Generische Verkettung maschineller Ansätze der Bilderkennung durch Wissenstransfer in verteilten Systemen

Am Beispiel der Aufgabengebiete INS und ACTEv der Evaluationskampagne TRECVid

Christian Roschke[1]

Abstract: Technologischer Fortschritt im Bereich multimedialer Sensorik und zugehörigen Methoden zur Datenaufzeichnung, Datenhaltung und -verarbeitung führt im Big Data-Umfeld zu immensen Datenbeständen in Mediatheken und Wissensmanagementsystemen. Dabei Zugrundliegende State of the Art-Verarbeitungsalgorithmen werden oftmals problemorientiert entwickelt, wobei sich aufgrund der enormen Datenmengen nur bedingt zuverlässig Rückschlüsse auf Güte und Anwendbarkeit ziehen lassen. So gestaltet sich auch die intellektuelle Erschließung von großen Korpora schwierig, da die Datenmenge für valide Aussagen nahezu vollumfänglich semi-intellektuell zu prüfen wäre, was spezifisches Fachwissen aus der zugrundeliegenden Datendomäne ebenso voraussetzt wie zugehöriges Verständnis für Datenhandling und Klassifikationsprozesse. Ferner gehen damit gesonderte Anforderungen an Hard- und Software einher, welche in der Regel suboptimal skalieren, da diese zumeist auf Multi-Kern-Rechnern entwickelt und ausgeführt werden, ohne dabei eine notwendige Verteilung vorzusehen. Folglich fehlen Mechanismen, um die Übertragbarkeit der Verfahren auf andere Anwendungsdomänen zu gewährleisten. Die Arbeit [Ro21] nimmt sich diesen Herausforderungen an und fokussiert auf die Konzeptionierung und Entwicklung einer verteilten holistischen Infrastruktur, die die automatisierte Verarbeitung multimedialer Daten im Sinne der Merkmalsextraktion, Datenfusion und Metadatensuche innerhalb eines homogenen Systems ermöglicht.

1 Einführung

Der technologische Fortschritt im Bereich multimedialer Sensorik und zugehörigen Methoden zur Datenaufzeichnung, Datenhaltung und -verarbeitung führt im *Big Data*-Umfeld u.a. zu immensen Datenbeständen in Mediatheken, Medienarchiven und Wissensmanagementsystemen. Zugrundliegende *State of the Art*-Verarbeitungsalgorithmen werden oftmals problemorientiert entwickelt. Aufgrund der enormen Datenmengen lassen sich nur bedingt zuverlässig Rückschlüsse auf Güte und Anwendbarkeit ziehen. So gestaltet sich auch die intellektuelle Erschließung von großen Korpora schwierig, da die Datenmenge für valide Aussagen nahezu vollumfänglich zumindest semi-intellektuell zu prüfen wären, was spezifisches Fachwissen aus der zugrundeliegenden Datendomäne ebenso voraussetzt wie zugehöriges Verständnis für Datenhandling und Klassifikationsprozesse in der Informations- und Kommunikationstechnik. Ferner gehen damit gesonderte Anforderungen an Hard- und Software einher, welche sich in der Regel schlecht skalieren, da die-

[1] christian.roschke@hs-mittweida.de

se zumeist auf Multi-Kern-Rechnern entwickelt und ausgeführt werden ohne dabei eine notwendige Verteilung vorzusehen. Demzufolge fehlen auch Mechanismen, um die Übertragbarkeit der Verfahren auf andere Anwendungsdomänen mit geringen Aufwänden zu gewährleisten.

Der Fokus der Arbeit liegt in der Konzeptionierung und Entwicklung einer verteilten holistischen Infrastruktur, die die automatisierte Verarbeitung multimedialer Daten im Sinne der Merkmalsextraktion, Metadatenanreicherung, Metadatenfusion und Metadatensuche mit Hilfe von modularen Komponenten innerhalb eines homogenen aber zugleich verteilten Systems ermöglicht. Dabei sind Ansätze aus den Domänen des *Maschinellen Lernens*, der *Verteilten Systeme*, des *Datenmanagements* und der *Virtualisierung* zielführend miteinander zu verknüpfen, um auf große Datenmengen angewendet, evaluiert und optimiert werden zu können.

2 Herausforderungen zur Optimierung des Wissenstransfers zwischen SMAs

Der Prozess des *Information Retrieval* besteht nach *Kürsten* [Kü12] und *Ritter* [Ri14] aus mehreren Phasen. Nach dem Upload von multimedialen Daten (Text, Bild oder Video) in ein Datenrepository folgt die Annotation ausgewählter Datensamples, um eine Extraktion von Metadaten mit Hilfe maschineller Lernverfahren und automatisierter Methoden zu realisieren, deren Leistungsfähigkeit sich mittels komponentenbasierter Evaluation und Optimierung steigern lässt.

