6D Posenschätzung mit gelernten, dichten Korrespondenzvorhersagen[1]

Eric Brachmann[2]

Abstract: Diese Arbeit befasst sich mit der der Schätzung von Position und Orientierung von Objekten oder Szenen aus einzelnen Kamerabildern (Posenschätzung). Es wird ein Verfahren vorgestellt, welches etablierte Lösungsstrategien der Computer Vision mit neuen Verfahren des maschinellen Lernens kombiniert. Zunächst wird das Korrespondenzproblem zwischen Eingabebild und Zielobjekt gelöst, dann wird die gesuchte Pose durch robuste, geometrische Optimierung ermittelt. Nur Teile des Verfahrens werden anhand von Trainingsdaten gelernt, was zu Generalisierung und Interpretierbarkeit des Systems führt. Etablierte Algorithmen, insbesondere der *RANSAC*-Algorithmus aus der robusten Optimierung, werden so erweitert, dass ein Trainieren des Gesamtsystems möglich ist. Mit dem DSAC-Algorithmus (*Differentiable RANSAC*) stellt diese Arbeit Forschern auf dem Gebiet des maschinellen Lernens ein neues, vielseitiges Werkzeug zur Verfügung.

1 Einleitung

In den letzten Jahrhunderten hat die Menschheit einen technischen Fortschritt erlebt, der in vielen Regionen der Welt zu einer Steigerung der Lebensqualität führte. Eine Grundlage für diesen Fortschritt wurde vor ca. 200 Jahren in der industriellen Revolution gelegt. Im Moment beobachten wir erneut einen rasanten Umbruch in der Menschheitsgeschichte, auch digitale Revolution genannt, ausgelöst durch die rapide Entwicklung von Computern. Berechnungen und Simulationen erfolgen in Sekundenbruchteilen, riesige Datenmengen können gespeichert und verarbeitet werden und das Internet verbindet Menschen und Geräte weltweit. Durch aktuelle Entwicklungen in der Forschung zur künstlichen Intelligenz (KI) können Maschinen immer komplexere Aufgaben immer selbstständiger lösen. Schon bald könnten Smart Homes, voll-automatische Fabriken und Warenhäuser, computergestützte Chirurgie oder autonomes Fahren Realität werden. Einige dieser Errungenschaften hätten das Potential, die Lebensqualität und Lebenserwartung weiter zu steigern, etwa durch Reduzierung von Verkehrsunfällen. Gleichzeitig wird unser bisheriges Verständnis von (menschlicher) Intelligenz und der soziale und wirtschaftliche Status von (menschlicher) Arbeit in Frage gestellt.

Soll eine Maschine eine komplexe Aufgabe autonom lösen, muss sie in den meisten Fällen ihre Umgebung wahrnehmen und interpretieren. Die Computer Vision ist ein Gebiet innerhalb der KI-Forschung, welches sich mit dem Verstehen von Bildern beschäftigt, d.h. mit dem Extrahieren von semantischen Informationen aus visuellen Daten. Für uns Menschen ist diese Fähigkeit so allgegenwärtig und selbstverständlich, dass uns die komplexen Prozesse der Bildentstehung, die visuelle Vielfalt und Mehrdeutigkeit unserer

[1] Englischer Titel der Dissertation: “Learning to Predict Dense Correspondences for 6D Pose Estimation”

[2] Universität Heidelberg, eric.brachmann@tu-dresden.de

alltäglichen Umgebung nicht bewusst sind. Das Aussehen von Gegenständen ändert sich enorm, je nach Blickwinkel, Beleuchtungssituation, Reflektionen oder teilweiser Verdeckung. Während man in den Anfangszeiten der Computer Vision versuchte, stabile Muster händisch zu definieren, durch die man trotz all dieser Faktoren Objekte in Bildern erkennen konnte, setzt man heute zunehmend auf das maschinelle Lernen. Aufgrund eines Trainingdatensatzes soll sich ein technisches System die komplexen Zusammenhänge zwischen Objektattributen und deren visueller Erscheinung selbständig erschließen. Neue Methoden des maschinellen Lernens, insbesondere des *Deep Learnings* mittels neuronalen Netzen mit Millionen lernbaren Parametern, haben die Leistungsfähigkeit von technischen Systemen für viele Aufgaben massiv verbessert. Etwa in der Erkennung von Gesichtern oder Verkehrsschildern sind diese Systems dem Menschen unter bestimmten Umständen inzwischen überlegen [St12].

