{max} der Knoten im PSTG. Um nun wiederum eine obere Schranke für einzelne Knoten des PSTG zu erhalten, bestimmt *SysWCEC* die WCET jedes Knotens. Der Energiebedarf einzelner Knoten berechnet sich folglich durch $WCEC = P_{max} \cdot WCET$. Durch diese Art der WCEC-Analyse ist es möglich, *sichere* Schranken von Knoten zu bestimmen. Jedoch ist die *Genauigkeit* der WCET-Schranke unbekannt. Für diese Frage nach dem Analysepessimismus von *SysWCEC* muss der Pessimismus von WCET-Werten abgeschätzt werden. Das Vorgehen bei der Abschätzung von Pessimismus ist die zentrale Fragestellung des folgenden Abschnitts.

3 Validierung von Worst-Case-Analysen

Der folgende Abschnitt 3.1 erläutert das grundsätzliche Problem bei der Bestimmung der Genauigkeit von Worst-Case-Analysen. Anschließend beschreibt Abschnitt 3.2 die Lösung dieses Problems durch den Benchmark-Generator *GenE*.

3.1 Problem der Validierung von Worst-Case-Analysen

Durch die verwendeten Abstraktionen in den Analysealgorithmen liegt die ausgegebene obere Schranke – bei korrekter Implementierung – über dem tatsächlichen Worst-Case (wie in Abbildung 2). Existierende Verfahren zur Abschätzung des Analysepessimismus verwenden Benchmark-Suites auf der Ebene von Quellcode (z. B. C). Das grundlegende Problem ist hierbei, dass bei diesen Evaluationsverfahren das absolute Vergleichsmaß, also der tatsächliche Worst-Case, fehlt. Dieser tatsächliche Worst-Case sowie alle relevanten Programmfakten (wie Schleifenobergrenzen) sind jedoch nicht automatisiert bestimmbar [KKZ13]. Auch eine manuelle Extraktion dieser Fakten ist aufwendig und fehleranfällig. Durch dieses fehlende Wissen haben existierende Evaluationsansätze nur eine geringe Aussagekraft, weil sie den absoluten Grad der Über- oder (im Fall eines Softwarefehlers) Unterschätzung nicht bestimmen können.

3.2 *GenE* – Benchmark-Generator für Worst-Case-Analysen

Der Generator *GenE* ist eine Lösung für das Problem der Evaluation und Validierung von Worst-Case-Analysen [Wä15b, Wä17]. *GenE* generiert Benchmark-Programme auf eine Art, sodass alle Programmfakten bekannt sind. Mit diesem Wissen ist der tatsächliche Worst-Case bestimmbar, der wiederum als Vergleichsmaß für Analysewerkzeuge dient. Die Funktionsweise von *GenE* lässt sich anhand einer Metapher veranschaulichen: Benchmarks sind wie ein Irrgarten für Analysatoren, welche einen Weg durch den Irrgarten finden müssen. Selbst wenn ein Analysator möglicherweise einen Weg (Lösung) finden würde, so ist unbekannt, ob dieser der optimale Weg ist. *GenE* verfolgt den genau umgekehrten Ansatz: Zunächst wird ein Pfad bestimmt und darauf folgend werden Verzweigungen entlang dieses Pfades eingefügt. Dadurch wird der Irrgarten um den vorab bestimmten Pfad gebaut. Durch den generativen Ansatz von *GenE* ist die optimale Lösung (tatsächlicher Worst-Case) des Irrgartens konstruktionsbedingt bekannt.

Programm-Patterns Für den Generierungsprozess verwendet *GenE* eine Vielzahl von verschiedenen *Programm-Patterns*. Diese haben die Eigenschaft, dass sie Wissen über die Programmfakten (wie ihren Worst-Case-Pfad) besitzen. Weiterhin sind Programm-Patterns miteinander kombinierbar für die Erzeugung von neuen Benchmarks. Dadurch generiert *GenE* beispielsweise Schleifen innerhalb von eingabedatenabhängigen Verzweigungen. *GenE* implementiert eine Vielzahl von unterschiedlichen Pattern für Pfade, Schleifen, arithmetische Operationen oder Variablen. Um die Generierung von unrealistischem Code für Worst-Case-Analysen zu vermeiden, verwendet *GenE* Pattern aus realen Industrieanwendungen sowie aus existierenden Benchmark-Suites.

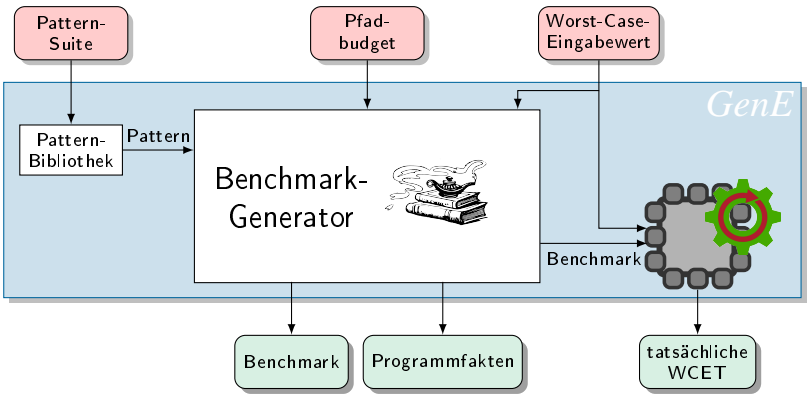


Abb. 5: Der *GenE*-Generator erzeugt Benchmarks so, dass der schlimmste Fall des Ressourcenbedarfs bekannt ist. Dieser Fall dient als Vergleichsmaß für statische Worst-Case-Analysewerkzeuge.

Pattern-Suites Ein weiteres Problem bei der bisherigen Verwendung von Benchmarks ist die Tatsache, dass existierende Benchmarks im Bereich der Worst-Case-Analyse üblicherweise mehrere verschiedene Pattern enthalten. Daraus resultiert das Problem, dass verschiedene Faktoren zum Pessimismus in der ausgegebenen Worst-Case-Schranke beitragen. *GenE* löst dieses Problem, indem die Auswahl der verwendeten Programm-Pattern konfigurierbar ist. Durch diese spezifischen *Pattern-Suites* kann *GenE* die individuellen Stärken und Schwächen von Analysatoren evaluieren.

Ein- & Ausgaben Diese Konfiguration der Pattern-Suite ist eine der Eingaben für *GenE*, wie in Abbildung 5 dargestellt. Das Pfadbudget konfiguriert die Komplexität des generierten Benchmarks (durch die Angabe der Anzahl von Instruktionen entlang des Worst-Case-Pfades). Der Worst-Case-Eingabewert ist ein entscheidender Wert für den generierten Benchmark: Wird das Programm mit diesem Wert ausgeführt, so wird der Worst-Case-Pfad abgelaufen. Die Vermessung dieses Pfades ergibt letztendlich den tatsächlichen Worst-Case-Ressourcenbedarf. *GenE* zielt auf die Evaluation von WCET-Analysatoren ab. Basierend auf *GenE* entwickelten Eichler et al. einen Generator für vollständige Echtzeitsysteme [Ei18] sowie einen speziellen Generator für WCEC-Analysatoren [EWSP19].

Evaluation & Validierung von aiT WCET-Werkzeug Evaluationen mit dem WCET-Analysator aiT der Firma AbsInt zeigten einige Benchmarks, bei denen der ausgegebene Wert geringer als die tatsächliche WCET war. AbsInt konnte mithilfe der Benchmarks von *GenE* dieses Verhalten bestätigten, welches durch ein fehlerhaftes Hardwaremodell verursacht wurde. Nachdem die Unterschätzungen bekannt wurden, korrigierte AbsInt diese Fehler in einer überarbeiteten Version von aiT, in der keine Fehler mehr gefunden werden konnten. In Anbetracht der Tatsache, dass statischen WCET-Werkzeuge für äußerst sicherheitskritische Echtzeitsysteme (wie den Airbus A380) verwendet werden, ist die strukturierte Evaluation und Validierung durch einen Testfall-Generator wie *GenE* von hoher Relevanz, um die Qualität der Analysatoren zu verbessern.

GenE ermöglicht durch die spezifischen Pattern-Suites das Aufspüren von individuellen Schwächen in der WCET-Analyse. Jedoch ist bei der Verwendung von sicheren Schranken

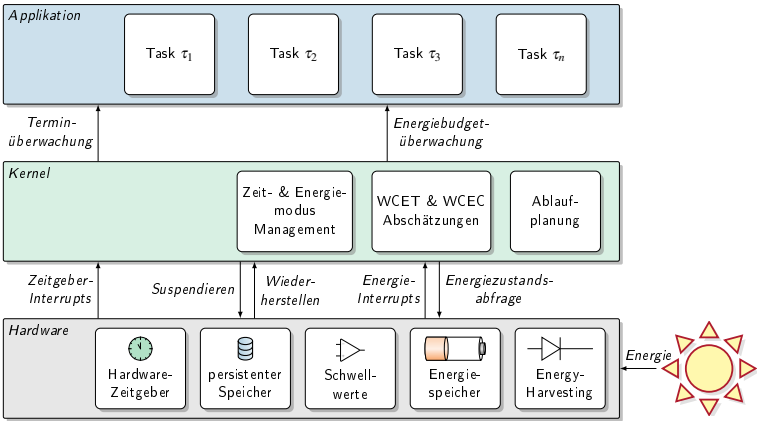


Abb. 6: *EnOS* ermöglicht Echtzeitgarantien während dedizierter Phasen und sorgt durch die vorausschauende Ablaufplanung für sichere Zustände während Stromausfällen.

Pessimismus zur Laufzeit unvermeidbar. Wenn allerdings der Betriebssystemkern Wissen über den erwarteten Grad des Pessimismus besitzt, so kann der Kern diesen effizient ausnutzen. Der nächste Abschnitt über den *EnOS*-Kern erläutert diesen Zusammenhang näher.

4 Betrieb energiebeschränkter Echtzeitsysteme

Beim Betrieb von energiebeschränkten Echtzeitsystemen ergeben sich mehrere Herausforderungen (siehe Abschnitt 4.1). Basierend darauf stellt Abschnitt 4.2 *EnOS* vor.

4.1 Herausforderungen beim Betrieb energiebeschränkter Echtzeitsysteme

Analysepessimismus zeigt sich zur Laufzeit durch ungenutzte Ressourcen. Besonders in eingebetteten Systemen möchte man dies verhindern. Weiterhin muss die Ablaufplanung des energiebeschränkten Echtzeitsystems die beiden Ressourcen Zeit und Energie berücksichtigen, um Ausführungs- und Laufzeitgarantien für kritische Aufgaben zu gewährleisten. Der in Abschnitt 4.2 vorgestellte Betriebssystemkern *EnOS* adressiert die spezielle Klasse der *energieneutralen* Systeme, welche sich dadurch auszeichnet, dass die verwendete Energie vollständig aus der Umgebung entnommen wird und in aufladbaren Batterien gespeichert wird. Mechanismen für diese Art der Energieernte (engl. energy harvesting) umfassen beispielsweise Solarzellen. Durch die unregelmäßige Energiezufuhr muss das System mögliche Stromausfälle tolerieren. Wenn sich der Batteriezustand dem Ende entgegen neigt, muss das System rechtzeitig in einen sicheren Zustand übergehen, da dieser Übergang selbst Energie erfordert. Weiterhin benötigt das System Wissen darüber, wann ausreichend Energie geerntet wurde, um nach einem Stromausfall wieder sinnvolle Arbeit unterbrechungsfrei auszuführen und Phasen mit Echtzeitanforderungen zu ermöglichen.

4.2 *EnOS* – Betriebssystemkern für energieneutrale Echtzeitsysteme

Abbildung 6 zeigt eine Übersicht des Designs. Hardwareseitig verwendet *EnOS* eine Überwachung des Ressourcenbedarfs, welche beispielsweise beim Über-/Unterschreiten eines Schwellwerts der verfügbaren Energie dem Kern einen Interrupt zustellt. *EnOS* benutzt nicht-flüchtigen Speicher für das Sichern von Daten während eines Stromausfalls.

Auf der Softwareseite ist die offline berechnete Budgetierung von verschiedenen Betriebsmodi basierend auf WCEC- und WCET-Abschätzungen von zentraler Bedeutung. Diese Modi werden zur Laufzeit ausgeführt, entsprechend der verfügbaren Energie. Durch die vorausschauende Planung erlaubt *EnOS* Phasen mit Echtzeitgarantien. Nahe der leeren Batterie garantiert *EnOS* den Übergang in einen sicheren Zustand durch pessimistische (aber verlässliche) WCEC-Schranken. Dagegen verwendet *EnOS* in Phasen von hoher Energieverfügbarkeit optimistische Abschätzungen, welche den Einfluss des Analysepessimismus mildern und Ressourcen effizienter ausnutzen. Diese sind möglicherweise Unterschätzungen des tatsächlichen WCEC-Werts, jedoch garantiert die Ressourcenüberwachung unter allen Umständen die Ausführung von kritischen Aufgaben. Bei der Bestimmung von optimistischen Budgets und des erwarteten Grades an Analysepessimismus ist wiederum *GenE* hilfreich (siehe Abschnitt 3). Das Wissen über WCEC-Werte für Aufgaben ermöglicht *EnOS* die Beantwortung der Frage, wann ausreichend Energie geerntet wurde, um den Betrieb nach einem Stromausfall vorausschauend wieder aufzunehmen.

5 Zusammenfassung

Die Dissertation [Wä20] zielt auf den verlässlichen Betrieb von Systemen ab, die sich durch Energie- und Zeitbeschränkungen kennzeichnen. Der erste der drei Hauptbeiträge ist der Analysator *SysWCEC*, welcher Schranken des Energiebedarfs von Aufgaben bestimmt. *SysWCEC* beinhaltet eine Methodik zur Modellierung von temporär aktiven Leistungsverbrauchern. Weiterhin betrachtet diese Analyse die Ablaufplanung mit festen Prioritäten, die Verwendung von Betriebsmitteln, synchrone Aufgaben und asynchrone Interrupts. Die Bestimmung des Analysepessimismus ist ein fundamentales Problem bei Worst-Case-Analysen, welches die Dissertation mit dem Benchmark-Generator *GenE* löst. Das Aufdecken von Softwarefehlern in einem verbreiteten WCET-Werkzeug unterstreicht die Signifikanz eines solchen Generators zur Validierung von Analysatoren. Um zur Laufzeit einen Betrieb unter Zeit- und Energieschranken zu ermöglichen, stellt die Dissertation den Kern *EnOS* vor. Durch das Wissen über den erwarteten Analysepessimismus lässt sich dessen Einfluss mildern. *EnOS* verwendet zur Laufzeit ungenutzte Ressourcen für unkritische Aufgaben und garantiert schlussendlich die Ausführung kritischer Aufgaben unter Verwendung von verlässlichen Schranken des Ressourcenbedarfs.

> Der Quellcode von *SysWCEC*, *GenE* und *EnOS* ist online verfügbar:
> `https://gitlab.cs.fau.de/{syswcec, gene, enos}`

Literaturverzeichnis

[Co18] Cohen, A. et al.: Inter-Disciplinary Research Challenges in Computer Systems for the 2020s. Bericht, USA, 2018.

[Ei18] Eichler, C.; Distler, T.; Ulbrich, P.; Wägemann, P.; Schröder-Preikschat, W.: TASKers: A Whole-System Generator for Benchmarking Real-Time-System Analyses. In: Proc. of WCET '18. S. 6:1–6:12, 2018.

[EWSP19] Eichler, C.; Wägemann, P.; Schröder-Preikschat, W.: GenEE: A Benchmark Generator for Static Analysis Tools of Energy-Constrained Cyber-Physical Systems. In: Proc. of CPS-IoTBench '19. 2019.