Beide Autoren fokussieren in ihren Arbeiten auf die Optimierung von *Single Model Architekturen* (SMA) in jeweiligen Problemdomänen. Unter dem Begriff SMA werden in diesem Zusammenhang Architekturen verstanden, die hauptsächlich auf einem Modell basieren, dabei zwar multiple Ansätze des maschinellen Lernens anzuwenden vermögen, aber explizit für ein Problemfeld kreiert wurden und zumeist auf einem spezifischen Einzelrechner-System funktionieren. Der dabei untersuchte Wissenstransfer erfolgt in einer kontrollierten und fest definierten Umgebung innerhalb eines Modells. Das zu transferierende Wissen ist dabei als Sammlung von Metainformationen sowie Hyperparametern zu verstehen, wobei der Wissenstransfer ausschließlich innerhalb einer SMA erfolgt bzw. genutzt wird, um mithilfe von Ensemble-Ansätzen die Güte der Detektionsergebnisse zu verbessern.

Nach *Ritter* [Ri14, §§4,5] funktionieren diese Ansätze bei homogeneren Problemstellungen wie der strukturellen Videoanalyse und dem dortigen automatisierten Auffinden von Kameraeinstellungen und Übergängen (*Shot Detection*) oder bei der Lokalisation von Gesichtern oder Personen in Einzelbildern oder Videos. Im Gegensatz zu diesen homogeneren Problemfeldern umfassen die TRECVid Aufgabengebiete *Instance Search* (INS) und *Activities in Extended Video* (ACTEv) wesentlich komplexere, teilweise aus mehreren einfacheren Problemfeldern zusammengefasste Aufgabenstellungen. INS stellt die Suche nach unscharf vorgegebenen Objektklassen dar, wobei neben der Detektion und Erkennung auch die Identifizierung bestimmter Personen an definierten Orten im Fokus steht.

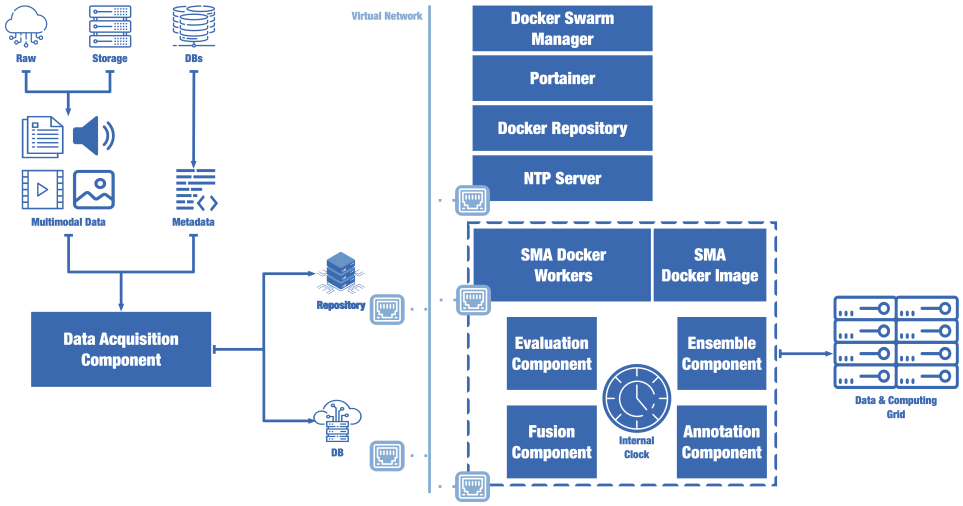

Abb. 1: Vereinfachte Darstellung der Gesamtarchitektur von EMSML

Dies bedingt eine Kombination von Objekterkennung, Ortserkennung und Gesichtserkennung. Die Suchprozesse sind darüber hinaus ad-hoc durchzuführen und beliebige Topics ohne anfragespezifisches Training zu verarbeiten. Bei ACTEv hingegen sind Entitäten und deren Aktivitäten innerhalb eines Datensatzes zu identifizieren. Dabei wird auf die Interaktion zwischen detektierten Objektklassen fokussiert und es stehen wenig *Ground Truth*-Daten zur Verfügung. Die Komplexität ist darüber hinaus dadurch gegeben, dass die Kombinationen von Objekterkennung, Tracking und Aktivitätserkennung essentiell sind, um die gestellten Suchanfragen behandeln zu können. Überdies werden die Datensätze vier bis acht Wochen und die Suchanfragen zwei Wochen vor der finalen Abgabe bereitgestellt, was eine zeitlich eingeschränkte Bearbeitungszeit bedingt.