Die Dissertation [Br18a] beschäftigt sich mit dem Schätzen der Position und der Orientierung von Objekten aus einzelnen Bildern. Aufgrund des komplexen Bildentstehungsprozesses ist die Lösung dieser Aufgabe für technische Systeme schwierig, gleichzeitig erfordern Anwendungen wie *Augmented Reality* eine hohe Stabilität und Präzision der Ergebnisse. Reines *Deep Learning*, d.h. die Lösung der gesamten Aufgabe durch ein neuronales Netz, liefert nur enttäuschende Ergebnisse. Die Dissertation stellt ein präzises, skalierbares und vielseitig einsetzbares Verfahren zur Posenschätzung vor, welches Methoden des maschinellen Lernens mit traditionellen Ansätzen der Computer Vision kombiniert.

1.1 Problemdefinition

Die Dissertation beschäftigt sich mit der Posenschätzung von Objektinstanzen. Im Unterschied zu einer Objektklasse bezeichnet eine Instanz ein bestimmtes physisches Objekt, einzigartig in Material und Form. Beispielsweise handelt es sich bei *Auto* um eine Objektklasse und bei *roter VW Golf VII* um eine Objektinstanz. Weiterhin sind Objektinstanzen in dieser Arbeit auf Starrkörperobjekte beschränkt, d.h. ihre Form ändert sich nicht. Im Unterschied dazu stehen artikulierte oder verformbare Objekte wie Laptops.

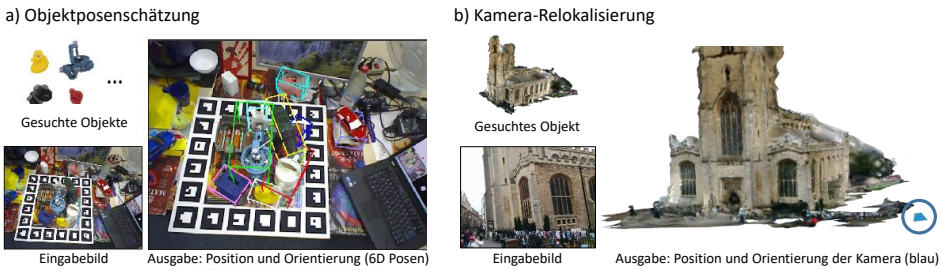


Abb. 1: **Posenschätzung. a)** Das System schätzt Position und Orientierung mehrerer sichtbarer Objekte, dargestellt als farbige 3D-Rahmen. **b)** Das System schätzt Position und Orientierung der Kamera, von der ein Bild einer bekannten Umgebung (eine Kirche in Cambridge, UK) aufgenommen wurde.

Gegeben ist ein Kamerabild I, das eine oder mehrere bekannte Objektinstanzen zeigt. Es handelt sich entweder um ein RGB-Bild einer üblichen Farbkamera, oder ein RGB-D-Bild einer speziellen Tiefenkamera, z.B. einer Kinect-Kamera, was die Posenschätzung erheblich vereinfacht. Für jedes Objekt, welches dem System bekannt ist, liefert es einen Sicherheitswert, ob das jeweilige Objekt zu sehen ist. Weiterhin bestimmt es für jedes Objekt die Pose $\mathbf{h}$ bestehend aus der Position $\mathbf{t}$ relativ zur Kamera und der Orientierung θ des Objekts, siehe Abb. 1 a). Position und Orientierung haben jeweils 3 Freiheitsgrade, d.h. bei der Pose $\mathbf{h}$ handelt es sich um einen 6D-Vektor.

Statt um einen Gegenstand kann es sich bei dem gesuchten Objekt auch um eine ganze Umgebung handeln, etwa eine bestimmte Wohnung oder ein Gebäude, siehe Abb. 1 b). In diesem Fall wird die Pose der Kamera relativ zum Objekt geschätzt, auch Kamera-Relokalisierung genannt. Objekt-Posenschätzung und Kamera-Relokalisierung sind methodisch äquivalent und werden in dieser Arbeit gleichermaßen behandelt.