[KKZ13] Knoop, J.; Kovács, L.; Zwirchmayr, J.: WCET Squeezing: On-demand Feasibility Refinement for Proven Precise WCET-bounds. In: Proc. of RTNS '13. S. 161–170, 2013.

[Ou19] Ouyang, H. et al.: Symbiotic cardiac pacemaker. Nature Communications, 10, 2019.

[SSP11] Scheler, F.; Schröder-Preikschat, W.: The Real-Time Systems Compiler: Migrating event-triggered systems to time-triggered systems. Software: Practice and Experience, 41(12):1491–1515, 2011.

[Wä15a] Wägemann, P.; Distler, T.; Hönig, T.; Janker, H.; Kapitza, R.; Schröder-Preikschat, W.: Worst-Case Energy Consumption Analysis for Energy-Constrained Embedded Systems. In: Proc. of ECRTS '15. S. 105–114, 2015.

[Wä15b] Wägemann, P.; Distler, T.; Hönig, T.; Sieh, V.; Schröder-Preikschat, W.: GenE: A Benchmark Generator for WCET Analysis. In: Proc. of WCET '15. S. 33–43, 2015.

[Wä16a] Wägemann, P.; Distler, T.; Janker, H.; Raffeck, P.; Sieh, V.: A Kernel for Energy-Neutral Real-Time Systems with Mixed Criticalities. In: Proc. of RTAS '16. S. 25–36, 2016.

[Wä16b] Wägemann, P.; Distler, T.; Raffeck, P.; Schröder-Preikschat, W.: Towards Code Metrics for Benchmarking Timing Analysis. In: Proc. of RTSS WiP '16. 2016.

[Wä17] Wägemann, P.; Distler, T.; Eichler, C.; Schröder-Preikschat, W.: Benchmark Generation for Timing Analysis. In: Proc. of RTAS '17. S. 319–330, 2017.

[Wä18a] Wägemann, P.; Dietrich, C.; Distler, T.; Ulbrich, P.; Schröder-Preikschat, W.: Whole-System WCEC Analysis for Energy-Constrained Real-Time Systems (Artifact). Dagstuhl Artifacts Series (DARTS '18), 4(2):7:1–7:4, 2018.

[Wä18b] Wägemann, P.; Dietrich, C.; Distler, T.; Ulbrich, P.; Schröder-Preikschat, W.: Whole-System Worst-Case Energy-Consumption Analysis for Energy-Constrained Real-Time Systems. In: Proc. of ECRTS '18. S. 24:1–24:25, 2018.

[Wä18c] Wägemann, P.; Distler, T.; Janker, H.; Raffeck, P.; Sieh, V.; Schröder-Preikschat, W.: Operating Energy-Neutral Real-Time Systems. ACM TECS, 17(1):11:1–11:25, 2018.

[Wä20] Wägemann, P.: Energy-Constrained Real-Time Systems and Their Worst-Case Analyses. Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2020.

[Wi08] Wilhelm, R. et al.: The Worst-case Execution-time Problem – Overview of Methods and Survey of Tools. ACM TECS, 7(3):1–53, 2008.

Peter Wägemann ist Post-Doktorand an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Im September 2020 promovierte Peter Wägemann mit Auszeichnung an der Technischen Fakultät der FAU im Bereich der Informatik. Neben einer Lehrtätigkeit im Bereich von Echtzeitsystemen sowie dem individuellen Prototyping, umfasst seine Arbeit die Forschung auf den Gebieten der eingebetteten Systeme und deren statischer Programmanalyse.

Zufällige Hypergraphen für Hashing-basierte Datenstrukturen[1]

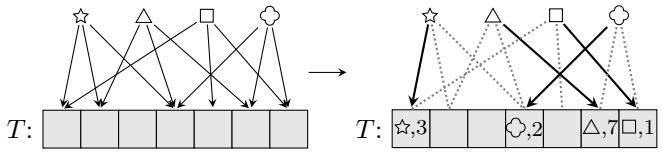
Stefan Walzer[2]

Abstract: Für Wörterbücher und verwandte Datentypen gibt es Implementierungsansätze, bei denen Hashfunktionen mehrere zufällige Möglichkeiten zur Speicherung jedes Schlüssels vorsehen. Solche Verfahren weisen oft einen scharfen Schwellwert bzgl. erzielbarer Auslastungsfaktoren auf, aus dem sich die Speichereffizienz der Datenstruktur ergibt. Hypergraphen spielen in Schwellwertanalysen – nicht aber in dieser Zusammenfassung – eine zentrale Rolle.

Drei Ergebnisse der Arbeit liefern Schwellwerte für Varianten von Cuckoo-Hashtabellen, darunter die Doppel-Hashing-Variante und eine Variante mit nicht-ausgerichteten Blöcken. Drei weitere Ergebnisse untersuchen neue Varianten von Retrieval-Datenstrukturen, aus denen sich unter anderem neue speichereffiziente Alternativen zu Bloom-Filtern ergeben.

1 Orientierbare, lösbare und schälbare Konfigurationen

Man stelle sich vor, die Zuordnung $f = \{☆ \mapsto 3, △ \mapsto 7, □ \mapsto 1, ♢ \mapsto 2\}$ soll mithilfe einer Tabelle T gespeichert werden. Um die Informationen schnell abrufen zu können und zugleich etwas Flexibilität zu wahren, sind jedem *Schlüssel* aus $\{☆, △, □, ♢\}$ mehrere Positionen in T zufällig zugeordnet. Die Information zu jedem Schlüssel muss an genau einer dieser Positionen abgelegt werden und jede Position darf höchstens ein Schlüssel/Wert-Paar enthalten. Im folgenden Beispiel ist es möglich, alle Schlüssel zu platzieren, daher nennen wir die Konfiguration links **orientierbar**. Eine passende *Orientierung* ist rechts gezeigt.

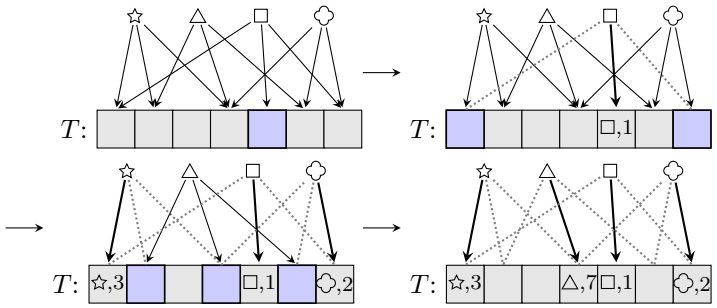


(1)

Im vorliegenden Fall kann eine Orientierung sogar mit einem Greedy-Algorithmus gefunden werden, der nach Tabellenpositionen sucht, für die nur ein (verbleibender) Schlüssel infrage kommt. Zunächst kann so nur □ an Position 5 platziert werden, anschließend werden ☆ und ♢ platziert und zuletzt hat △ drei Positionen ganz für sich allein und wird in irgendeine davon platziert. Falls die Greedy-Strategie funktioniert, nennen wir die Konfiguration **schälbar**.

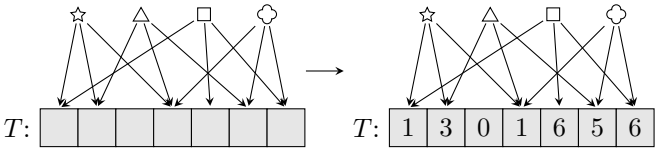
[1] Englischer Titel der Dissertation: „*Random Hypergraphs for Hashing-Based Data Structures*“.

[2] Institut für Informatik, Universität zu Köln, walzer@cs.uni-koeln.de (seit 1.12.2020).



(2)

Betrachten wir nun noch eine eigentümliche Art, die Information in f zu speichern:



(3)

Bildet man das bitweise exklusive oder (XOR) der Zahlen in den Positionen eines Schlüssels, erhält man hier den Wert, der dem Schlüssel zugewiesen ist. Für ◌ ist das $1 \oplus 5 \oplus 6 = (001)_2 \oplus (101)_2 \oplus (110)_2 = (010)_2 = 2$ also $f(◌)$. Wir nennen die Konfiguration links **lösbar**, denn die gewünschten Gleichungen können für alle Schlüssel gleichzeitig erfüllt werden (egal welche Werte f den Schlüsseln zuordnet). Es ist nicht schwierig zu beweisen, dass aus Schälbarkeit Lösbarkeit folgt und aus dieser wiederum Orientierbarkeit folgt.

Die Dissertation beleuchtet – grob gesagt – folgende Frage.

Q: *Angenommen Schlüsseln werden Positionsmengen unabhängig und zufällig zugewiesen. Unter welchen Umständen ist damit zu rechnen, dass die entstehende Konfiguration schälbar, lösbar oder orientierbar ist?*

2 Motivation: Hashing-basierte Datenstrukturen

Die Frage **Q** stellt sich im Kontext von hashing-basierten Datenstrukturen. Jedem Schlüssel ist durch eine **Hashfunktion** – einem einfachen pseudozufälligen Algorithmus – ein *Hashwert* zugeordnet, den wir gemäß eines Hashing-Schemas (siehe Seite 4) als Positionsmenge in T interpretieren.

Ein **Wörterbuch** ist ein abstrakter Datentyp zum Speichern einer Abbildung f. Eine **Cuckoo-Hashtabelle** [PR04; Fot+05] implementiert ein Wörterbuch und ist durch eine Hashfunktion sowie eine wie in **(1)** befüllte Tabelle gegeben. Beim *Suchen* nach dem Schlüssel △ erhalten wir durch Auswerten der Hashfunktion die Positionen $\{2,4,6\}$. An $T[6]$ erkennen wir dann $f(△) = 7$. Die Suche nach einem abwesenden Schlüssel ◇ geht ähnlich vonstatten. Nachdem wir ihn in keiner seiner zulässigen Positionen finden, können

wir schließen, dass er nicht zum Definitionsbereich von f gehört. Cuckoo-Hashing basierte Wörterbücher sind intensiv untersucht und meist besonders speichereffizient.

Eine **Retrieval-Datenstruktur** [DP08; BPZ13] ist eine *partielle* Implementierung des Wörterbuchdatentyps. Diese erlaubt es zwar die Funktion f auszuwerten, aber falls der angefragte Schlüssel nicht zum Definitionsbereich von f gehört, dürfen beliebige Werte herauskommen. Zumeist sind Retrieval-Datenstrukturen durch eine Hashfunktion und eine wie in **(3)** befüllte Tabelle gegeben. Für einen Schlüssel $\triangle$ im Definitionsbereich von f erhalten wir durch Auswerten der Hashfunktion dessen Positionsmenge $\{2,4,6\}$. Durch bitweises XOR der dort gespeicherten Zahlen, hier $3 \oplus 1 \oplus 5 = 7$, erhalten wir $f(\triangle)$. Für einen abwesenden Schlüssel $\diamond$ würden wir eine analoge Berechnung ausführen, ohne zu bemerken, dass dieser nicht zum Definitionsbereich von f gehört. So gesehen liefern Retrieval-Datenstrukturen nur die bedingte Information der Form „*falls* $\triangle$ zum Definitionsbereich gehört, *dann* gilt $f(\triangle) = 7$“. Ein Vorteil ist, dass in T keine Schlüssel gespeichert werden müssen, was eine enorme Platzersparnis bedeuten kann, wenn Werte eine wesentlich kürzere Bitdarstellung haben als Schlüssel.

Die vielleicht wichtigste Anwendung von Retrieval-Datenstrukturen liegt in der Konstruktion von **Filtern**, genauer *Approximate-Membership-Query-Filtern*. Für eine Menge M, sagen wir $M = \{\star, \triangle, \square, \hexagon\}$, interessieren uns Fragen der Form „ist $x \in M$?“. Wir wählen eine Hashfunktion f mit Wertebereich $R = \{0, \dots, 2^r - 1\}$, sagen wir $r = 3$. Die Zuordnung $\{x \mapsto f(x) \mid x \in M\}$ speichern wir nun als Retrieval-Datenstruktur, merken uns aber zusätzlich f als Hashfunktion. Für jeden Schlüssel gibt es nun zwei Wege, um einen Wert aus R zu berechnen: Erstens durch direkte Auswertung von f und zweitens durch Befragen der Retrieval-Datenstruktur. Für Schlüssel aus M erhalten wir übereinstimmende Resultate (z.B. 3 für $\star$). Für einen Schlüssel $\diamond \notin M$ ist das dagegen nur mit Wahrscheinlichkeit $1/|R| = 2^{-r}$ der Fall, da $f(\diamond)$ bei der Konstruktion der Retrieval-Datenstruktur keine Rolle gespielt hat. Auf eine Frage „ist $x \in M$?“ können wir also „$x \notin M$“ antworten, wenn keine Übereinstimmung vorliegt und „$x \in M$ ist wahrscheinlich“ sonst.
Seit der ursprünglichen Idee von Bloom [Blo70] entwickelte sich eine kaum noch zu überschauende Literatur zu Bloom-Filter-Varianten. Trotz der Möglichkeit falsch-positiver Antworten sind diese aufgrund ihres geringen Speicherverbrauchs und ihrer schnellen Anfragezeit aus der Praxis kaum noch wegzudenken [BM03]. Zum Beispiel können in Datenbanksystemen durch die zuverlässigen negativen Antworten oft aufwändige Anfragen an große Hintergrundspeicher vermieden werden.

3 Hashing-Schemata und Schwellwertphänomene

Orientierbarkeit wie in **(1)** ermöglicht also die Konstruktion von Cuckoo-Hashtabellen und Lösbarkeit wie in **(3)** die Konstruktion von Retrieval-Datenstrukturen. Schälbarkeit ermöglicht in beiden Fällen die Konstruktion durch einen einfachen *Schälalgorithmus*. Frage **Q** von vorhin hat aus Datenstrukturperspektive also folgende Entsprechung:

Q′: *Unter welchen Umständen ist zu erwarten, dass die Konstruktion von Cuckoo-Hashtabellen und Retrieval-Datenstrukturen gelingt? Unter welchen Umständen ist zu erwarten, dass ein Schälalgorithmus verwendet werden kann?*

Wir wollen die Frage $\mathbf{Q}'$ noch weiter konkretisieren. Mit „Umständen“ meinen wir zweierlei. Erstens das Verhältnis $c = \frac{m}{n} \in [0,1]$ zwischen der Anzahl m von Schlüsseln und der Größe n der Tabelle T. Wir nennen c **Auslastungsfaktor**. Zweitens meinen wir das **Hashing-Schema**. Dieses legt fest, wie der rohe Hashwert eines Schlüssels x – für gewöhnlich eine lange Bitfolge – als Positionsmenge $e(x) \subseteq \{1,\dots,n\}$ in T interpretiert wird. Das Hashing-Schema kann dabei für $e(x)$ eine innere Struktur vorsehen. Folgende Schemata sind üblich:

Voll-zufällig. Gegeben ist ein $k \in \mathbb{N}$. Die Menge $e(x)$ ist eine uniform zufällige Teilmenge von $[n]$ der Größe k. Das obige Beispiel ist voll-zufällig mit $k = 3$.

Doppel-Hashing. Gegeben ist ein $k \in \mathbb{N}$. Die Menge $e(x)$ ist eine zufällige arithmetische Folge mit k Gliedern, das heißt $e(x) = \{a, a+b, a+2b, \dots, a+(k-1)b\}$ wobei a und b zufällig sind und Indizes modulo n zu verstehen sind. Die Motivation ist hier eine Entlastung der Hashfunktion, die nur zwei statt k Zufallswerte produzieren muss.