Zur Bearbeitung derart komplexer und heterogener Problemfelder in adäquater Rechenzeit werden skalierbare und verteilte Ansätze benötigt, die in der Lage sind, multiple SMAs verteilt über diverse Rechnersysteme anzuwenden und einen Wissenstransfer zwischen diesen zu realisieren. Diesbezüglich sind Aspekte des *Datenhandling & Datenmanagement*, der *Steuerung & Virtualisierung*, der *Modellierung & Erstellung von Domänenwissen* sowie der *Annotation, Evaluation und Optimierung* zu beachten und in einer generischen sowie holistischen Infrastruktur abzubilden. Eine dafür geeignete Architektur stellt das in der Arbeit entwickelte und in Abbildung 1 vereinfacht dargestellte *Evaluations- und Managementsystem für Maschinelles Lernen* (EMSML) dar.

Im Rahmen von *Datenhandling und Datenmanagement* ist das Gesamtsystem mittels der *Data Acquisition Component* in der Lage, multimodale Daten sowie Metadaten aus unterschiedlichen Quellen zu importieren und für die persistente Speicherung aufzubereiten. Im Sinne der Vereinheitlichung heterogener Daten findet diesbezüglich eine Umbennung und Transformation in homogene Formate statt. Die Speicherung erfolgt letztendlich auf Dateiebene in Repositories und auf Metadatenebene in Datenbanken. Dabei kann das Wissen,

aus den in den jeweiligen Problemdomänen ursprünglich zur Verfügung gestellten Daten, gesammelt und durch zentrale sowie ausfallsichere Zugriffssysteme problemübergreifend transferiert werden.

Die *Steuerung und Virtualisierung* erfolgt mittels *virtualisiertem Netzwerk*, das den Zugriff auf die Infrastruktur ermöglicht. SMAs tauschen sich darüber aus und können zur Verarbeitung bereits gesammeltes Wissen verwenden. Über eine *Docker Registry* ist das System darüber hinaus in der Lage, Zustandsinformationen einzelner *Single Model Architekturen* in Form von *Docker Images* bereitzustellen und zu verteilen. Die einzelnen SMAs selbst werden als *Docker Container* virtualisiert, wobei eine Verteilbarkeit der Businesslogik auf diverse Rechenknoten stattfinden kann. Die Verteilung dieser Logik übernimmt der *Docker Swarm Manager*, der jeden *Docker Container* als zentrale *Broker-Instanz* verwaltet sowie den Wissenstransfer auf technischer Ebene unter Bereitstellung notwendiger Ressourcen ermöglicht. Mittels *Portainer* lässt sich die Wissensbasis darüber hinaus intellektuell verwalten und zusätzlich mit semantischen Informationen, Prozessabläufen etc. anreichern.

Die Modellierung & Erstellung von Domänenwissen erfolgt über den Einsatz von *Apache Spark*, wobei *DataFrames* zur Verarbeitung bereitgestellt werden und eine Abfrage über *Spark SQL* realisiert wird. Dabei lässt sich das übermittelte Wissen in Form von Metadaten während der Übertragung beliebig modellieren, filtern und problemspezifisch anpassen. Die Übermittlung selbst erfolgt in diesem Kontext batchweise oder in Echtzeit und realisiert somit eine Verarbeitung im *Big Data*-Umfeld. Diesbezüglich können parallele Anfragen mehrerer *Docker-Container* beliebig kombiniert oder ausgetauscht werden, wobei die Verwaltung über den *Swarm Manager* unter Einbezug einer mittels NTP synchronisierten Zeit koordiniert wird.

Die *Annotations*-Komponente innerhalb von EMSML stellt eine webbasierte Annotationsplattform für Audio- und Videoanalysealgorithmen zur Verfügung, wobei die Güte von Prozessketten und SMAs zusätzlich zu automatisch erzeugten Metriken intellektuell eingeschätzt werden kann. Darüber hinaus lässt sich damit zusätzliches semantisches Wissen hinzufügen, wobei Annotationen von Benutzern zeitgenau erfassbar sind. Die automatisierte *Evaluation* wird über Vergleichssysteme innerhalb von EMSML realisiert. Diesbezüglich findet eine Aufgliederung komplexer Probleme in weniger komplexe Teilprobleme statt. Dabei sind einzelne SMAs anzuwenden und mittels klassischer Metriken im Kontext der Detektionsgüte einschätzbar. Überdies ist dabei der Einsatz von Kreuzvalidierung zur *Optimierung* der Hyperparametereinstellungen innerhalb einzelner oder mehrerer SMAs realisierbar, was eine Identifizierung des besten Ansatzes für jeweils ein Teilproblem und durch geeignete Kombination eine möglichst optimale Lösung des Anwendungsproblems zulässt.