1.2 Anwendungen

Die Anwendungsmöglichkeiten von Posenschätzung sind vielfältig. Kamera-Relokalisierung kann die Navigation von autonomen Fahrzeugen unterstützen, wenn GPS nicht zuverlässig funktioniert oder die Genauigkeit nicht ausreicht. Für Navigation in geschlossenen Räumen steht GPS außerdem oft nicht zur Verfügung. Soll ein Roboter mit Objekten interagieren, sie etwa in einem automatisierten Warenhaus greifen, muss deren Lage im Raum exakt bestimmt werden. *Augmented Reality*, also die Verschmelzung von realen und virtuellen Inhalten, erfordert die genaue Registrierung von Objekten und Umgebung mit der AR-Anzeige. Gleichermaßen wird in der Wahrnehmungspsychologie die Aufmerksamkeit von Probanden mittels tragbaren Eye-Trackern untersucht. Durch Kamera-Relokalisierung können in diesem Kontext Aufmerksamkeitskarten, etwa von sicherheitskritischen Arbeitsplätzen, erstellt werden. Im Folgenden wird eine Anwendung der Posenschätzung für computergestützte Chirurgie näher erläutert.

Die moderne Medizin ermöglicht besonders schonende Eingriffe durch laparoskopische, also minimal-invasive, Chirurgie, siehe Abb. 2 a). Dabei operiert der Chirurg mit speziellen Instrumenten durch die geschlossene Bauchdecke. Der Chirurg kann sich lediglich durch eine sehr eingeschränkte endoskopische Sicht orientieren, siehe Abb. 2 b), was diese Eingriffe sehr kompliziert macht. Durch das Tracken der 6D-Posen der Operationsinstrumente wäre es möglich, den Chirurgen zu unterstützen, etwa um Distanzen in der Bauchhöhle zu messen, Schnitttiefen zu bestimmen oder Operationsphasen zu erkennen. Wäre es darüber hinaus möglich, die Endoskop-Kamera innerhalb der Bauchhöhle zu lokalisieren, könnten dem Chirurgen Navigationshilfen angeboten werden, etwa über *Augmented Reality*, siehe Abb. 2 c).

Bei dem soeben erläuterten Anwendungsszenario von Posenschätzung in der laparoskopischen Chirurgie handelt es sich um eine Vision. Im Moment existiert kein Verfahren für das Intrumenten-Tracken oder die Kamera-Relokalisierung, das innerhalb des menschlichen Körpers verlässlich funktioniert. Mit dieser Arbeit werden bestehende Verfahren in

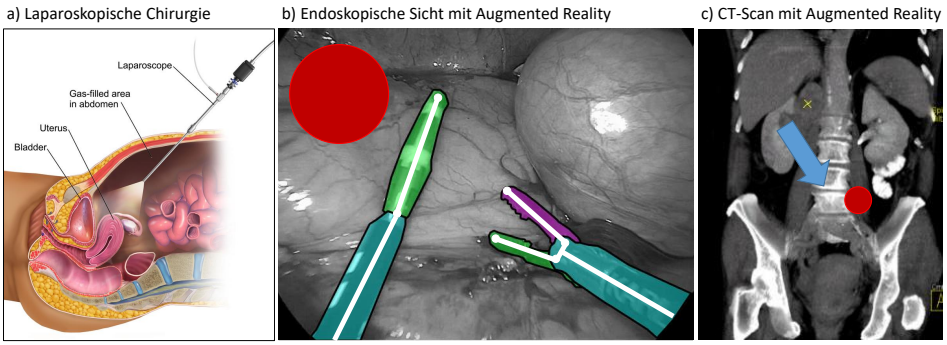
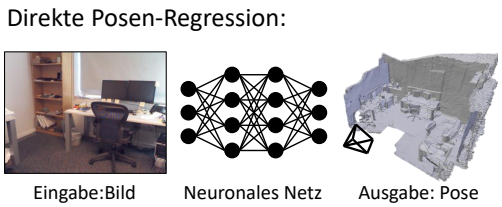