(Nicht) ausgerichtete Blöcke. Gegeben sind $k, \ell \in \mathbb{N}$. Die Menge $e(x)$ enthält k zufällig gewählte *Blöcke* von jeweils ℓ aufeinanderfolgenden Positionen, also $k\ell$ Positionen insgesamt. Hintergrund hier ist die schnellere Zugriffszeit für zusammenhängende Speicherplätze durch Cache-Effekte. Oft wird verlangt, dass ℓ die Tabellengröße teilt und die Blöcke gemäß ganzzahliger Vielfache von ℓ *ausgerichtet* sind. *Nicht-ausgerichtete Blöcke* hingegen dürfen beliebig liegen.

Die Relativierung „ist zu erwarten“ in $\mathbf{Q}'$ lässt Raum für Unsicherheit. Die zufälligen Hashwerte könnten dergestalt sein, dass zu viele Schlüssel um dieselben Positionen in T konkurrieren und die Konstruktion der Datenstruktur fehlschlägt. Bei sehr großen Datenmengen mit Millionen von Schlüsseln rückt diese Unsicherheit allerdings in den Hintergrund. Es ergibt sich ein **Schwellwert** c^* für den Auslastungsfaktor sodass für $c < c^*$ die Konstruktion mit hoher Wahrscheinlichkeit gelingt, für $c > c^*$ dagegen mit hoher Wahrscheinlichkeit fehltschlägt.

Aus den Schwellwerten für Orientierbarkeit, Lösbarkeit und Schälbarkeit ergibt sich die Speichereffizienz, die Cuckoo-Hashtabellen und Retrieval-Datenstrukturen mit dem jeweiligen Hashing-Schema für große Datenmengen zuverlässig erreichen können und ob ein Schälalgorithmus die Konstruktion zuverlässig bewerkstelligen kann. Die Leitfrage der Dissertation hat daher folgende mathematische Konkretisierung:

$\mathbf{Q}''$: *Was sind die Schälbarkeits-, Lösbarkeits- und Orientierbarkeitsschwellwerte verschiedener Hashing-Schemata? Welche Schemata haben Schwellwerte nahe* 1*?*

4 Zentrale Ergebnisse

Die Dissertation hat 6 Hauptergebnisse. Jedes lässt sich als Schwellwertanalyse eines Hashing-Schemas auffassen. Im Fall der Ergebnisse D, E und F waren die Hashing-Schemata bereits bekannt – nicht aber vollständige Analysen. Die Hashing-Schemata aus A, B und C sind neu und wir wollen diese Ergebnisse hier stärker hervorheben.

4.1 Ergebnisse A, B und C

Höhere Schälbarkeitsschwellwerte durch Spatial Coupling. Der größte Schälbarkeitsschwellwert, der sich mit dem voll-zufälligen Hashing-Schema erreichen lässt, ist $c^* \approx 0.82$ für $k = 3$. Die Dissertation stellt einen einfachen Trick vor, der zu einem höheren Schwellwert führt: Für einen Parameter $0 < \varepsilon < 1$ besteht $e(x)$ nicht aus k zufälligen Positionen aus ganz $[n]$, sondern aus k zufälligen Positionen aus einem zufälligen Intervall der Größe εn. Die Positionen in $e(x)$ liegen also insbesondere nahe beieinander.

Da eine ähnliche Idee in der Kodierungstheorie unter dem Namen „Spatial Coupling“ kursiert, nennen wir dieses Hashing-Schema entsprechend. Ergebnis A lautet nun wie folgt:

Satz A *Der Schälbarkeitsschwellwert des Spatial-Coupling Hashing-Schemas konvergiert, für $\varepsilon \to 0$, gegen den Orientierbarkeitsschwellwert c_k^* des voll-zufälligen Hashing-Schemas.*

Da c_k^* für $k \to \infty$ gegen 1 konvergiert, lassen sich so Schälbarkeit und Auslastungsfaktoren beliebig nahe 1 vereinbaren. Auch für kleine k, etwa $k = 3$, ist der Schälbarkeitsschwellwert mit $c_3^* \approx 0.9179$ deutlich größer als im voll-zufälligen Fall.

Lösbarkeitsschwellwerte nahe 1. Wir betrachten ein neues Hashing-Schema mit *zwei ausgerichteten Block-Teilmengen*, eine auf Lösbarkeit zugeschnittene Variante von ausgerichteten Blöcken. Bei Blockgröße ℓ enthält $e(x)$ nicht alle 2ℓ Positionen der beiden Blöcke, sondern nur eine uniform zufällige Teilmenge davon.

Satz B *Das Hashing-Schema mit zwei ausgerichteten Block-Teilmengen bei Blockgröße $\ell \geq \log_2 n$ hat einen Lösbarkeitsschwellwert von* 1.[1]

Tatsächlich kann man sogar mit *einer* Block-Teilmenge Lösbarkeitsschwellwerte nahe 1 erreichen, wenn der Block nicht ausgerichtet ist.

Satz C *Das Hashing-Schema mit* einer *nicht-ausgerichteten Block-Teilmenge bei Blockgröße $\ell \geq \frac{\log_2 n}{\varepsilon}$ hat einen Lösbarkeitsschwellwert von $1 - O(\varepsilon)$.*

4.2 Anwendungen von A, B und C für platzsparendes Retrieval und Filter

Wir wollen die Relevanz der Ergebnisse A, B und C am Beispiel von Filtern erklären, die sich aus den entsprechenden Retrieval-Datenstrukturen ergeben.

Besonders einfach fällt dies für **Ergebnis A**, denn in der Literatur gibt es bereits Filter, die auf dem voll-zufälligen Hashing-Schema mit $k = 3$ basieren und einen Schälalgorithmus zur Konstruktion verwenden. Diese heißen Xor-Filter und werden als platzsparende Alternativen zu Bloom-Filtern gehandelt [GL19]. Durch Wechsel auf das Spatial-Coupling Hashing-Schema verbessert sich der Auslastungsfaktor von etwa 81% auf etwa 91%, ohne dass sich ansonsten etwas Wesentliches an der Datenstruktur ändert. Mit Wahl von $k > 3$ kann der Auslastungsfaktor bei leichter Verschlechterung der Laufzeit theoretisch beliebig nahe an 100% gebracht werden. Es gibt bereits Implementierungen unter dem Namen „Fuse-Filter“, die allerdings noch an einem ärgerlichen Problem leiden: Die hohe

[1] Das Ergebnis der Dissertation ist noch detaillierter und erlaubt $n = m + O(\log n)$ oder sogar $n = m$.

Speichereffizienz beginnt sich erst bei großen Datensätzen von 10^5 und mehr Schlüsseln einzustellen.

Ergebnisse B und C liefern Retrieval-Datenstrukturen und damit auch Filter mit Auslastungsfaktoren nahe 100% und geringen Anfragezeiten. Um den Sprung in die Praxisrelevanz zu schaffen, ist allerdings noch die Hürde einer geringen Konstruktionszeit zu nehmen. In beiden Fällen gilt es hier ein lineares Gleichungssystem über dem Körper $\mathbb{F}_2$ (ohne Rückgriff auf Schälbarkeit) zu lösen. Die Variablen entsprechen den Tabelleneinträgen und jeder Schlüssel liefert eine Gleichung, die sich gemäß $e(x)$ auf die Variablen bezieht. Abb. 1 zeigt die Form dieser Gleichungssysteme in Matrixdarstellung.

Für das Gleichungssystem gemäß **Ergebnis B** (Abb. 1 **(a)**) können wir leider selbst unter Verwendung diverser Tricks zur Lösung dünn-besetzter Gleichungssysteme keine überzeugenden Geschwindigkeiten erreichen. Die Konstruktionszeit ist in unseren Experimenten rund 10-mal höher als bei gängigen Alternativen.

Viel besser sieht es aus für Filter gemäß **Ergebnis C**. Die theoretische Konstruktionszeit liegt bei $O(n/\varepsilon^2)$, was ausnutzt, dass das Gleichungssystem (Abb. 1 **(b)**) nach einer Umsortierung der Zeilen schon beinahe Stufenform hat. In der Praxis ist die Konstruktionszeit durch nahezu perfekte Cache-Effizienz (für moderate $\varepsilon \approx 5\%$) sogar *besser* als bei manchen Ansätzen, die mit einem „$O(n)$“ Schälalgorithmus Schnelleres vermuten ließen.

(a)
$$\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 010 & & 110 & \\ 100 & & & 110 \\ 111 & & & 010 \\ 010 & 110 & & \\ & 111 & & 001 \\ & & 001 & \end{pmatrix} \end{array} \cdot \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

(b)
$$\begin{array}{c} \begin{array}{cccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \end{array} \\ \begin{pmatrix} & 0 & 1 & 1 & 0 & 0 & & & & & & & \\ 1 & 0 & 1 & 0 & 0 & & & & & & & & \\ & & & 1 & 0 & 0 & 0 & 1 & & & & & \\ & & & & & & & & 1 & 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 & 1 & 1 & & & & & & & \\ & & & & & & 1 & 1 & 1 & 0 & 0 & & \end{pmatrix} \end{array} \cdot \vec{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

Abb. 1: Form der zu lösenden Gleichungssysteme bei Konstruktion einer Retrieval Datenstruktur.
(a) Mittels Ergebnis *B* mit $n = 12$, $L = 3$ und $r = 1$.
(b) Mittels Ergebnis *C* mit $n = 13$, $L = 5$ und $r = 1$.

In Abb. 2 zeigen wir Konstruktionszeiten und Speicher-Overhead der Retrieval-Datenstrukturen aus A, B und C sowie dreier Konkurrenten für $n = 10^7$ und $r = 1$. Speicher-Overhead ε bedeutet einen Speicherverbrauch von $(1+\varepsilon)\text{OPT}$, wobei OPT das informationstheoretische Optimum ist. Ein Auslastungsfaktor von c führt zu $\varepsilon \approx \frac{1}{1-c} - 1$.

4.3 Ergebnisse D, E und F

Ergebnisse D, E und F betrachten bekannte Hashing-Schemata, deren Schwellwerte (und weitere Eigenschaften) bereits empirisch untersucht waren. Die Dissertation liefert eine mathematische Analyse und exakte Schwellwerte und vertieft somit das theoretische Verständnis der Hashing-Schemata.

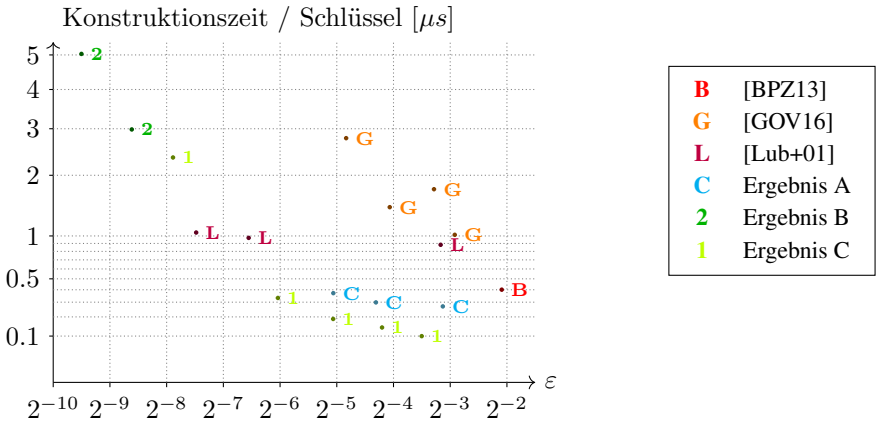


Abb. 2: Speicher-Overhead und Konstruktionszeiten unserer und fremder Retrieval-Datenstrukturen.

Ergebnis D. Wir bestimmen exakte Schwellwerte für das für das Hashing-Schema mit nicht-ausgerichteten Blöcken. Bei gleichem k und ℓ liegen diese höher als bei der Variante mit ausgerichteten Blöcken. Für $k = 2$ gibt es hierzu Vorarbeiten [LP09].

Ergebnis E. Wir beweisen, dass das voll-zufällige Hashing-Schema und Doppel-Hashing (bei gleichem k) die selben Orientierbarkeitsschwellwerte aufweisen. Dies ergänzt einen unvollständigen Beweis [Lec13].

Ergebnis F. Wir bestimmen Orientierbarkeitsschwellwerte für ein spezielles Hashing-Schema aus [MS17], das für dynamisch wachsende Hashtabellen geeignet ist.

5 Methoden

An den Bildern in **(1)**–**(3)** sieht man schon, wie eine Konfiguration aus Schlüsseln und Tabellenpositionen als bipartiter Graph modelliert werden kann. Noch nützlicher ist eine Terminologie mit Hypergraphen. Jede Tabellenposition entspricht einem Knoten und jeder Schlüssel einer Menge von Tabellenpositionen – also einer Hyperkante. Diskutieren wir Lösbarkeit, betrachten wir meist Inzidenzmatrizen wie in Abb. 1. Große Teile der Arbeit sprechen daher die Sprache der Graphentheorie und der linearen Algebra. Hinzu kommen die Standardwerkzeuge der Wahrscheinlichkeitstheorie.

Für Ergebnis B genügt das bereits. Die anderen Ergebnisse benötigen ausgefalleneres Rüstzeug, darunter die folgenden beiden Theorien. Diese tragen dort, wo sie zum Einsatz kommen, einen erheblichen Teil der Beweislast.

Die *Objective Method* und Lokal-Schwache Konvergenz [AS04; Lel12]. Es ist möglich, dass eine Folge $(G_n)_{n\in\mathbb{N}}$ von Graphen mit divergierender Knotenzahl $|V(G_n)| = n$ bezüglich lokaler Grapheigenschaften schwach konvergiert.[2] Für die gemäß eines Hashing-Schemas

[2] Das heißt formal: Ist $X_{n,r}$ die r-Nachbarschaft eines zufälligen Knotens von G_n, so konvergiert $(X_{n,r})_{n\in\mathbb{N}}$ in Verteilung für alle $r \in \mathbb{N}$.

gewonnene Folge von (Hyper-)Graphen wachsender Größe ist dies überwiegend der Fall. Im Geiste der *Objective Method* [AS04] arbeiten wir in asymptotischen Analysen mit einem entsprechenden unendlichen Grenzobjekt anstelle von unendlichen Folgen endlicher Graphen. Dies erlaubt es uns einen Satz von Lelarge [Lel12] zu verwenden, der Orientierbarkeitsschwellwerte über Eigenschaften solcher Grenzobjekte charakterisiert. Der Satz spielt eine zentrale Rolle für Ergebnisse D, E und F, sowie eine Nebenrolle für Ergebnis A.

Schwellwert-Saturierung mit Spatial-Coupling [KRU15]. Wir verwenden eine Idee aus dem scheinbar fernen Gebiet der *Kodierungstheorie*. Eine zentrale Frage betrifft dort den Anteil an Redundanz, den ein Sender in eine Nachricht einbauen muss damit diese vom Empfänger dekodiert werden kann, obwohl eine gewisse Rate an Information während der Übertragung verloren geht. Die Klasse der *Low-Density-Parity-Check-Codes* kann mittels ähnlicher Hypergraphen modelliert werden wie unsere Hashing-Schemata und es ergeben sich ähnliche Schwellwertphänomene für die erreichbar Informationsdichte. Der Schwellwert ist für den schnellen und einfachen *Belief-Propagation*-Decoder (BP-Decoder) meist niedriger als für den ideale *Maximum-Aposteriori-Probability*-Decoder (MAP-Decoder).