3 Wissenschaftliche Ergebnisse

Diese Arbeit ordnet Methoden und Strategien des State of the Art im Bereich des Maschinelles Lernen in der Mediendomäne unter Inklusion der klassischen Lernmethoden zzgl. Transfer-Lernen, Ensemble-Lernen und zugehöriger Leistungemaße ein und betrachtet aktuelle Technologien und Frameworks zur Detektion von Mustern in den Domänen zur Gesichts-, Aktivitäts- und Ortserkennung. Zudem werden diese ins Verhältnis zum aufkommenden Datenmanagement gesetzt und mit den Erfordernissen und Architekturmodellen verteiler Systeme und Virtualisierungslösungen verbunden.

Um den Anforderungen verteilter Systeme gerecht zu werden, wird das von Ritter [Ri14] abgeleitete Schema zur generischen Mustererkennung dabei in einem ersten Schritt um drei Strategien zur Datenfusion auf Pixel-Level, Feature-Level und Decision-Level geeignet erweitert. Zur Demonstration der Allgemeingültigkeit erfolgt eine Eingruppierung der einzelnen Prozessschritte von neuronalen Faltungsnetzen (CNN). Aus der intensiven Betrachtung der Fachliteratur werden insgesamt zehn Hauptkriterien abgeleitet, welche notwendig sind, um an multimedialen wissenschaftlichen Evaluationskampagnen unter Einsatz von verteilter Systemen in verschiedenen Problemdomänen teilnehmen zu können.

Der holistische Systementwurf der in der Arbeit erstellten Gesamtarchitektur des *Evaluations- und Managementsystems für Maschinelles Lernen* (EMSML) besteht aus elf Hauptkomponenten und 30 Anforderungen, welche über technische und qualitative Anforderungen hinaus Aspekte und Konzepte zum Datenhandling, zur Ausgestaltung der *Single Model Architekturen* und zum Wissenstransfer enthalten. Ein domänenübergreifender Workflow identifiziert sieben Basisschritte, um kompatible und verteilbare SMAs zu erstellen, die problemspezifisch entworfen wurden, deren (Teil-)Ergebnisse aber problemübergreifend zur Lösung komplexerer Problemklassen beitragen können.

Eine tiefgreifenden Analyse der Entwicklung der Problemdomäne *Instance Search* und des *Related Work* offenbart über die Historie der wissenschaftlichen Evaluationskampagne einen steigenden Schwierigkeitsgrad, der sich in zunehmend fordernden Kategorien und höherem Automatisierungsgrad widerspiegelt. Konkret sei hierbei darauf verwiesen, dass zu Beginn des Wettbewerbs eher einzelne Objekte aufzufinden waren, für die vage Umschreibungen wie *cuved plastic bottle of ketchup* und vier bis fünf Beispielbilder existierten. Im Verlauf der Zeit und der Entwicklung der zugrundeliegenden Technologien in den vergangenen für diese Arbeit relevanten Evaluationsperioden trat hingegen die Suche nach eher selten auftretenden Nebendarstellern mit unterschiedlichen Aktivitäten oder unterschiedlichen Orten in den Fokus. Im Vergleich zur traditionellen Objekterkennung, begründet sich die Schwierigkeit des Problems in vier wesentlichen Punkten: (a) Eine äußerst geringe Anzahl an Trainingsbeispielen. (b) Das zu Trainingszwecken verwendbare Videomaterial ist zwei Stunden lang und für einen Großteil der Suchanfragen nicht repräsentativ, da dies willkürlich ausgewählt wurde (Video 0 als erstes Video der gesamten Kollektion) und in der Regel weder das gesuchte Objekt noch dessen Ort oder Handlung enthält. (c) Der Testkorpus ist mit 462 Stunden Länge überproportional groß und darf in keinem Fall für die Entwicklung von Algorithmen und deren Evaluation und Optimierung eingesetzt werden.