Abb. 2: **Computergestützte Chirurgie (Vision). a)** Schematische Darstellung einer minimalinvasiven Operation im Bauchraum. **b)** Sicht durch das Endoskop. Die Pose der Instrumente wird getrackt. Über *Augmented Reality* wird die Zielposition des Eingriffs als Navigationshilfe eingeblendet (rot). **c)** In einem präoperativen CT-Scan wird die geschätzte Endoskop-Position (blau) und die Zielposition des Eingriffs (rot) gezeigt.

vielen Aspekten verbessert, z.B. hinsichtlich der Robustheit gegenüber Objektverdeckung oder der Genauigkeit der Kamera-Relokalisierung. Andere wichtige Aspekte bleiben offen für zukünftige Forschung, etwa Relokalisierung in deformierbaren Umgebungen (etwa der Bauchhöhle), oder das Tracken von stark reflektiven Objekten (etwa Operationsinstrumenten).

Forschungsbeitrag



Direkte Posen-Regression

Methode (Jahr)	Genauigkeit
PoseNet v.1 [KGC15] (2015)	45cm, 10°
Spatial LSTM [Wa17] (2017)	31cm, 10°
PoseNet v.2 [KC17] (2017)	23cm, 8°
PoseNet v.3 [Br18c] (2018)	22cm, 8°
MapNet+ [Br18c] (2018)	19cm, 7°
Traditioneller Ansatz [SLK12] (2012)	5.1cm, 2.5°
Ansatz dieser Arbeit [ER18a] (2018)	3.6cm, 1.1°

Abb. 3: **Direkte Posen-Regression. Links:** Ein neuronales Netz erzeugt die gewünschte Ausgabe direkt. **Rechts:** Direkte Regression mit neuronalen Netzen erzielt in der Kamera-Relokalisierung keine guten Ergebnisse. Diese Arbeit kombiniert neuronale Netze mit traditionellen Ansätzen und ermöglicht damit eine sehr hohe Genauigkeit.

Im Jahr 2012 gewann ein *Convolutional Neural Network* (CNN) den bedeutenden ImageNet Wettbewerb für Bildklassifizierung mit einem großen Abstand zu allen konkurrierenden Methoden. Seitdem hat das sogenannte *Deep Learning*, also das maschinelle Lernen mittels neuronaler Netze einen beispiellosen Siegeszug in der Computer Vision und einigen anderen Wissenschaftszweigen angetreten. Für Aufgaben wie Bildklassifizierung,

2D-Objekt-Detektion und semantischer Segmentierung sind neuronale Netze im Moment unangefochten in ihrer Leistungsfähigkeit. Diese Methoden sind dabei oft sehr ähnlich aufgebaut. Das Eingabebild durchläuft die verschiedenen Schichten des neuronalen Netzes, wobei es schrittweise in die gewünschte Ausgabe transformiert wird, etwa in Wahrscheinlichkeiten für verschiedene Bildklassen. Das neuronale Netz erzeugt die gewünschte Ausgabe also direkt und unmittelbar aus der Eingabe, im folgenden *direkte Regression* genannt, siehe auch Abb. 3, links. Während dieses Vorgehen für viele Problemstellungen hervorragende Resultate erzielt, sind die entsprechenden Ergebnisse für 6D-Posenschätzung enttäuschend. Abb. 3, rechts führt die Ergebnisse einiger aktueller Ansätze von direkter Regression für das Problem der Kamera-Relokalisierung an. Die Genauigkeit stagniert seit 2017 bei ca. 20cm für die Lokalisierung innerhalb eines Zimmers und ist damit für Anwendungen wie *Augmented Reality* nicht brauchbar. Das ist insbesondere überraschend, da wesentlich ältere, traditionelle Ansätze (z.B. feature-basiertes Matching [SLK12]), welche keinerlei Form des maschinellen Lernens verwenden, diese Genauigkeit bei weitem übertreffen. Traditionellen Ansätze kombinieren von Menschen erdachte Bild-Features mit der Optimierung von geometrischen Bedingungen unter Kenntnis bestimmter Aspekte der Bildentstehung. Diese Verfahren haben den weiteren Vorteil, dass die geometrische Konsistenz der Vorhersagen geprüft werden kann. Damit sind die Vorhersagen zu gewissen Maße interpretierbar und mit einer Abschätzung ihrer Zuverlässigkeit verknüpft. Bei Ansätzen der direkten Regression handelt es sich dagegen zumeist um eine *Black Box*. Ihre Voraussagen resultieren aus einem nicht einsehbaren bzw. nicht interpretierbaren Prozess und werden ohne Sicherheitsabschätzung getroffen. Eine Voraussage kann also nur schwer bezüglich ihrer Vertrauenswürdigkeit abgeschätzt werden, was in kritischen Anwendungen wie dem autonomen Fahren ein essentielles Problem darstellt.