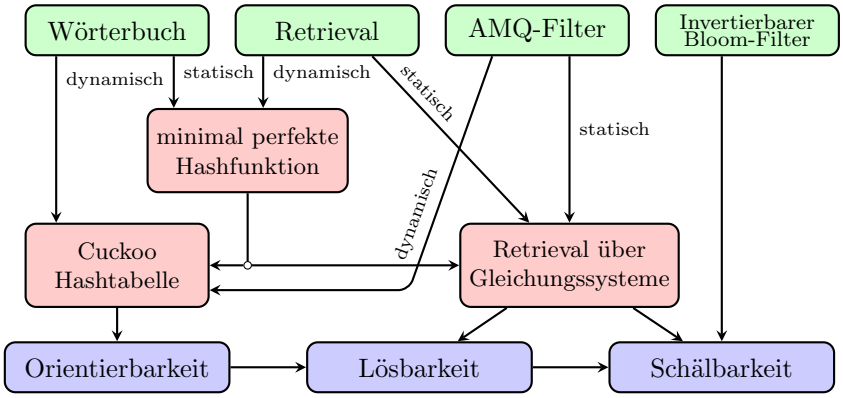
Spatial Coupling versieht nun Codes mit einer linearen Geometrie, wodurch diese an Symmetrie verlieren. Unter gewissen Umständen hebt das den BP-Schwellwert $c^{\triangle}$ auf den MAP-Schwellwert c^* an. Übersetzt auf unseren Fall hebt die Technik den Schälbarkeitsschwellwert $c^{\triangle}$ auf den Orientierbarkeitsschwellwert c^* an – genau wie in Ergebnis A behauptet. Das Phänomen ist unintuitiv. Folgendes gleichermaßen unintuitives physikalisches Phänomen kann uns als Analogie dienen. Sei $c^* = 0\,^\circ\mathrm{C}$ und $c^{\triangle}$ die „Temperatur, bei der Wasser gefriert“. Sehr reines Wasser kann (im Gegensatz zu Leitungswasser) auf Temperaturen weit unter $c^* = 0\,^\circ\mathrm{C}$ *unterkühlt* werden (bis zu $-48.3\,^\circ\mathrm{C}$) und dabei flüssig bleiben. Das liegt an fehlenden Kristallisationskeimen, an denen Eiskristalle wachsen könnten. Erst Unregelmäßigkeiten im Wasser *erhöhen* $c^{\triangle}$ auf c^*, sodass das Wasser bei $0\,^\circ\mathrm{C}$ gefriert.

Im Falle von Ergebnis A ergeben sich geeignete Unregelmäßigkeiten an den beiden Enden der Tabelle. Verklebte man sie zu einem Kreis (mit Indizes modulo n und Intervallen, die die Naht überspannen), bliebe die Schwellwertverbesserung völlig aus.

6 Fazit

Die Dissertation untersucht Hashing-Schemata zur Konstruktion von Wörterbüchern und verwandten Datentypen. Folgende Abbildung gibt eine Übersicht. Dort sind auch zwei Datenstrukturen erwähnt, die in dieser Zusammenfassung nicht zu Sprache gekommen sind. Die zugrundeliegenden Algorithmen sind größtenteils einfach. Die Herausforderung liegt hauptsächlich in der Analyse der erzielbaren Speichereffizienz, was ein mathematisches Verständnis von Schwellwertphänomenen bzgl. Orientierbarkeit, Lösbarkeit und Schälbarkeit voraussetzt.

Die Ergebnisse D, E und F der Arbeit liefern eine theoretische Analyse existierender Hashing-Schemata zur Verwendung bei Cuckoo-Hashtabellen. Ergebnisse A, B und C sind neue Hashing-Schemata, die sich (unter anderem) zur Konstruktion von Retrieval-Datenstrukturen und Filtern eignen. Der Autor ist erfreut darüber, trotz der theoretischen



Natur der Arbeit eine Nähe zur Anwendung zu erleben: Tatsächlich wurde ein auf Ergebnis C aufbauender Filter mittlerweile unter dem Namen Ribbon-Filter [Dil20] in Facebooks Datenbanksystem RocksDB als experimentelle Alternative zu den herkömmlichen (Blocked-Bloom-)Filtern eingebaut. Ergebnis A ist mit Fuse-Filtern auf der Schwelle dazu.

Literatur

[AS04] David Aldous und J. Michael Steele. The Objective Method: Probabilistic Combinatorial Optimization and Local Weak Convergence. In: *Probability on Discrete Structures*. Springer, 2004, S. 1–72. ISBN: 978-3-662-09444-0. DOI: 10.1007/978-3-662-09444-0_1.

[Blo70] Burton H. Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. In: *Commun. ACM* 13.7 (1970), S. 422–426. DOI: 10.1145/362686.362692.

[BM03] Andrei Z. Broder und Michael Mitzenmacher. Network Applications of Bloom Filters: A Survey. In: *Internet Mathematics* 1.4 (2003), S. 485–509. DOI: 10.1080/15427951.2004.10129096.

[BPZ13] Fabiano Cupertino Botelho, Rasmus Pagh und Nivio Ziviani. Practical Perfect Hashing in Nearly Optimal Space. In: *Inf. Syst.* 38.1 (2013), S. 108–131. DOI: 10.1016/j.is.2012.06.002.

[Dil20] Peter Dillinger. *RIBBON: A practical and near-optimal static Bloom alternative for RocksDB*. 2020. URL: https://www.youtube.com/watch?v=XfwxUBL8xT8&t=1h16m10s (besucht am 20. 11. 2020).

[DP08] Martin Dietzfelbinger und Rasmus Pagh. Succinct Data Structures for Retrieval and Approximate Membership (Extended Abstract). In: *Proc. 35th ICALP (1)*. 2008, S. 385–396. DOI: 10.1007/978-3-540-70575-8_32.

[Fot+05] Dimitris Fotakis, Rasmus Pagh, Peter Sanders und Paul G. Spirakis. Space Efficient Hash Tables with Worst Case Constant Access Time. In: *Theory Comput. Syst.* 38.2 (2005), S. 229–248. DOI: 10.1007/s00224-004-1195-x.

[GL19] Thomas Mueller Graf und Daniel Lemire. Xor Filters: Faster and Smaller Than Bloom and Cuckoo Filters. In: *CoRR* (2019). arXiv: 1912.08258.

[GOV16] Marco Genuzio, Giuseppe Ottaviano und Sebastiano Vigna. Fast Scalable Construction of (Minimal Perfect Hash) Functions. In: *Proc. 15th SEA*. 2016, S. 339–352. DOI: 10.1007/978-3-319-38851-9_23.

[KRU15] S. Kudekar, T. J. Richardson und R. L. Urbanke. Wave-Like Solutions of General 1-D Spatially Coupled Systems. In: *IEEE Trans. Inf. Theory* 61.8 (2015), S. 4117–4157. DOI: 10.1109/TIT.2015.2438870.

[Lec13] Mathieu Leconte. Double hashing thresholds via local weak convergence. In: *Proc. 51st Allerton*. 2013, S. 131–137. DOI: 10.1109/Allerton.2013.6736515.

[Lel12] Marc Lelarge. A New Approach to the Orientation of Random Hypergraphs. In: *Proc. 23rd SODA*. 2012, S. 251–264. DOI: 10.1137/1.9781611973099.23.

[LP09] Eric Lehman und Rina Panigrahy. 3.5-Way Cuckoo Hashing for the Price of 2-and-a-Bit. In: *Proc. 17th ESA*. 2009, S. 671–681. DOI: 10.1007/978-3-642-04128-0_60.

[Lub+01] Michael Luby, Michael Mitzenmacher, Mohammad Amin Shokrollahi und Daniel A. Spielman. Efficient Erasure Correcting Codes. In: *IEEE Trans. Inf. Theory* 47.2 (2001), S. 569–584. DOI: 10.1109/18.910575.

[MS17] Tobias Maier und Peter Sanders. Dynamic Space Efficient Hashing. In: *Proc. 25th ESA*. 2017, 58:1–58:14. DOI: 10.4230/LIPIcs.ESA.2017.58.

[PR04] Rasmus Pagh und Flemming Friche Rodler. Cuckoo Hashing. In: *J. Algorithms* 51.2 (2004), S. 122–144. DOI: 10.1016/j.jalgor.2003.12.002.

[Wal20] Stefan Walzer. Random Hypergraphs for Hashing-Based Data Structures. Diss. Technische Universität Ilmenau, 2020. URL: https://www.db-thueringen.de/receive/dbt_mods_00047127.

Stefan Walzer wurde 1988 in Frankfurt an Main geboren. Von 2008-2015 studierte er Informatik und Mathematik am Karlsruhe Institut für Technologie und hat in seinen Bachelor- und Masterarbeiten unter Betreuung von Maria Axenovich die diskrete Mathematik lieb gewonnen. In seiner Zeit als Doktorand an der TU Ilmenau von 2015-2020 forschte er mit Martin Dietzfelbinger an hashbasierten Datenstrukturen. Im Anschluss an seine Promotion im November 2020 hat er eine PostDoc-Stelle an der Universität zu Köln in der Gruppe von Christian Sohler angetreten. Privat interessiert er sich für Philosophie und ist Liebhaber von Brett- und Computerspielen.

Die Erforschung der Wahrnehmung durch die Augen: Von Blickverfolgung zur visuellen Markantheit und mentalen Bildsprache[1]

Xi Wang[2]

Abstract: Einerseits strömt eine große Menge visueller Information durch die Augen in das Gehirn. Andererseits, verraten die Augen auch eine beträchtliche Menge an Informationen. Die komplexe Kombination der verschiedenen Aufgaben der Augen bietet wertvolle Möglichkeiten sowohl menschliche visuelle Wahrnehmung, kognitive Prozesse und mentale Zustände zu verstehen, als auch dieses Wissen in praktischen Anwendungen zu nutzen. Diese Arbeit untersucht die Rolle von Augenbewegungen während dem nach innen als auch dem nach außen gerichteten Informationsfluss. Insbesondere fokussiert die Arbeit darauf wie Menschen 3D Objekte während aktiver Wahrnehmung betrachten und wie sie ihre Augen bewegen während sich nichts betrachten. Des Weiteren untersucht diese Arbeit wie Augenbewegungen in praktischen Anwendungen genutzt werden können. Diese Arbeit liefert Beiträge an der Schnittstelle zwischen Psychologie, Computergrafik und Mensch-Maschine-Interaktion und stellt entsprechende Bausteine für zukünftige Studien zur Verfügung.

Einführung

Jeden Tag wird eine enorme Menge an visuellem Inhalt vom Menschen konsumiert dessen gesamte Verarbeitung sich auf den Input der Augen verlässt - den Sinnesorganen des visuellen Systems. Durch Augenbewegungen gesammelte sensorische Eingaben liefern Informationen für viele kognitive Prozesse. Die Untersuchung von Augenbewegungen hat eine lange Geschichte in der umfangreiche Untersuchungen durchgeführt wurden, um Augenbewegungen während des Lesens, der Wahrnehmung von Szenen, der visuellen Suche und der Entscheidungsfindung zu untersuchen.

Gleichzeitig fungieren Augenbewegungen auch als Ausgabesignal, das dem zentralen Befehlsgeber im Gehirn folgt. Wie von Land und Hayhoe [LH01] zusammengefasst, besteht die Rolle von Augenbewegungen aus Lokalisierung, Richtung, Führung und Kontrolle, die alle durch Top-Down-Einflüsse gesteuert werden. Darüber hinaus zeigen Studien zur Pupillometrie, dass die Reaktionen der Pupillen mit der Wahrnehmung auf hoher Ebene verbunden sind und mit Veränderungen der kognitiven Zustände zusammen hängen. Viele Menschen glauben, dass es mehr zu erfahren gibt, wenn man sich in die Augen schaut - ganz nach dem alten Sprichwort: Augen sind das Fenster zur Seele.

Durch die bessere Zugänglichkeit zu entsprechenden Hardwaregeräten, ist die Aufzeichnung von Augenbewegungen in vielen Szenarien möglich geworden. Diese Fortschritte

[1] Englischer Titel der Dissertation:"Exploring Perception through The Eyes: from Eye Tracking to Visual Saliency and Mental Imagery".

[2] Technische Universität Berlin, nicole.xiwang@gmail.com

eröffnen neue Möglichkeiten, Augenbewegungen in natürlichen Umgebungen zu untersuchen und in praktischen Anwendungen zu nutzen. Die zwei Rollen unserer Augen sind eng miteinander verbunden. Der Versuch die Rolle der Augen als Signaleingabe von der Ausgabe-Funktionalität zu entkoppeln stellt eine grundlegende Herausforderung dar, um die gesamte Informationsverarbeitung besser zu verstehen.

In dieser Arbeit [Wa20] untersuchen wir sowohl den nach innen als auch den nach aussen gerichteten Anteil der Augen am Informationsfluss. Insbesondere untersuchen wir, wie Menschen 3D-gedruckte Objekte im Raum betrachten, wobei der Schwerpunkt auf der Rolle der Informationsaufnahme liegt (Teil II). Der 3D-Druck erlaubt es echte physikalische Reize zu realisieren, während die in Teil I beschriebenen Methoden es uns ermöglichen die Blickpositionen auf den gedruckten Objekten genau abzuschätzen. Im Gegensatz zu früheren Studien, die nur flache Bilder als Stimuli betrachten, bringt uns die Verwendung echter physischer Objekte einen Schritt weiter zum Verständnis des natürlichen Betrachtungsverhaltens. Des Weiteren untersuchen wir in Teil III die Möglichkeit, Informationen aus Augenbewegungen anhand des LAN-Paradigmas (Looking-at-Nothing) abzulesen, welches die enge Verbindung zwischen Augenbewegungen und mentalen Bildern aufzeigt. Augenbewegungen werden zuerst als neue Modalität in einer Bildsuche verwendet. Später schlagen wir eine Berechnungsmethode basierend auf Augenbewegungen vor, um experimentell zu untersuchen, was im episodischen Gedächtnis priorisiert wurde.

Zusammenfassend sind die drei sich ergänzenden Themen dieser Arbeit: (i) wie man Blickpositionen im 3D-Raum schätzt, (ii) wie Menschen 3D Objekte betrachten und (iii) wie Menschen ihre Augen bewegen, während sie sich bestimmte Inhalte vorstellen.

Teil I Videobasiertes Eye Tracking in 3D

Genauigkeit der monokularen Blickverfolgung auf 3D-Geometrie. Das Verständnis des Betrachtungsverhaltens von Menschen beim Betrachten von Objekten spielt eine wichtige Rolle in Anwendungen wie Datenvisualisierung, Szenenanalyse, Objekterkennung und Bilderzeugung. Das Betrachtungsverhalten kann durch Messen von Fixierungen mittels Eye Tracking analysiert werden. In der Vergangenheit wurden solche Experimente, insbesondere für Objekterkundungsaufgaben, mit flachen 2D-Stimuli durchgeführt, die auf einem Bildschirm dargestellt wurden. Da der visuelle Aufmerksamkeitsmechanismus des Menschen in 3D-Umgebungen entwickelt wurde, kann die Tiefe einen wichtigen Einfluss auf das Betrachtungsverhalten haben. Da unser Forschungsziel darin besteht, das Betrachtungsverhalten des Menschen für tatsächlich dreidimensionale Reize zu untersuchen, müssen wir in der Lage sein, 3D-Blickpositionen mit hoher Genauigkeit zu verfolgen.

Standard-Eye-Tracking-Setups bestimmen nur die Blickrichtung des Menschen. Der gebräuchlichste Ansatz zur Bestimmung der Betrachtungstiefe besteht darin, einen binokularen Eyetracker zu verwenden und die Augenvergenz zu messen. Die experimentelle Bestimmung der Tiefe anhand der binokularen Vergenz ist jedoch von Natur aus schlecht konditioniert. Obwohl nichtlineare Abbildungen verwendet werden können, um den Fehler zu reduzieren, erfordern diese eine komplexe Kalibrierung und rechenintensive Optimierung der Abbildung, während sie dennoch zu relativ großen Ungenauigkeiten führen.