Außerhalb des Evaluations-Turnus wurden im Rahmen der Arbeit in den Teilbereichen *Objektdetektion*, *Personenklassifikation*, *Ortsdetektion* sowie *Aktivitätserkennung* domänenspezifische Experimente durchgeführt. Innerhalb der Domäne der *Objektdetektion* ließen sich für die *Personenerkennung* aus dem BBC *EastEnders*-Datensatz 1.000 Schlüsselbilder extrahieren und intellektuell annotieren. Der dabei entstehende Datensatz besteht aus 250 Bildern mit einer Person, 250 Bildern mit zwei Personen und 500 Bildern ohne Personen. In vier Verarbeitungsabläufen konnte mit den Softwareframeworks *Detectron* und *Yolo9000* untersucht werden, ob die Personenerkennung innerhalb von EMSML unabhängig von der verwendeten Technologie durchführbar ist und durch die Verwendung von Verteilungsstrategien in diesem Bereich eine Verbesserung der Performanz erzielt werden kann. Dabei wurden im Rahmen der Verteilung rechnerspezifische und prozessspezifische Performanzwerte sowie die Detektionsgüte erfasst. Diesbezüglich ist erkennbar, dass *Detectron* im Kontext der *Personendetektion* am bereitgestellten Datensatz bessere Ergebnisse als *Yolo9000* erzielen kann und eine höhere *Accuracy* aufweist. Darüber hinaus lassen sich die beiden heterogenen Technologien mittels EMSML einsetzen und durch die Kombination der Ergebnisse Verbesserungen erzielen. Die Fusion auf *Ergebnis-Level* ermöglicht diesbezüglich eine Steigerung der Accuracy von 81 % bei *Detectron* und 72 % für *Yolo9000* auf kumulativ 94 %. Für die Verarbeitung aller 1.00 Bilder wurden mit *Detectron* ca. 100 Sekunden und mit *Yolo9000* ca. 200 Sekunden benötigt. Bei der Verteilung der Verarbeitung auf die beiden Knoten verhielten sich die Knoteneigenen Parameter ähnlich zu den lokalen Versuchen; allerdings wurde die Zeit durch die Parallelverarbeitung signifikant verringert, was auf eine hervorragende Skalierungsfähigkeit von EMSML schließen lässt.

Im Rahmen der *Personenklassifikation* konnte mittels *Decision-Level Fusion* eine Verbesserung der Accuracy von 88 % bei *FaceNet* sowie 88 % bei *OpenFace* auf 95 % bei einer Verarbeitungszeit für 1.000 Bilder von 30 Sekunden für *FaceNet* und 49 Sekunden bei *OpenFace* erreicht werden. Die *Ortsdetektion* basierte auf einem ähnlichen Ansatz, wobei 1.000 Schlüsselbilder von Orten genutzt wurden. Dabei ließen sich mittels Ergebnisfusion der Frameworks *Places365* und *TuriCreate* eine Steigerung der Accuracy von 77 % und 72 % auf 83 % erzielen, wobei die Verarbeitungszeiten 115 respektive 200 Sekunden betrugen.

Experimente im Umfeld der *Aktivitätserkennung* forcierten die Verwendung von Ähnlichkeitsmodellen und einem Bildklassifizierer die auf Einzelbilder von einzelnen Aktivitäten trainiert wurden und sich ebenfalls auf einen Testdatensatz von 1.000 intellektuell zusammengestellten Images anwenden ließen. Dabei konnte eine Accuracy von 69 % und 75 % erreicht werden, wobei Fusionierungsansätze eine Verbesserung dieser auf 77 % ermöglichten. Dabei konnten Verarbeitungszeiten von 215 Sekunden (Ähnlichkeitsmodell) und 60 Sekunden (Bildklassifizierer) erreicht werden.

Die domänenübergreifenden Experimente fokussierten auf die Bearbeitung des Gesamtproblems, wobei im Jahr 2018 Personen an einem Ort und im Folgejahr Personen bei der Ausführung spezifischer Aktivitäten zu identifizieren waren. Dabei sollten jeweils 30 Topics im BBC *EastEnders*-Datensatz gesucht und die Ergebnisse gesammelt werden. Die Übermittlung an das NIST fand anschließend in Form einer XML statt. Diesbezüglich

ermöglichte das EMSML in mehreren Durchläufen im Jahr 2018 3.478 und 3.550 der 11.717 und im Jahr 2019 zwischen 874 bis 939 von 6.592 aller relevanter Schlüsselbilder zu identifizieren. Dabei konnte 2018 eine MAP von maximal 11 % und 2019 eine MAP von maximal 0,82 % in automatisierten Verarbeitungen erzielt werden. Der Einsatz intellektueller Nachbearbeitung ermöglichte 2018 darüber hinaus eine Steigerung der MAP auf 25,2 %. Gesamtheitlich betrachtet ließen sich über beide Jahre 39.186.091 Einzelbilder der 464 Stunden Videomaterialien verarbeiten und über 250 Millionen Metadaten sammeln. In diesem Kontext umfasste die Verarbeitungszeit für die komplexe Problemstellung durchschnittlich zwischen 18 und 24 Tagen sowie die Suchzeit über alle Topics 0,14 *ms* pro Bild.

Ähnlich detailliert wird die Entwicklung in der noch jungen Domäne der Aktivitätsklassifizierung aus multiplen Kameraaufnahmen (ACTEv) innerhalb der TRECVid Evaluationskampagne aufgearbeitet. Erwähnenswert ist, dass die Definition der Ergebnisklassen und der Evaluationsmetrik im Verlauf der ersten beiden Wettbewerbsjahre innerhalb der Evaluationsperioden mehrfach auf die Resultate ausgewählter Teams angepasst wurden, was einen nachhaltigen Systemaufbau bzw. die Erzielung valider vergleichbarer Ergebnisse deutlich erschwerte, weshalb die Arbeit naturgemäß auf die offiziellen Ergebnisse Bezug nimmt und eigene Korrekturen explizit ausweist.