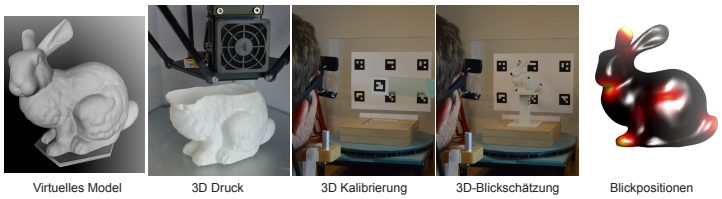


Abb. 1: Wir schätzen die genaue 3D-Blickpositionen, indem wir digitale Fertigung, Marker-Tracking und monokulares Eye-Tracking kombinieren.

Wir stützen unseren Ansatz auf eine Zuordnung zwischen den von einem Eye-Tracker erfassten Blickrichtungen und der physischen Welt, indem wir nicht nur feststellen, welches Objekt betrachtet wird, sondern auch die genaue 3D-Blickposition auf dem jeweiligen Objekt bestimmen. Die Hauptkomponenten, um eine so genaue Verfolgung zu erreichen, sind: (1) 3D-Stimuli werden durch digitale Fertigung erzeugt, sodass ihre Geometrie mit hoher Genauigkeit bekannt und auch in digitaler Form verfügbar ist, ohne die dargestellte Geometrie einzuschränken; (2) Marker werden in die 3D-Stimuli integriert, um die 3D-Position der Stimuli relativ zum Kopf zuverlässig und genau abzuschätzen; (3) Ein einfaches Kalibrierungsverfahren, mit dem die perspektivische Zuordnung von 3D-Positionen zu monokularen Pupillenpositionen genau berechnet werden kann; (4) Ein Fehlermodell für das Mapping ermöglicht die Berechnung plausibler Positionen auf dem 3D-Stimulus.

Die Ergebnisse zeigen, dass wir für typische Geometrien eine Winkelauflösung von 0,8° und zuverlässige Tiefenwerte innerhalb von $1,5\%$ des wahren Werts erhalten. Dies erreichen wir sogar in der Nähe von Silhouetten, wo die Tiefenschätzung auf Grund der starken Neigung der Geometrie besonders schwierig ist - und das trotz der Nutzung nur eines monokularen Eyetrackers und einem einfachen 11-Punkt-Kalibrierungsverfahren.

Der mittlere Vergenzpunkt ist unter Projektion fehlerbehaftet. Menschen neigen dazu, beide Augen ungefähr auf den gleichen Punkt im 3D-Raum zu richten. Vergenz ist die Bewegung beider Augen aufeinander zu oder voneinander weg, abhängig von der relativen Änderung vom vorherigen zum aktuellen Ziel. Oft wird angenommen, dass die Fixierungspunkte der beiden Augen perfekt ausgerichtet sind. Es es wurde aber auch gezeigt, dass die Augen zuerst divergieren, bevor sie während der Fixierungen am Blickpunkt konvergieren.

Es wurden viele Ansätze vorgeschlagen, um die Blickrichtung für jedes Auge im physischen Raum basierend auf aufgezeichneten Pupillenpositionen durch Eye-Tracking-Geräte abzuschätzen. Mit diesen Blickvektoren ist es möglich, den Blickpunkt auf reale dreidimensionale Reize zu rekonstruieren, indem ein oder beide Strahlen mit dem fixierten Objekt im Raum geschnitten werden (vorausgesetzt, seine Geometrie ist bekannt [Wa17]). Alternativ können wir versuchen, den Punkt zu finden, an dem sich die beiden Vektoren im Raum kreuzen. Allerdings schneiden sich zwei Blickvektoren im 3D-Raum normalerweise nicht. Selbst wenn der Betrachter mit beiden Augen einen Punkt im Raum betrachtet, enthalten die vom Eyetracker bereitgestellten Augenstrahlen aus verschiedenen Gründen Fehler: (i) Daten von Eye-Trackern weisen sowohl systematische als auch variable Fehler

(z.B. Rauschen) auf; (ii) Die Ungenauigkeit ist nicht konstant, sondern variiert mit der Pupillengröße und der Quantisierung der Hornhautreflexion (engl. “corneal reflection”, CR) in der Augenkamera; (iii) Idealerweise wird die Blickrichtung des Menschen nur so gesteuert, dass das Objekt in die Fovea gebracht wird, die eine nicht zu vernachlässigende Ausdehnung von $1.5 - 2^\circ$ hat; (iv) Es ist bekannt, dass beim binokularen Sehen viele Beobachter ein dominantes Auge haben, das genauer auf das Ziel gerichtet ist (in etwa 70% der Fälle das rechte Auge), und ein schwächeres Auge, das erheblich vom Ziel abweichen kann; (v) Die resultierende unbekannte und wahrscheinlich nichtlineare Abbildungsfunktion von verfolgten Pupillen- und CR-Zentren im Augenvideo zu Sichtlinien wird unter Verwendung von Polynomen niedriger Ordnung approximiert.

Aus diesen Gründen sind die beiden von Eye-Trackern erzeugten projizierten Augenstrahlen im Allgemeinen schief und haben keinen gemeinsamen Schnittpunkt im Raum. Um einen Punkt zu berechnen, der sich dem erwarteten Schnittpunkt der Strahlen annähert, besteht die natürlichste und am häufigsten verwendete Lösung darin, den Punkt zu berechnen, der in 3D den geringsten Abstand zu beiden Strahlen aufweist. Hier nennen wir diesen Punkt den Vergenzpunkt. Wir leiten die notwendigen Gleichungen für diese Berechnung ab und simulieren damit die Rekonstruktion von Vergenzpunkten bei systematischen (Genauigkeit) und variablen (Präzision) Fehlern. Anschließend entwickeln wir die mathematische Beschreibung der Ungenauigkeit von Blickvektoren, mit der die Auswirkung von Offsets und Rauschen auf den geschätzten Schnittpunkt simuliert wird. Schließlich präsentieren wir eine Methode, um den Schnittpunkt von Augenstrahlen besser abzuschätzen und die Position des fixierten Objekts in 3D zu rekonstruieren, wenn verrauschte Vergenzdaten vorliegen. Die gesamte Analyse für die Blickverfolgung gilt nicht nur im Raum, sondern auch auf ebenen Flächen, sofern die projektive Abbildung des Untergrunds im Modell verwendet wird.

Teil II Vergleichen von Augenbewegungen auf dem Bildschirm mit Augenbewegungen in 3D

Messung der visuellen Markantheit an 3D-Druckobjekten. Visuelle Markantheit beschreibt die Idee, dass bestimmte Merkmale eines visuellen Stimulus stärker hervorstechen als andere und eher die Aufmerksamkeit eines Betrachters auf sich ziehen. Für flache Reize wie 2D-Bilder haben zahlreiche Experimente gezeigt, dass menschliche Beobachter ihren Blick eher auf solche visuell hervorstechenden Merkmale richten. In der Literatur wird häufig davon ausgegangen, dass die in flachen Stimuli gefundene Markantheit mit der zugrunde liegenden 3D-Szene in Beziehung gesetzt werden kann. Obwohl diese Annahme intuitiv erscheinen mag, wurde sie bisher nicht experimentell validiert.

Um diese Annahme zu evaluieren, haben wir ein Experiment durchgeführt, in dem untersucht wurde, ob für *echte 3D-Stimuli* visuell hervorstechende Merkmale vorhanden sind. Wir verwendeten 3D-gedruckte Objekte als Stimuli und verfolgten die Blicke der Beobachter, während sie die Oberflächen der Objekte inspizierten.

Als nächstes analysierten wir die Fixierungsdaten, um zwei Fragen zu beantworten. Zuerst wurde getestet, ob menschliche Beobachter Konsistenz in ihren Fixierungsmustern für dasselbe Objekt zeigen. Eine solche Konsistenz wäre zu erwarten, wenn für physische Objekte visuell hervorstechende Merkmale vorhanden sind und wenn diese Merkmale das Fixierungsverhalten steuern. Zweitens wurde getestet, ob ein algorithmisches Modell der visuellen Markantheit, bekannt als *mesh saliency*, menschliche Fixierungen genau vorhersagen kann. Um die Konsistenz zwischen verschiedenen menschlichen Beobachtern zu testen, wurde eine neue Analysemethode vorgeschlagen, bei der die beobachteten Fixierungsmuster auf einem Objekt gegen Sequenzen von Fixierungen getestet werden, die nicht mit der Geometrie zusammenhängen, aber psychophysisch plausibel sind. Diese Testsequenzen wurden aus Fixierungssequenzen erzeugt, die für dasselbe Subjekt, aber *unter Verwendung eines anderen Objekts* aufgezeichnet wurden.

Unsere Ergebnisse zeigen eine höhere Übereinstimmung bei der Fixierung auf demselben Objekt zwischen Beobachtern im Vergleich zu generierten Testsequenzen. Dies weist auf das Vorhandensein visuell hervorstechender Merkmale auf 3D-Objekten hin. Das Ergebnis legt auch nahe, dass die Blickrichtungen systematisch und sinnvoll mit dem externen Reiz variieren, was eine notwendige Voraussetzung für weitere Analysen des menschlichen Betrachtungsverhaltens für 3D-Reize ist.

Da die Daten aussagekräftig zu sein scheinen, haben wir sie verwendet, um die Vorhersagekraft von *mesh saliency* [LVJ05] zu untersuchen - dem bisher einzigem Modell, das von einem psychophysischen Experiment unterstützt wird. Um Mesh Saliency mithilfe von Fixierungsdaten für echte 3D-Stimuli zu validieren, haben wir die algorithmischen Vorhersagen mit Permutationen der Werte über die Knoten des Netzes verglichen. Wenn Mesh Saliency tatsächlich ein guter Prädiktor für die Fixierungspositionen wäre, würde es eine bessere Leistung erreichen als die Permutationen. Unsere Ergebnisse zeigen, dass dies nicht der Fall ist.

Wir glauben, dass unser Experiment ein wichtiger erster Test für die Annahme ist, dass theoretische Konzepte der menschlichen Wahrnehmung, die aus Experimenten mit 2D-Bildern abgeleitet wurden, auch für die Wahrnehmung von 3D-Objekten gelten.

Verfolgen des Blicks auf Objekte in 3D: Wie sehen Menschen den Hasen wirklich? Ein großer Teil der Geometrieverarbeitung in der Computergrafik basiert auf *wahrnehmungsbasierten Metriken* und *visuell hervorstechenden Formmerkmalen*. Interessanterweise basieren die meisten Ansätze ausschließlich auf geometrischen oder informationstheoretischen Maßen. Diejenigen, die auf Experimenten basieren, verwenden fast ausschließlich Renderings von auf einem Bildschirm dargestellten Formen zur Bewertung [Bu11, Ki10]. Wir stellen fest, dass Daten von menschlichen Beobachtern, die physikalische Manifestationen von 3D-Formen untersuchen, einen festeren Boden für Rechenmodelle der menschlichen Wahrnehmung bieten würden.

Die Darstellung des Stimulus auf einem Bildschirm führt zu einem einfachen Versuchsaufbau. Es wurde argumentiert, dass nur die visuelle Wahrnehmung auf der Netzhaut von Bedeutung ist, so dass die Beschränkung der Reize auf Bilder ausreichen könnte, um die Markantheit von Merkmalen zu erkennen. Dieser Standpunkt wird immer mehr in Frage

gestellt. Wenn 3D-Formen auf virtuelle Umgebungen beschränkt wären, z.B. nur auf Bildschirmen erscheinen würden, dann bieten bildschirmbasierte Experimente natürlich den erforderlichen Einblick.

Das Sammeln von Punkten auf realen 3D-Formen basierend auf dem menschlichen Betrachtungsverhalten ist wesentlich aufwändiger als das Experimentieren mit einem Bildschirm zur Präsentation. Wir glauben, dass eine wichtige Frage ist, ob überhaupt eine Markantheit bzgl. schwacher geometrische Merkmale vorliegt. Dies würde bedeuten, dass ein Bereich auf einer Form von verschiedenen menschlichen Beobachtern, für verschiedene Oberflächenreflexionseigenschaften und über verschiedene Blickrichtungen hinweg fixiert wird. Aus diesem Grund haben wir uns bemüht, verschiedene Blickrichtungen (7 Richtungen 15° Grad voneinander entfernt) und Materialien (diffuses Pulver und vergleichsweise glänzender Kunststoff) für eine Reihe verschiedener Formen zu variieren. Die Beleuchtung ist auf eine diffuse Lichtquelle an einem festen Ort beschränkt. Die Daten werden im Allgemeinen nützlich sein, um vorhandene Rechenmodelle für geometrische Markantheit zu bewerten und solche Modelle (ähnlich zu den jüngsten Ansätzen für Bilder [KAB17]) direkt aus den Daten zu generieren.

Daten, die in diesem Setup von über 70 menschlichen Beobachtern gesammelt wurden, scheinen darauf hinzudeuten, dass hervorstechende Merkmale von der Blickrichtung abhängen, nicht jedoch von den beiden verschiedenen Materialien, die wir verwendet haben. Die visuelle Inspektion von Regionen, die in allen Blickrichtungen fixiert wurden, scheint mit semantisch bedeutsamen Teilen verbunden zu sein. Diese Beobachtungen deuten darauf hin, dass es schwierig ist, die visuelle Markantheit allein anhand geometrischer Merkmale vorherzusagen. Basierend auf diesen Beobachtungen bauen wir ein Convolutional Neural Network (CNN) auf, das die aus unseren Experimenten erzeugten Blickdichtekarten für eine bestimmte Form vorhersagen kann. In Übereinstimmung mit unseren experimentellen Ergebnissen lässt es sich nicht über Formen hinweg verallgemeinern, ist jedoch bei der Vorhersage von Markantheit besser als geometrische Ansätze wie Mesh Saliency [LVJ05]. Zusammenfassend leisten wir folgende Beiträge: (i) Wir entwickeln ein Setup für Eye-Tracking-Experimente an realen 3D-Formen, einschließlich einer genauen Registrierung, einem Kalibrierungsverfahren und der automatischen Zuordnung von binokularen Eye-Tracking-Daten zur Oberfläche von 3D-Formen; (ii) Wir stellen den ersten großen Datensatz mit Fixierungen an 3D-Formen zur Verfügung. Der Datensatz ist nützlich für die Bewertung von Wahrnehmungsmetriken und Markantheitsmaßen; (iii) Wir entwickeln eine neuartige Methode zur Analyse von Fixierungsverteilungen auf 3D-Formen; (iv) Wir zeigen, dass die Stabilität von Merkmalen von der Entfernung bzgl. der Betrachtungswinkel abhängt; (v) Wir entwickeln einen Ansatz des maschinelles Lernens, der es ermöglicht, die visuelle Markantheit des menschlichen visuellen Systems auf Objekten basierend auf ansichtsabhängigen Geometrieinformationen vorherzusagen.

Teil III Vergleichen von Augenbewegungen während der Codierung mit Augenbewegungen während des Erinnerns

Das Phänomen des Looking-at-nothing und die Datenerfassung. Die Erforschung von Augenbewegungen während des visuellen Erinnerns hat eine lange Geschichte. Frühe Arbeiten [Ja32, Pe10] haben intensive Aktivität der Augenbewegungen während des visuellen Erinnerns mit den entsprechenden mentalen Bildern verknüpft. Eine große Anzahl neuerer Studien [La14, SMK16] hat gezeigt, dass Menschen ihre Augen spontan bewegen, wenn sie sich an eine Szene aus dem Gedächtnis erinnern, und dass solche Augenbewegungsmuster sehr der räumlichen Anordnung der Elemente in der zurückgerufenen Szene ähneln. Dieser Effekt wurde für Teilnehmer demonstriert, die sich visuelle Szenen einprägen und diese später auf einem leeren Bildschirm abrufen sowie für Teilnehmer, die sich verbale Informationen in Verbindung mit einen räumlichen Hinweis merken und diese später abrufen, während Sie auf einen leeren Bildschirm schauen.

Frühere Studien zielten darauf ab, die Funktionalität von Augenbewegungen während des „Looking-at-Nothing" zu bewerten, insbesondere, ob sie beim Abrufen des Gedächtnisses eine funktionale Rolle spielen. Die Ergebnisse legen nahe, dass Augenbewegungen als „räumliche Indizes" fungieren, die bei der Erinnerung an das räumliche Layout einer Szene hilfreich sein können. Ob aber solche Augenbewegungen als zusätzliche Hinweise das Abrufen des Gedächtnisses erleichtern können, ist noch fraglich. Zur Unterstützung einer solchen funktionellen Rolle wurde in Experimenten, in denen die Teilnehmer während des Rückrufs nur ein Fixierungskreuz betrachten dürfen, eine beeinträchtigte episodische Gedächtnisleistung berichtet.

Studien zeigen, dass solche Augenbewegungen stark mit der räumlichen Anordnung des zurückgerufenen Inhalts korrelieren und als Erinnerungshinweise fungieren indem sie den Abrufvorgang erleichtern. Um dieses Phänomen zu untersuchen und die in den Augenbewegungen enthaltenen Informationen zu untersuchen, erstellen wir einen Datensatz mit aufgezeichneten Augenbewegungen während ein Foto sofort aus dem Gedächtnis abgerufen wird.

Das durch Gaze Tracking enthüllte mentale Bild Stellen Sie sich vor, Sie denken an Ihre Urlaubsfotos vom letzten Sommer. Nach 5 Sekunden erscheint ein Foto aus diesem Urlaub vor Ihnen, das dem Moment sehr ähnlich ist, an den Sie sich gerade erinnert haben.

Wir glauben, dass eine solche scheinbar magische Aufgabe durch die Verwendung von Augenbewegungen während der mentalen Bildgebung möglich ist. Die starke Ähnlichkeit zwischen Augenbewegungen während der Wahrnehmung und denen während des Erinnerns scheint die Möglichkeit zu eröffnen, Bildaugenbewegungen für die rechnergestützte Bildsuche zu verwenden: Das Bild wird ausgewählt, bei dem die Ähnlichkeit zwischen Augenbewegungen während der tatsächlichen Wahrnehmung und während des Erinnerns am größten ist (siehe Abbildung 2). Angesichts der sehr robusten Ergebnisse zu Augenbewegungen während des Erinnerns mag die rechnergestützte Bildsuche auf der Grundlage dieser Augenbewegungen zunächst als triviale Aufgabe erscheinen, es gibt jedoch vier grundlegendene Herausforderungen: (i) Obwohl Augenbewegungen beim Abrufen eines Bildes eine funktionale Rolle spielen, sind sie *nicht* identisch mit den Augenbewegun-

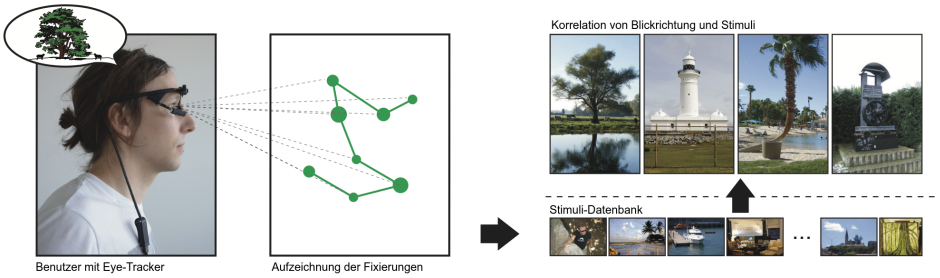


Abb. 2: Wir verwenden Augenbewegungen einer Person, die sich an ein Bild erinnert, während sie nichts betrachtet (in diesem Fall eine weiße Wand), um eine Reihe potenziell übereinstimmender Bilder aus einer Datenbank abzurufen.

gen, die während der Codierung des Bildes ausgeführt wurden; (ii) Es wurde gezeigt, dass Augenbewegungen während des Erinnerns möglicherweise durch verdeckte Aufmerksamkeit ausgelöst werden, was die Funktionalität von Augenbewegungen beim Looking-at-Nothing erklären kann; (iii) Es ist auch bekannt, dass einige Menschen das Abrufen eines Bildes mit geschlossenen Augen bevorzugen; (iv) Viele frühere Studien berichten, dass Augenbewegungen beim Looking-at-Nothing aufgrund des Fehlens eines Referenzrahmens eine große Variation enthalten.

Diese Technik der Bildsuche setzt voraus dass eine starke Ähnlichkeit zwischen den Augenbewegungen bei realen und imaginären Bildern bei verschiedenen Betrachtern besteht. Angesichts der Tatsache, dass eine solche Ähnlichkeit besteht, ihre Stärke jedoch nicht quantifiziert wurde, bewerten wir verschiedene Abrufszenarien. In allen Fällen fragen wir im Wesentlichen: Wie gut können Bilder rechnerisch von anderen Bildern allein auf Grundlage von Augenbewegungen unterschieden werden?

Wir entwickeln zwei Arten von Abrufalgorithmen für diese Szenarien. Wir beschränken uns auf die Verwendung räumlicher Histogramme der Daten und betrachten eine erweiterte Version der Earth-Mover-Distanz (EMD). Zumindest für die Szenarien, die genügend Daten liefern, verwenden wir auch tiefe neuronale Netze. Im Allgemeinen stellen wir fest, dass das Abrufen möglich ist, obwohl die Daten aus dem LAN eine Herausforderung darstellen und die resultierende Leistung je nach Szenario und Beobachter erheblich variiert.

Basierend auf den vielversprechenden Ergebnissen in einer Laborumgebung machen wir einen ersten Schritt in Richtung einer realen Anwendung: Mehrere Teilnehmer wurden mit einem mobilen Eyetracker in ein inszeniertes „Museum" geschickt. Nach ihrer Tour bitten wir sie, sich an einige Bilder zu erinnern, während sie ein leeres Whiteboard betrachten. Wir stellen fest, dass der Medianrank in einem Klassifizierungsansatz klein ist, was zeigt, dass die Idee für die praktische Anwendung vielversprechend ist.

Eine methodische Untersuchung des spontan priorisierten Bildinhalts. Aus Studien zur Veränderungsblindheit wissen wir, dass die Teilnehmer in einigen Situationen direkt auf ein aufgaben-irrelevantes Objekt schauen und dennoch keine Spur im Gedächtnis registriert werden kann. In dem bekannten Gorilla-Experiment in der Psychologie berich-

teten Beobachter, dass der Gorilla selbst dann nicht bemerkt wurde, wenn er über einen längeren Zeitraum fixiert wurde. Es ist bekannt, dass nicht alle Fixierungen gleich sind und nur einige zu einer Erinnerung an das fixierte Objekt führen. Dies ist theoretisch wichtig, hat darüber hinaus aber auch praktische Auswirkungen auf eine Vielzahl von Forschungsfeldern.

Das Hauptziel in diesem Teil der Arbeit ist, basierend auf Augenbewegungen beim Looking-at-Nothing die Fixierungen, an die man sich eher erinnert, von denen zu unterscheiden, an die man sich weniger erinnert. Wir nehmen die Übereinstimmung einer Fixierung in der Explorationsphase mit einer Fixierung aus dem Rückruf als Hinweis darauf, dass der entsprechende Szeneninhalt vorhanden ist und somit beim Rückruf aus dem episodischen visuellen Gedächtnis priorisiert wurde.

Diese Methode basiert auf der gut etablierten Erkenntnis des LAN-Paradigmas, dass sich unsere Augen bewegen, während wir ein Bild abrufen. Es gibt jedoch eine grundlegende Einschränkung des LAN-Paradigmas: Ohne den vorhandenen Stimulus gibt es keinen anderen physischen Bezugsrahmen als die Grenzen des Bildschirms. Die Orte der Fixierungen während des Rückrufs weisen eine signifikante lokale Verschiebung auf, d.h. die räumliche Reproduktion der Fixierungspositionen enthält Fehler. Diese Verformung des Bildraums wurde in der Literatur konsistent berichtet und beinhaltet das Schrumpfen, Verschiebung und überhaupt keine Augenbewegungen.

Die Ergebnisse des vorgeschlagenen Verfahrens legen nahe, dass die relative Bedeutung von Szenenelementen während des Abrufs in einigen Fällen anders als bei der Auswahl der Fixierungsziele priorisiert zu sein scheint. Nicht alle Fixierungen während der Codierung werden spontan aus dem episodischen Gedächtnis abgerufen. In den Bereichen Computervisualisierung und Mensch-Computer-Interaktion ist es häufig ein Ziel, ein spezifisches Design zu erstellen welches Unsere Ergebnisse deuten jedoch darauf hin, dass es einen Unterschied zwischen fixierter und erinnerter visueller Information gibt. Dies kann zu einer neuen Deutung der visuellen Relevanz in entsprechenden Anwendungen führen.

Zusammenfassung

Diese Arbeit untersucht die Rolle von Augenbewegungen bei der Informationsverarbeitung sowohl nach innen als auch nach außen. Augenbewegungen fungieren als eine Kombination verschiedener Rollen und es ist schwierig, diese voneinander zu trennen. Dies bleibt eine große Herausforderung für die zukünftige Arbeiten. Wir glauben, dass die Integration von Augenbewegungsaufzeichnungen und Daten aus anderen Modalitäten (z.B. Sprache, Geste, EEG usw.) neue und interessante Forschungsrichtungen bietet welche sowohl Grundlagenforschung als auch praktische Anwendungen umfassen. Die jüngsten Entwicklungen beim maschinellen Lernen, insbesondere des Deep Learnings, bieten leistungsstarke Rechenwerkzeuge, um die Vorteile von Big Data zu nutzen. Solche Berechnungsmethoden bieten die notwendigen Mittel, um den menschlichen Geist besser zu verstehen und das Verständnis zu nutzen, um Anwendungen zu entwickeln, welche die verschiedene Rollen von Augenbewegungen kombinieren um intelligente Werkzeuge zur Unterstützung des täglichen Lebens zu entwickeln.

Literaturverzeichnis

[Bu11] Bulbul, Abdullah; Capin, Tolga; Lavouè, Guillaume; Preda, Marius: Assessing Visual Quality of 3-D Polygonal Models. IEEE Signal Processing Magazine, 28(6):80–90, Nov 2011.

[Ja32] Jacobson, Edmund: Electrophysiology of mental activities. The American Journal of Psychology, 44(4):677–694, 1932.

[KAB17] Kruthiventi, Srinivas SS; Ayush, Kumar; Babu, Radhakrishnan Venkatesh: Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 2017.

[Ki10] Kim, Youngmin; Varshney, Amitabh; Jacobs, David W.; Guimbretière, François: Mesh Saliency and Human Eye Fixations. ACM Trans. Appl. Percept., 7(2):12:1–12:13, Februar 2010.

[La14] Laeng, Bruno; Bloem, Ilona M.; D'Ascenzo, Stefania; Tommasi, Luca: Scrutinizing visual images: The role of gaze in mental imagery and memory. Cognition, 131(2):263 – 283, 2014.

[LH01] Land, Michael F.; Hayhoe, Mary: In what ways do eye movements contribute to everyday activities? Vision Research, 41(25):3559 – 3565, 2001.

[LVJ05] Lee, Chang Ha; Varshney, Amitabh; Jacobs, David W.: Mesh Saliency. ACM Trans. Graph., 24(3):659–666, Juli 2005.

[Pe10] Perky, Cheves West: An experimental study of imagination. The American Journal of Psychology, 21(3):422–452, 1910.

[SMK16] Scholz, Agnes; Mehlhorn, Katja; Krems, Josef F: Listen up, eye movements play a role in verbal memory retrieval. Psychological research, 80(1):149–158, 2016.

[Wa17] Wang, Xi; Lindlbauer, David; Lessig, Christian; Alexa, Marc: Accuracy of Monocular Gaze Tracking on 3D Geometry. In (Burch, Michael; Chuang, Lewis; Fisher, Brian; Schmidt, Albrecht; Weiskopf, Daniel, Hrsg.): Eye Tracking and Visualization. Springer International Publishing, Cham, S. 169–184, 2017.

[Wa20] Wang, Xi: Exploring perception through the eyes : from eye tracking to visual saliency and mental imagery. Doctoral thesis, Technische Universität Berlin, Berlin, 2020.

Xi Wang ist derzeit Postdoktorand an der ETH Zürich. Sie schloss ihre Doktorarbeit an der TU Berlin mit Auszeichnung ab. Ihre Forschung befasst sich mit der menschlichen Wahrnehmung und ihren Anwendungen auf Computer Vision, Computergrafik und Mensch-Computer-Interaktionen. Ihre Projekte umfassen die Untersuchung der Vergenz von Augenbewegungen im 3D-Raum und welche Stellen realer Objekte Menschen besonders betrachten, sowie die Verwendung eines mentales Bildparadigmas, um mentale Bilder aus Augenbewegungen abzuleiten, während Menschen nichts betrachten. Ihr Hauptinteresse ist aus der menschlichen Wahrnehmung zu lernen und das Gelernte in Rechenmodellen und Anwendungen anzuwenden.

Qualitative Analyse des Wissenstransfers bei der Paarprogrammierung[1]

Franz Zieris[2]

Abstract: Bei der Paarprogrammierung (PP) arbeiten zwei Softwareentwickler/innen an einem Computer eng zusammen an einer technischen Aufgabe. Praktiker erhoffen sich davon eine Reihe von Vorteilen, wie etwa schnelleren Fortschritt, höhere Qualität und den Austausch von Wissen. Während die bisherige Forschung oft auf unmittelbar messbare Effekte aus Laborsituationen fokussiert war, die auftretenden großen Streuungen aber nicht erklären konnte, richtet sich meine Forschung auf das Verstehen der zu Grunde liegenden Mechanismen. Ich habe Videoaufzeichnungen von 27 industriellen PP-Sitzungen qualitativ analysiert und eine *Grounded Theory* des Wissenstransfers bei der PP erarbeitet: Zentral in PP-Sitzungen ist *aufgaben-spezifisches Wissen über das Softwaresystem*. Paare gleichen zunächst ihr diesbezügliches Vorwissen an, bevor sie gemeinsam fehlendes Wissen aufbauen. Transfer von *Wissen über Softwareentwicklung allgemein* spielt hingegen eine viel kleinere Rolle und erfolgt erst, wenn das Paar seine Bedürfnisse nach System-Wissen geregelt hat. Paare, die ihr gemeinsames Verständnis pflegen, können kurze, aber sehr produktive *Fokusphasen* haben; ist es zu schwach, droht hingegen ein *Zusammenbruch* des Paarprozesses.

1 Einführung

Die Idee, sich als Softwareentwickler/in nicht allein, sondern zu zweit über ein Programmier-Problem zu beugen, ist vermutlich so alt wie das Handwerk selbst. In den 1990er Jahren wurde sie u.a. von Coplien als *Muster* beschrieben [Co98, S. 294] und erlangte durch Beck im Rahmen der agilen Entwicklungsmethode *eXtreme Programming* [Be99] unter dem Namen *Pair Programming* bzw. *Paarprogrammierung* (PP) große Bekanntheit.

Durch diesen von engster Zusammenarbeit und ständiger Kommunikation geprägten Entwicklungsstil erhoffen sich Praktiker in der Industrie eine Reihe von Vorteilen [BN08]:

- Zwei Entwickler/innen verfügen über mehr Wissen, können so mehr Ideen produzieren, an komplizierteren Aufgaben arbeiten als jede/r allein und weniger Defekte und bessere Entwürfe in kürzerer Zeit hervorbringen.
- Fehlendes Wissen für Defektsuche und Systemverstehen lässt sich zu zweit schneller und zuverlässiger aneignen und im Gedächtnis behalten.
- Die Entwickler/innen können für zukünftige Aufgaben voneinander oder gemeinsam Neues lernen, was das Risiko von Wissensinseln im Team senkt.