Die Zielstellung der domänenbasierten Experimente war bei ActEV ähnlich ausgerichtet, wie bei INS. Dabei ließen sich die Gebiete *Personen- und Fahrzeugerkennung*, *Tracking* sowie *Aktivitätserkennung* individuell untersuchen.

Im Bereich der *Personen- und Fahrzeugerkennung* konnten aus dem Datensatz VIRAT-V1 1.000 Schlüsselbilder extrahiert und annotiert werden. Die Frameworks *Detectron* und *Yolo9000* beinhalten bereits vortrainierte Modelle, die eine Erkennung von Personen und Fahrzeugen gestattet. Im Rahmen der Experimente stellte sich heraus, dass für den bereitgestellten Datensatz *Detectron* in der Personenerkennung bessere Ergebnisse erzielen konnte, wobei *Yolo9000* hingegen bei der Fahrzeugerkennung bessere Resultate erzielte. Darüber hinaus ermöglichte die *Decision-Level Fusion* eine Genauigkeit von jeweils 90 % auf 93 % zu steigern und die Defizite der beiden Frameworks auszugleichen. Die Verarbeitungszeit aller 1.000 Bilder wurde mit ca. 300 Sekunden für *Detectron* und 500 Sekunden für *Yolo9000* bestimmt. Die im Vergleich zu *Instance Search* längere Bearbeitungszeit lässt sich durch die wesentlich höhere Auflösung von 1920×1080 Pixel des VIRAT-Datensatzes begründen.

Die Experimente im Bereich *Tracking* basierten auf zehn 100 Sekunden langen Videoclips aus dem VIRAT-Datensatz. Die enthaltenen Objekte wurden intellektuell annotiert und die Begrenzungsrahmen aus dieser Annotation mit den (x,y)-Koordinaten der oberen linken Ecke sowie Breite und Höhe gespeichert. Die resultierenden Videoclips wurden anschließend in Einzelbilder zerlegt, wobei sich bewegende Objekte über mehrere Einzelbilder repräsentiert wurden. Dabei konnte ermittelt werden, dass die Gesamtergebnisse im Sinne der Precision 65 %, Recall 75 % und Accuracy 83 % teilweise mittelmäßige bis gute Werte aufweisen, was darauf schließen lässt, dass sich der eigens entwickelte Tracking-Algorithmus für die gestellte Aufgabenstellung eignet. Die intellektuelle Sichtung der Ergebnisse zeigte, dass der Tracker Schwächen bei Überdeckungen und sich verändernder

Geschwindigkeiten von Objekten aufweist. Die Verarbeitung aller 30.000 Bilder dauerte im Durchschnitt 800 Sekunden.

Für die Domäne der *Aktivitätserkennung* wurden aus dem VIRAT-Datensatz weitere 200 Videos mit jeweils 10 Sekunden Länge extrahiert. Dabei konnten die drei Verfahren zur (1) *Posenerkennung via LSTM*, (2) *Extraktion von Bewegungsvektoren* sowie (3 zur) *Größenveränderung der Bounding Boxes* getestet werden, um relevante Aktivitäten automatisiert zu klassifizieren. Der Ansatz der Bewegungsvektoren zeigte in diesem Kontext die besten Ergebnisse im Rahmen von Aktivitäten die auf Bewegungsänderungen fokussierten (bspw. Linksabbiegen, Rechtsabbiegen, Kehrtwende). Der Ansatz der Beobachtung von Größenänderungen von Bounding Boxes gestattet grundsätzlich die Erkennung zur Öffnung eines Kofferraums, wobei dieser Ansatz beim Nachziehen eines Objekts schlechtere Ergebnisse aufwies. Durch Verwendung eines mittels synthetischer Daten trainierten LSTMs ließ sich darüber hinaus eine Aktivitätserkennung von Personen-bezogenen Aktivitäten realisieren. Über alle Aktivitätsklassen hinweg war es damit möglich, eine durchschnittliche Güte von 53 % Precision, 73 % Recall und 71 % Accuracy zu erreichen, wobei die Defizite der einzelnen Ansätze kompensierbar waren. Die Verarbeitungszeiten variieren dabei stark und sind abhängig vom gewählten Ansatz. Der LSTM-Ansatz benötigte erwartungsgemäß doppelt so viel Zeit wie die Ansätze (2) und (3), die mit durchschnittlich $0,023$ Sekunden in etwa gleich lang für die Verarbeitung eines Bildes benötigten.