[1] Englischer Titel der Dissertation: “Qualitative Analysis of Knowledge Transfer in Pair Programming” [Zi20]

[2] Freie Universität Berlin, zieris@inf.fu-berlin.de

Obwohl die Paarprogrammierung seit den 1990er Jahren Gegenstand zahlreicher wisenschaftlicher Untersuchungen war, gibt es bislang keine klaren Erkenntnisse darüber, in dem welchem Grad sich diese Vorteile tatsächlich einstellen und welche Rahmenbedingungen dafür entscheidend sind. Wie ich in Abschnitt 2 erläutern werde, liegen die Gründe dafür in einem simplistischen und teils dogmatischen Verständnis der Paarprogrammierung, sowie einer Versteifung auf quantitative Untersuchungen unter Laborbedingungen. In Abschnitt 3 beschreibe ich, wie ich mit einem qualitativen Forschungsansatz in der industriellen Praxis erhobene Daten analysiert habe – vorrangig Videoaufzeichnungen von Paarprogrammierungssitzungen, ergänzt durch Feldbeobachtungen und Interviews. Ich gebe einen kurzen Einblick in einen Teil meiner Ergebnisse und wie ich diese mit Praktikern validiert habe (Abschnitt 4) bevor ich in Abschnitt 5 meine Arbeit zusammenfasse.

2 Überblick über die Erforschung der Paarprogrammierung

Wissenschaftliche Studien zur Paarprogrammierung fallen grob in zwei Kategorien. Auf der einen Seite sind Untersuchungen in kontrollierten Umgebungen, in denen oft zufällig zusammengesetzte Paare kleine Programmieraufgaben lösen, um mit Einzelarbeitenden quantitativ verglichen zu werden (Abschnitt 2.1). Auf der anderen Seite stehen qualitative Analysen natürlicher Situationen, in denen komplexe Aufgaben von selbstorganisierten und spontan gebildeten Paaren bearbeitet werden (Abschnitt 2.2).

Ich klammere in der folgenden Diskussion Studien zur Paarprogrammierung aus, die nur auf Fragebögen und Interviews basieren, da durch die Art der Datenerhebung dort Detailtiefe und Nähe zu realen Ereignissen nicht gewährleistet sind.

2.1 Quantitative Studien zur Paarprogrammierung: Ergebnisse und Probleme

Über die Jahre wurden kontrollierte Experimente durchgeführt, in denen die Arbeitsgeschwindigkeit und -qualtität von Paaren mit der von Alleinarbeitenden verglichen wurde. Hannay et al.'s Meta-Analyse über 18 solcher Experimente [Ha09] zeigt zwar im Mittel einen signifikaten positiven Effekt der Paarprogrammierung auf *Qualität* und *Bearbeitungsdauer in Zeitstunden*,[3] allerdings *keinen* statistisch signifikanten Effekt auf den *Aufwand in Personenstunden*.[4] Bedeutsamer ist aber noch die von Hannay et al. ermittelte *große Heterogenität* zwischen den Effekten der Primärstudien, die darauf hinweist, dass diese Studien womöglich wesentlich verschiedene Dinge gemessen haben und nicht vergleichbar sind. Die Schlussfolgerung von Hannay et al.: *Paarprogrammierung* sei nicht per se besser oder schlechter, sondern hänge von noch nicht verstandenen situativen Faktoren ab, wie etwa Aufgabenkomplexität, Entwicklererfahrung (im Programmieren *und* im Paarprogrammieren), Motivation und Teamklima.

[3] Kleine bzw. mittlere Effektgröße, 95%-Vertrauensbereich von Hedges $g = [0.07, 0.60]$ bzw. $g = [0.13, 0.94]$.
[4] 95%-Vertrauensbereich: $g = [-1.18, 0.13]$

Allerdings konnte selbst das großangelegte Experiment von Arisholm et al. [Ar07] trotz vieler Versuchspersonen (295 professionelle Softwareentwickler) keinen klaren moderierenden Einfluss von *Aufgabenschwere* (konkret: einfache vs. komplexe Architektur) und *Programmiererfahrung* auf die Bearbeitungsdauer und Korrektheit der Lösungen von Paaren im Vergleich zu Solos nachweisen. Ingesamt gehen die Probleme quantitativer PP-Studien jedoch noch über nicht verstandene Moderator-Variablen hinaus:

1. Es gibt zu viele potenzielle Einflussgrößen als dass sie realitischerweise in kontrollierten Experimenten durchgeprüft werden könnten. Das obige Experiment ist trotz immensen Aufwands an der Analyse nur *zweier* Faktoren gescheitert.
2. Die Experimentsituationen unterscheiden sich relevant von der industriellen Praxis, sodass etwaige Ergebnisse ohnehin nicht übertragbar wären: Die Versuchspersonen haben oft wenig Erfahrung mit dem Paar-Arbeitsmodus; werden an unbekannte Systeme an vorgegebene Spielzeugaufgaben[5] gesetzt, ohne entscheiden zu können, ob sie Paararbeit hier überhaupt für sinnvoll halten; und das mit ihnen zugewiesenen Partnern, anstatt dass sich Paare dynamisch aus dem Projektalltag ergeben.
3. Es wird stillschweigend angenommen, dass "Paarprogrammierung" etwas kanonisches ist, das Entwickler/innen einfach "tun" können. Wie aber geht Paarprogrammierung 'richtig'? Diskutiert man erst eine Reihe von Ideen und wählt dann sorgfältig eine aus? Oder folgt man der u.a. von Williams & Kessler [WK02] vorgeschlagenen Rollenteilung in aktiven "Driver" und kontrollierenden "Navigator"? Verfolgt man die Gedanken eines Partners so lange bis man in eine Sackgasse gerät und wechselt dann die Führung? Oder verfolgt ein Partner nur still das Geschehen, bis ihm ein Problem auffällt? Verschiedene Paare werden verschiedenen Prozessmustern folgen.

In der Konsequenz ist es nicht zielführend Solo- und Paarprogrammierung in künstlichen Umgebungen nur anhand summativer Ergebnis-Metriken quantitativ zu vergleichen. Für ein Verständis darüber, wie und wann Paarprogrammierung funktioniert, sind *qualitative* Forschungsmethoden nötig, die einerseits den *Prozess* in den Blick nehmen und andererseits die Praktik dort untersuchen, wo sie letztlich – übrigens ohnehin unabhängig von experimentellen Ergebnissen – eingesetzt wird: In echten Projekten in der Industrie.

2.2 Qualitative Studien zur Paarprogrammierung: Ergebnisse und Probleme

Einige qualitativ-quantative Studien haben industrielle (z.B. [BRB08]) oder industrie-nahe (z.B. [WH09]) PP-Sitzungen aufgezeichnet, Dialoge transkribiert, gelabelt und Zählstatistiken ausgewertet. Diese Studien zeigen, dass es eine ständige Kommunikation zwischen den Partnern gibt, und dass es bei der Art der Sprachbeiträge *keine* systematischen Unterschiede

[5] Das "komplexe" Experiment-System von Arisholm et al. [Ar07] umfasste z.B. nur 12 Klassen mit 287 Zeilen Code; die Änderungen dauerten im Mittel gerade einmal 60 Minuten.

zwischen vermeintlichen “Drivern” und “Navigatoren” gibt. Allerdings benutzen diese Studien ebenfalls summative Metriken, bei denen ganze PP-Sitzungen als einzelne Datenpunkte unter Vernachlässigung zeitlicher Bezüge betrachtet werden. Ohne Prozessdimension sind diese Forschungsansätze ebenfalls nicht geeignet, die Paarprogrammierung zu erklären.

Qualitative PP-Studien, die auf Theoriebildung statt Hypothesentests ausgerichtet sind, gehen weiter und charakterisieren PP-Prozesse auf einer konzeptionellen Ebene. Sie zeigen z.B. dass die Kommunikation guter Paare nach Mustern verläuft (z.B. implizite vs. explizite Erklärungen [Pl15] oder *Restarting*, *Planing*, *Action* [ZHR13]). Ein zentrales Problem der Ergebnisse vieler solcher Arbeiten ist allerdings, dass sie eher *beschreibende Taxonomien* von PP-Aspekten sind als dass sie auf eine *erklärende Theorie* der PP hinarbeiten, und so kaum weiterführende Forschung erlauben. Auch fehlt oft eine Praxisorientierung, die einen Nutzen für die Anwendung in der Industrie zumindest in Aussicht stellt.

3 Zielsetzung, Datengrundlage und Forschungsmethode

Das Ziel meiner qualitativen Forschung ist es, zu verstehen wie Wissenstransfer bei der Paarprogrammierung tatsächlich funktioniert. Es geht nicht darum zu klären, ob Paarprogrammierung “besser” ist, sondern Praktikern, die paarprogrammieren möchten, Hilfestellung zu geben, um Probleme zu vermeiden und Potentiale zu nutzen. Meine qualitative Forschung folgt der *Grounded Theory Methodology* (GTM) nach Strauss & Corbin [SC90] und baut auf die Vorarbeit von Salinger [Sa13] auf, der ein “Vokabular” zur Charakterisierung der Basisaktivitäten in einem PP-Prozess entwickelt hat.

Datengrundlage sind seit 2007 von Kollegen und mir in 13 Firmen gesammelte Aufzeichnungen von PP-Sitzungen,[6] in denen professionelle Softwareentwickler/innen selbstbestimmt an ihren alltäglichen Aufgaben arbeiten. Reflektierende Interviews, Gruppendiskussionen und Workshops zur Bewertung und Einordnung der Erkenntnisse ergänzen das Material. Ich habe der GTM-Methode des *Theoretischen Samplings* [SC90] folgend, iterativ insgesamt 27 Sitzungen aus 10 Firmen mit verschiedenen Technologiestacks und Anwendungsdomänen, sowie 29 Entwickler/inne/n verschiedener Erfahrungsstufen gewählt (siehe Tab. 1).

Meine Analysen habe ich nicht auf Transkripten, sondern direkt auf den Videos durchgeführt. Beobachtungsnotizen und reflektierende Gespräche mit den Entwickler/inne/n am Tag nach einer Aufzeichnung dienten v.a. zur Einordnung der Geschehnisse und der Vervollständigung von Kontextinformationen. In meiner Analyse habe ich die GTM-Praktiken des *offenen*, *axialen*, und *selektiven Kodierens* [SC90] angewendet, wobei ich mich schrittweise von der Ebene tausender einzelner Äußerungen, über hunderte Wissenstransfer-Episoden (einige Sekunden bis Minuten), zu Episoden-Clustern bis hin zur Gesamtsitzungsdynamik vorarbeitete, bis eine *theoretische Sättigung* [SC90] meiner Konzepte erreicht war. Im Folgenden gebe ich einen kurzen Einblick in einen Teil meiner Ergebnisse.

[6] Bestehend aus Bildschirmvideo, Webcam und Audio; Details sind ausführlich in einem technischen Bericht [ZP20] beschrieben.

Firma mit Anwendungsdomäne und Programmiersprachen	**Sitzungen**		**#Entwickler**
A: Content-Management-System (Java, Objective-C, SQL)	1	02:22h	2
B: Social Media (PHP, JavaScript, SQL, HTML, CSS)	3	06:30h	2
C: Geoinformationssystem (Java)	5	07:49h	8
D: Customer-Relationship-Management (Java, XML)	1	02:24h	2
E: Logistik (C++, XML)	1	01:17h	2
J: Rundfunk-Datenmanagement (Java)	2	02:22h	2
K: Immobilienplattform (Java, SQL, CoffeeScript)	4	05:52h	3
M: Datenanalyse in Energie & Logistik (SQL)	1	00:25h	2
O: Online-Projektplanung (CoffeeScript)	4	05:11h	3
P: Online-Autoteilehändler (PHP, SQL)	4	05:41h	3
SUMME	27	39:53h	29

Tab. 1: Kontexte und Umfang der analysierten Sitzungsaufzeichnungen (Details in [ZP20]).

4 Überblick über die Ergebnisse

4.1 Paarhaftigkeit und Prozessflüssigkeit

Unter den analysierten Paaren gibt es große Unterschiede in Bezug auf die **Flüssigkeit** (*Fluency*) des Fortschrittes. Manche Paare haben **Fokusphasen** während derer es praktisch keine Sprechpausen gibt und die Partner gegenseitig ihre Gedanken, teilweise sogar Sätze und Programmcodezeilen ergänzen. Es ist unmöglich die Dynamik einer Fokusphase auf Papier greifbar zu machen. Als Näherung sind in Abb. 1 die Aktiväten und Sprachäußerungen einer Fokusphase im zeitlichen Verlauf dargestellt: Innerhalb von 60 Sekunden bespricht das Paar 11 (!) verschiedene Themen, spricht und editiert dabei parallel und nahezu ununterbrochen. Demgegenüber stehen Paare, deren Paarprozess einen **Zusammenbruch** erleidet, weil sie nicht mehr inhaltlich auf die Äußerungen ihres Partners Bezug nehmen, oder sogar in eine verlegene Starre verfallen (etwa bei A: *"Ah, der erwartet eine Zahl, kriegt aber ein Objekt!"* – B: *[bewegt stumm Mauscursor für 30 Sekunden ziellos über Bildschirm]*).

Maßgeblich für die Flüssigkeit eines Paarprogrammierprozesses ist die Leichtigkeit mit der die Partner gegenseitig die Intentionen ihrer Handlungen und Äußerungen verstehen können. Für diese **Paarhaftigkeit** (*Togetherness*) habe ich fünf Einflussfaktoren identifiziert:

1. Ein **gemeinsames Verständnis des Softwaresystems** erlaubt effiziente Kommunikation (z.B. durch kurze, aber verständliche Bezeichner – wie etwa *"Factory"* statt *'FeatureLayerAttributeTableCellRendererFactory'*); fehlt es, sind mehr Erklärungen nötig oder es kommt zu aufzuklärenden Missverständnissen.
2. Ein **gemeinsames Verständnis von Softwareentwicklung**, z.B. über gängige Architekturen, Entwurfsmuster und Lösungsansätze, erlaubt als implizites Wissen ebenfalls eine effiziente Kommunikation; andernfalls fällt es schwer Vorschläge des Partners zu bewerten (*"Ähm, okay? Mach mal weiter. Für mich ist das oberste Wissenskante!"*).

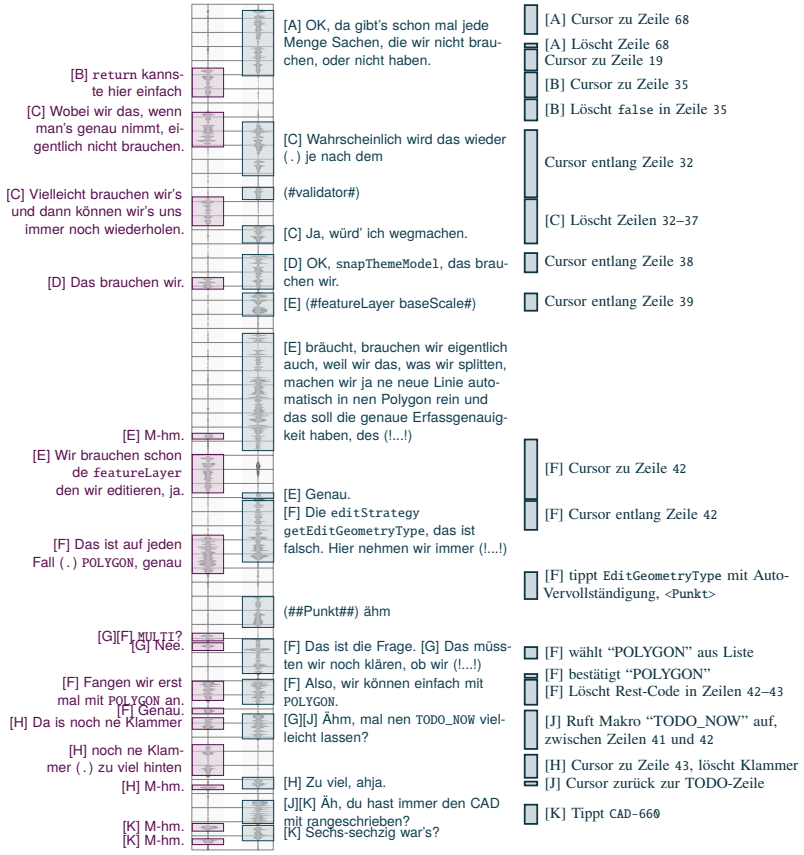


Abb. 1: Eine Fokusphase (60 Sekunden, Zeit von oben nach unten), in der das Paar elf Themen [A]–[K] bespricht (Spalten 1 und 2) und ein Partner direkt Code-Änderungen vornimmt (Spalte 3).

3. Ein **gemeinsamer Plan**, z.B. in der Sitzung getroffene strategische Entscheidungen, bietet den Hintergrund um taktische Vorschläge schnell zu verstehen und zu ergänzen (A: *"[öffnet Code] Okay, viel davon kann weg."* – B: *"Mach hier unten einfach 'return'."*); andernfalls kommt es zu Missverständnissen (A: *"Können wir das debuggen?"* – B: *"Haben wir doch gerade!"*).

4. **Gewahrsein der Arbeitsumgebung** ist z.B. bei verteilter Paarprogrammierung durch die räumliche Trennung reduziert, kann aber auch durch zu kleine Schrift ein Problem im lokalen Fall sein (*"Wo bist du gerade? In welcher Klasse?"*).

5. Eine **Sprachbarriere** kann durch Fremdsprachen bestehen (*"An 'offset' is a duration?"*) oder durch idiosynkratische Ausdrucksweisen (*"Wart mal kurz."* für *'Ich nehme mir jetzt die Tastatur und Maus.'*).

Gute Paare erkennen Defizite und kompensieren z.B. geringeres Gewahrsein (Faktor 4) durch Erläuterungen von Editieraktivitäten, oder nehmen sich die Zeit ein Idiom zu erklären, das den Partner irritiert hat (Faktor 2). Die Paarhaftigkeit ist keine statische Eigenschaft, sondern charaktersiert den momentanen Zusammenhalt und kann von den Entwickler/inne/n vernachlässigt (*"Mach mal. Ich sag, wenn ich wieder voll drin bin."*) oder durch Reparaturen wiederhergestellt werden (*"Warum hattest du das gerade gemacht?"*).

Die bislang nicht erklärte Varianz der experimentellen Ergebnisse (siehe Abschnitt 2.1) ergibt sich womöglich durch unterschiedlich große Defizite der Paarhaftigkeit der Versuchspaare sowie durch unterschiedliche Kompetenzen mit diesen umzugehen (in Experimentsituation mindestens Faktoren 2 und 3). Klären lässt sich das im Nachhinein nicht. Es bleibt dabei, dass für Forscher Prozesseigenschaften wie *Flüssigkeit* und *Paarhaftigkeit* verborgen bleiben, solange sie Paarprogrammierungssitzungen nur auf summative Metriken wie verstrichene Zeit oder erzeugte Codezeilen reduzieren.

4.2 Wissensbedürfnisse, Wissensbedarfe und eine fundamentale Sitzungsdynamik

Im Kern befasste sich meine Forschung mit der Frage, wie Paarprogrammierer Wissen transferieren und sich gemeinsam neu aneignen. Eine zentrale Beobachtung ist hierbei, dass Wissenstransfer in **jeder Sitzung** stattfindet, sei es als erklärtes *Ziel* der Sitzung, in der ein Partner das Softwaresystem zum ersten Mal sieht, oder als *Nebeneffekt* nachdem das Paar festgestellt hat, dass eine/r eine zu schließende Lücke im System- oder Programmierverständnis hat (Reparatur des Zusammenhalts, siehe Abschnitt 4.1).

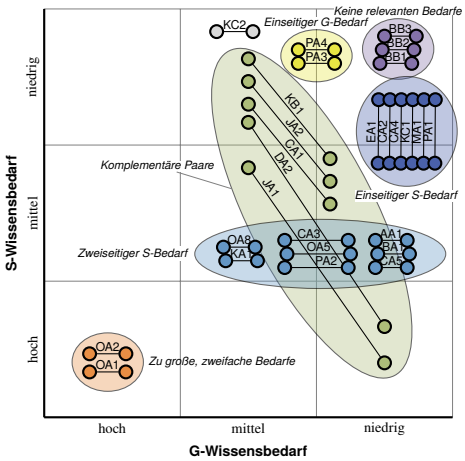


Abb. 2: Startkonstellationen der Wissensbedarfe in den analysierten Paarprogrammierungssitzungen

Ich unterscheide auf der einen Seite ein **Wissensbedürfnis** (*Knowledge Want*), das ein/e Entwickler/in in einer konkreten Situation verspürt und das eine Wissenstransfer-Episode motiviert, in der dann der Partner solange befragt, mit Erklärungen versorgt, oder in die Aneingung neuen Wissens durch Quellcode-Lesen o.ä. einbezogen wird, bis das Bedürfnis gestillt ist.

Der **Wissensbedarf** (*Knowledge Need*) ergibt sich dagegen aus inhaltlichen Anforderungen: Wie gut ist jedes Paarmitglied gewappnet, die Aufgabe erfolgreich zu bearbeiten? Der tatsächliche Wissensbedarf ergibt es sich erst im Laufe einer Sitzung; er kann sich als größer als gedacht herausstellen oder aber durch eine Eingrenzung der Aufgabe schrumpfen.

Basierend auf hunderten analysierten Episoden unterscheide ich zwei Arten von in industriellen Sitzungen relevanten Wissen: System-spezifisches **S-Wissen**, das Wissen über Anforderungen, Architektur und Entwurf, konkreten Technologieeinsatz, Quellcodeeigenschaften und über Defekte umfasst; sowie generisches **G-Wissen** über Softwareentwicklung an sich, das Entwurfsmuster, Programmiersprachen, Werkzeuge und Technologien abdeckt.

Paar-Sitzungen lassen sich nun danach charakterisieren, wie groß die Wissensbedarfe der Partner in den S- und G-Dimensionen *für ihre konkrete Aufgabe* sind, welchen Wissensbedürfnissen sie nachgehen und dadurch ihre Bedarfe erfüllen. Ingesamt habe ich sechs wiederkehrende Startkonstellationen identifiziert (Abb. 2), darunter z.B. *komplementäre Paare*, bei denen ein Partner über mehr aufgabenrelevantes G-Wissen verfügt (etwa bezogen auf Entwurfsmuster und Refactorings) und der andere über mehr S-Wissen (etwa Autorenwissen), oder den *einseitigen S-Bedarf*, der auftritt, wenn ein Partner schon mit der Aufgabe begonnen hat und bereits einen entsprechen S-Wissensvorsprung aufbauen konnte.

In der Zusammenschau der *aller* Sitzungsverläufe habe ich eine gemeinsame, **fundamentale Dynamik** entdeckt, die aus drei Phasen besteht. (1) Schließen der **Primären Wissenlücke**: Zunächst gleichen Paare einen etwaigen S-Wissensunterschied aus (*"Ich erzähl dir mal, was ich schon gemacht hab."* oder *"Kennst du nicht, ne? Ich zeig dir das Ganze erstmal."*). (2) Schließen der **Sekundären Wissenslücke**: Sie erkunden die Software durch Lesen oder mittels Debugger und erarbeiten sich so S-Wissen, das beiden noch fehlt. (3) Erst dann, wenn primäre und sekundäre Wissenslücke geschlossen sind, nutzen manche Paare die **Gelegenheit** zum Transfer von G-Wissen, die sich durch einen etwaigen Wissensunterschied bietet (*"Soll ich dir erklären wie OSGi-Classloading funtioniert?"*).

Es ist nun also eine empirische Beobachtung, dass in industrieller Paarprogrammierung der Transfer von system-unabhängigen G-Wissen erst erfolgt, wenn es keine offenen S-Wissensbedürfnisse mehr gibt, die Entwickler/innen also die für die Aufgabe relevanten Systemteile gut genug verstanden haben. Es zeigt sich eine klare Hackordnung der beiden Wissensarten: Die Aneignung von Systemverständnis war in fast allen Sitzungen eine Hauptsache. Unterschiede im Programmierwissen hingegen behinderten die Paare praktisch nicht, sondern blieben höchstens eine ungenutzte Gelegenheit. Wenn jedoch beiden Partnern zu viel G- und S-Wissen fehlt, wie es in zwei untersuchten Sitzungen der Fall war (Abb. 2, links unten), ist die Aufgabe ist schlicht zu schwer und das Paar macht nur wenig Fortschritt.

4.3 Anwendungsmöglichkeiten

Ich habe drei Praktiken zur Einbindung meiner Erkenntnisse in den Entwicklungsalltag eines Teams ausgearbeitet und in vier Firmen mit Workshops und Interviews evaluiert:

1. **Paare formen**: Teams können das G-S-Diagramm (Abb. 2) verwenden, um für konkrete Aufgaben strukturiert vorhandenes Vorwissen und günstige Paarkonstellationen zu besprechen (ob z.B. zwei Teammitglieder ein *komplementäres Paar* bilden).

2. **Sitzungsziel**: Müssen beide ihren kompletten S-Bedarf erfüllen, oder kann ein Teil der *primären Lücke* bleiben? Gibt es eine Gelegenheit zum Transfer von G-Wissen?
3. **Sitzungsreflektion**: Wurden die Wissensbedarfe erfüllt? Gibt es relevante S-/G-Lücken, die vorher nicht bekannt waren? Sollten diese im Team diskutiert werden?

Alle drei Ideen stießen bei den Praktikern auf breite Resonanz. Während Paare-formen und Sitzungsziel-festlegen wegen praktischer Beschränkungen wie Teamgröße, Entwicklerverfügbarkeit, und starrer Aufgaben wenig Wirkung entfalten konnte, waren die Sitzungsreflektionen ein Erfolg, da sie z.B. helfen, sich an erfolgreichen Wissensaustausch zu erinnern (*"Stimmt daran habe ich gar nicht gedacht, das war sogar richtig cool."*).

5 Zusammenfassung

In der Praxis dreht sich der Großteil einer jeden Paarprogrammierungssitzung um das Verstehen des Softwaresystems: Die mit Abstand meisten Wissenstransfer-Episoden haben System-Wissen zum Thema und Paare befassen sich *zuerst* mit ihren Bedürfnissen nach System-Wissen bevor Gelegenheit zum Austausch über allgemeinere Softwareentwicklungsthemen genutzt werden. Für das flüssige Voranschreiten innerhalb einer Sitzung ist es wichtig, dass das Paar sein gemeinsames Verständnis (u.a.) des Softwaresystems pflegt, da sonst ein Zusammenbruch des Paarprozesses droht.

Jahrelange Forschung unter Laborbedingungen hat Effekte der Paarprogrammierung auf Bearbeitungsdauer und Qualität nur in Tendenz und mit großer Streuung nachweisen, diese aber nicht erklären können. Nicht nur ist die Übertragbarkeit dieser Experimentalergebnissen ohnehin fragwürdig, wo doch der in der Praxis zentrale Rückgriff auf systemspezfisches Vorwissen und gemeinsames Verstehen eines komplexen Systems bei Spielzeugaufgaben komplett fehlt. Auch ist rückblickend nicht verwunderlich, dass quantitativ-orientierten Forschungsarbeiten Erklärungen fehlen, wenn sie nur summative Metriken im Blick hatten, nicht aber den eigentlichen Prozess, wodurch ihnen positive wie negative Ausprägungen entgangen sind, wie etwa Fokusphasen und Zusammenbrüche.

Literatur

[Ar07] Arisholm, E.; Gallis, H.; Dybå, T.; Sjøberg, D. I.: Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise. IEEE Transactions on Software Engineering 33/2, S. 65–86, 2007.

[Be99] Beck, K.: Extreme Programming Explained: Embrace Change. Addison-Wesley, 1999, ISBN: 0201616416.

[BN08] Begel, A.; Nagappan, N.: Pair Programming: What's in it for Me? In: Proc. Second ACM-IEEE Intl. Symp. on Empirical Software Engineering and Measurement. ACM, S. 120–128, 2008.

[BRB08] Bryant, S.; Romero, P.; du Boulay, B.: Pair Programming and the Mysterious Role of the Navigator. Intl. J. of Human-Computer Studies 66/7, S. 519–529, 2008.

[Co98] Coplien, J.: A Generative Development-Process Pattern Language. In (Rising, L., Hrsg.): The Patterns Handbook: Techniques, Strategies, and Applications. Cambridge University Press, S. 243–300, 1998, ISBN: 0-521-64818-1.

[Ha09] Hannay, J. E.; Dybå, T.; Arisholm, E.; Sjøberg, D. I.: The effectiveness of pair programming: A meta-analysis. Information and Software Technology 51/7, S. 1110–1122, 2009.

[Pl15] Plonka, L.; Sharp, H.; van der Linden, J.; Dittrich, Y.: Knowledge transfer in pair programming: An in-depth analysis. Intl. J. of Human-Computer Studies 73/, S. 66–78, 2015.

[Sa13] Salinger, S.: Ein Rahmenwerk für die qualitative Analyse der Paarprogrammierung, Diss., Fachbereich Mathematik und Informatik, Freie Universität Berlin, 2013.

[SC90] Strauss, A.; Corbin, J.: Basics of Qualitative Research. Grounded Theory Procedure and Techniques. Sage Publications, 1990, ISBN: 978-0803932500.

[WH09] Walle, T.; Hannay, J. E.: Personality and the Nature of Collaboration in Pair Programming. In: Proc. 3rd Intl. Symp. on Empirical Software Engineering and Measurement. IEEE, S. 203–213, 2009.

[WK02] Williams, L.; Kessler, R. R.: Pair Programming Illuminated. Addison-Wesley Professional, 2002, ISBN: 978-0-201-74576-4.

[ZHR13] Zarb, M.; Hughes, J.; Richards, J.: Industry-Inspired Guidelines Improve Students' Pair Programming Communication. In: Proc. 18th ACM Conf. on Innovation and Technology in Computer Science Education. S. 135–140, 2013.

[Zi20] Zieris, F.: Qualitative Analysis of Knowledge Transfer in Pair Programming, Diss., Fachbereich Mathematik und Informatik, Freie Universität Berlin, 2020.

[ZP20] Zieris, F.; Prechelt, L.: PP-ind: A Repository of Industrial Pair Programming Session Recordings, 2020, URL: https://arxiv.org/abs/2002.03121.

Franz Zieris wurde 1988 in Berlin geboren. Er studierte Informatik mit Psychologie im Nebenfach an der Freien Universität Berlin. Seit 2007 ist er freiberuflicher Softwareentwickler. Von 2012 bis 2020 war er Wissenschaftlicher Mitarbeiter der AG Software Engineering bei Prof. Dr. Lutz Prechelt, wo er mit qualitativen Forschungsmethoden industrienah zu Prozessthemen wie agiler Softwareentwicklung und Paarprogrammierung forschte und zu Themen der Softwaretechnik lehrte. Von 2012 bis 2018 leitete er das Open-Source-Projekt Saros (https://www.saros-project.org), das IDE-Plugins für verteilte Paarprogrammierung entwickelt.