Basierend auf den Erkenntnissen der domänenspezifischen Experimente ließ sich innerhalb der domänenübergreifenden Experimente das Gesamtproblem bearbeiten. In diesem Kontext waren 2018 das Auftreten von 12 Aktivitäten inklusive Objektpositionen und 2019 ausschließlich 18 Aktivitäten innerhalb der VIRAT-Datensätze zu identifizieren, wobei sich ermittelte Ergebnisse in Form einer JSON an das NIST übermitteln ließen. Dabei war es möglich 3.151.211, bzw. 3.550.215 Personen sowie zwischen 27.030.834 und 27.525.075 Fahrzeuge, innerhalb zugrundeliegender 2.094.413 Einzelbilder der VIRAT-Datensätze in einer Auflösung von 1920x1080px, zu ermitteln. Darüber hinaus konnte gezeigt werden, dass die Objektdetektion bei einer Objekt-P_{miss}-Rate von $0,32\,\%$ gut funktioniert und damit Ergebnisse im Rahmen der Aktivitätserkennung 2018 mit einer P_{miss}-Rate von durchschnittlich $0,94\,\%$ erzielbar waren. Überdies ermöglichte die Optimierung der Ansätze innerhalb des EMSML 2019 das Finden der bisher unerkannten Aktivitäten *Closing* und *Opening*. Weiterhin sank die P_{miss}-Rate der 12 Klassen von 2018 auf $0,91\,\%$, wobei 2019 18 Aktivitäten mit einer P_{miss}-Rate von $0,92\,\%$ ermittelt werden konnten. Über den kompletten Bearbeitungszeitraum beider Jahre ließen sich 90 Millionen Metadaten sammeln. Diesbezüglich umfasste die Verarbeitungszeit zwischen drei und neun Tagen, während die Suchzeit über alle Topics mit durchschnittlich $0,18$ *ms* pro Bild erzielt werden konnte.

Die Arbeit zeigt, dass die Adaption verschiedener moderner Mustererkennungsframeworks als *Single Model Architektur* durch EMSML prinzipiell ermöglicht wird und die Systemleistung durch Einsatz multipler SMAs sowie geeigneter Fusionsmethoden sowohl qualitativ als auch bzgl. der Geschwindigkeit verbessert werden kann. Für den Einblick in weitere relevante Ergebnisse sei auf die einzelnen Kapitel der Arbeit verwiesen.

4 Kapitelübersicht

Die Arbeit gliedert sich in sechs nachfolgend beschriebene Kapitel.

Kapitel 2 umfasst Methoden und Strategien der Bereiche *Maschines Lernen in der Mediendomäne*, *Verteilte Systeme* sowie *Datenmanagement & Virtualisierung*, wobei Definitionen und Begriffserklärungen im Kontext der Arbeit nähere Erläuterung finden. Darauf aufbauend werden aktuelle Technologien und Frameworks zur Detektion von Mustern analysiert, wobei eine Leistungsbewertung erfolgt und ein Kriterienkatalog abgeleitet wird.

In Kapitel 3 ermöglichen Vorüberlegungen die Analyse des zuvor erstellten Kriterienkatalogs hinsichtlich sich daraus ergebender Anforderungen an eine holistische Infrastruktur zur generischen Anwendung von Bilderkennungsalgorithmen im Kontext multipler Problemstellungen in der Domäne der Videoanalyse. Überdies lassen sich Konzepte ableiten, die die Erstellung einer Gesamtarchitektur gestatten. Das so entstehende holistische Framework EMSML bildet die Grundlage für die Umsetzung, Evaluation und Optimierung heterogener Technologien und Ansätze, wobei diesbezüglich ein domänenübergreifender Workflow zur Integration von Single Model Architekturen in die Infrastruktur beschrieben wird. Die so entstehenden Implementationen sind erfahrungsgemäß nur im Zusammenhang mit dem kontextspezifischen Anwendungsfällen konkret untersuchbar, wobei sich wissenschaftliche Evaluationskampagnen, wie beispielsweise TRECVid, aufgrund gegebener Evaluationsmetriken und Datensätze besonders zu eignen scheinen.

Kapitel 4 beinhaltet den ersten Anwendungsfall, das TRECVid-Aufgabengebiets *Instance Search*. In diesem Kontext sind Personen an bestimmten Orten oder bei definierten Aktivitäten zu identifizieren. Dabei wird in einer fachlichen Einordnung die historische Entwicklung des Aufgabengebiets und dessen Abgrenzung zur klassischen Objektdetektion dargestellt und die Spezifika der verwendeten Datensätze und Abgfragetypen analysiert. Weiterhin wird die komplexe Aufgabenstellung in Teilaspekte aufgegliedert und konzeptionelle Lösungsansätze vorgestellt. Anschließend lassen sich die Teillösungen zu einer übergreifenden Lösung kombinieren und aufzeigen, wie sich die gewählten Ansätze mittels der in Kapitel entwickelten Infrastruktur umsetzen lassen. Dabei steht insbesondere die Identifikation und anschließender Anwendung von Optimierungsparametern im Vordergrund, wobei eine Evaluation im Sinne der Funktionalität, Güte und Performanz erfolgt.

Das letzte Anwendungsszenario wird in Kapitel 5 beschrieben und umfasst das TRECVid-Aufgabengebiet *Activities in Extended Video*, wobei Personen oder Fahrzeuge bei der Ausführung definierter Aktivitäten wiederzufinden sind. Durch eine fachliche Einordnung und die Betrachtung der historischen Entwicklung kann dabei eine Abgrenzung zu *Instance Search* aufgezeigt werden. Analog zu Kapitel 4 werden in diesem Kontext zusätzlich die *Related Work* untersucht und der Forschungsfokus sowie die Zielsetzung der eigenen Arbeiten abgeleitet. Darüber hinaus wird aufgezeigt, welche Ansätze sich zur Lösung der komplexen Problemstellung eignen und welche Anpassungen bestehender Technologien zur Bearbeitung des Aufgabenkomplexes notwendig erscheinen. Diesbezüglich lässt sich ebenso eine Evaluation im Sinne der Funktionalität, Güte und Performanz durchführen.

Neben Kapitel 1, das den Anwendungskontext der Arbeit skizziert, ordnet das letzte Kapitel 6 die erzielten Ergebnisse ein und gibt einen Ausblick über zukünftige Verbesserungsmöglichkeiten.

Literaturverzeichnis

[Kü12] Kürsten, J.: A Generic Approach to Component-Level Evaluation in Information Retrieval. Dissertation, Technische Universität Chemnitz, Chemnitz, 2012. ISBN 978-3-941003-68-2.

[Ri14] Ritter, M.: Optimierung von Algorithmen zur Videoanalyse – Ein Analyseframework für die Anforderungen lokaler Fernsehsender. Dissertation, Technische Universität Chemnitz, Chemnitz, 2014. ISBN 978-3-944640-09-9.

[Ro21] Roschke, C.: Generische Verkettung maschineller Ansätze der Bilderkennung durch Wissenstransfer in verteilten Systemen – Am Beispiel der Aufgabengebiete INS und ACTEv der Evaluationskampagne TRECVid. Dissertation, Technische Universität Chemnitz, Chemnitz, 2021. ISBN 978-3-96100-142-2.

Christian Roschke wurde am 19. Mai 1987 in Altdöbern geboren und besuchte von 1999 bis 2006 das Dr.-Albert-Schweitzer Gymnasium in 03226 Vetschau, wo er die Allgemeine Hochschulreife mit der Note 1,8 erlangte. Nach der Schule absolvierte Herr Roschke ein Pflegepraktikum im Evangelischen Krankenhaus Luckau und begann im Sommersemester 2007 mit dem Studium der Medizin an der Justus-Liebig-Universität Gießen. Nach drei Semestern wechselte er die Studienrichtung und nahm ein Bachelor-Studium der *Angewandten Informatik* an der Technischen Universität Chemnitz im Wintersemester 2008 mit dem Schwerpunkt *Eingebettete Systeme* auf, welches er 2012 mit einer 1,7 abschloss. Nach der Erlangung des akademischen Grads *Bachelor of Science* begann Herr Roschke das Master-Studium *Data and Web Engineering* ebenfalls an der Technischen Universität Chemnitz sowie eine Tätigkeit als *Lehrkraft für besondere Aufgaben* (LfBA) an der Hochschule Mittweida. Den *Master of Science* erlangte Herr Roschke im Jahr 2016 mit der Note 1,6. Beruflich führte er seine Tätigkeit als LfBA bis Januar 2017 an der Hochschule Mittweida fort und trat anschließend eine Anstellung als *Akademischer Assistent* im BMBF-Projekt *Stärkung und Erweiterung des akademischen Mittelbaus* (SEM) bis 2020 eben dort an. Während seiner Beschäftigung als Akademischer Assistent begann Herr Roschke 2017 mit der Promotion *Generische Verkettung maschineller Ansätze der Bilderkennung durch Wissenstransfer in verteilten Systemen – Am Beispiel der Aufgabengebiete INS und ACTEv der Evaluationskampagne TRECVid*, die er im September 2021 mit dem Gesamtprädikat *magna cum laudae* verteidigte. 2021 wurde Herr Roschke als Professor für Digitale Transformation & Angewandte Medieninformatik an der Fakultät für Angewandte Computer- und Biowissenschaften der Hochschule Mittweida berufen.