Ein wesentlicher Beitrag dieser Arbeit besteht darin, traditionelle Ansätze der Computer Vision mit den neuen Möglichkeiten des maschinellen Lernens zu kombinieren, ohne die traditionellen Ansätze aber vollständig zu ersetzen. Diese Arbeit zeigt am Beispiel der 6D-Posenschätzung, dass diese Kombination in einer erhöhten Genauigkeit resultiert und gleichzeitig wünschenswerte Eigenschaften der traditionellen Methoden erhält, beispielsweise eine teilweise Interpretierbarkeit und abschätzbare Vertrauenswürdigkeit der Vorhersagen. Im Folgenden werden die zentralen Forschungsbeiträge der Arbeit aufgeführt.

- Ein neues Verfahren zur 6D-Posenschätzung, das traditionelle Ansätze aus der robusten Optimierung und projektiven Geometrie mit Werkzeugen des maschinellen Lernens vereint.
- Das Verfahren zeichnet sich durch hohe Genauigkeit und Vielseitigkeit aus: Es unterstützt RGB- sowie RGB-D-Bilder als Eingabe und schätzt die Pose von texturierten oder nicht-texturierten Gegenständen sowie von ganzen Umgebungen.
- Ein zentraler Algorithmus der robusten Optimierung, welcher oft in traditionellen Verfahren der Computer Vision Verwendung findet, ist *Random Sample Consensus* (RANSAC). Diese Arbeit beschreibt eine differenzierbare Variante von RANSAC die in Kombination mit *Deep Learning* verwendet werden kann.

2 Lernen von Bild-Objekt-Korrespondenzen

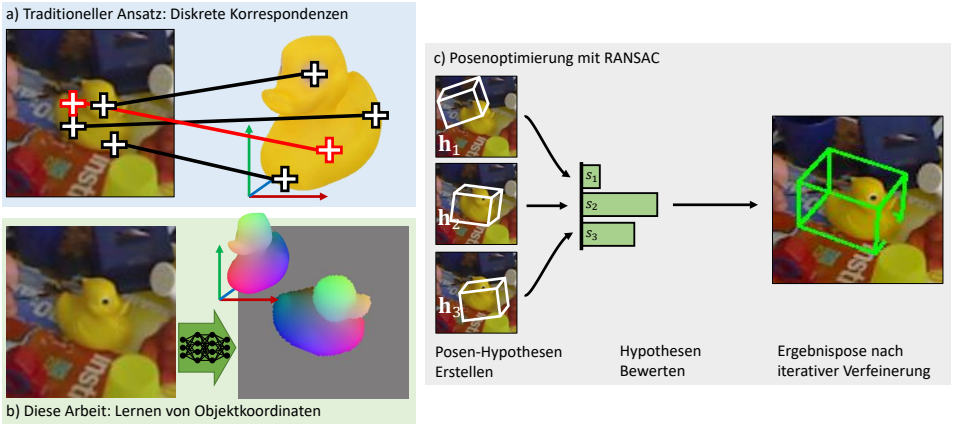


Abb. 4: **Objektkoordinaten-Regression. a)** Traditionell werden diskrete Korrespondenzen zwischen Bild und Objekt vorhergesagt. **b)** Diese Arbeit verwendet ein Lernverfahren um dichte, kontinuierliche Korrespondenzen zu schätzen. **c)** Die Posenschätzung erfolgt auf Grundlage der Korrespondenzen (a oder b) mit RANSAC.

Im traditionellen Ansatz erfolgt 6D-Posenschätzung in zwei Stufen. Zunächst wird mittels lokaler Bild-Features [Lo04] eine Anzahl von diskreten Korrespondenzen zwischen dem Eingabebild und dem Objekt gesucht, siehe Abb. 4 a). Am Beispiel der Abbildung konnten bestimmte Bildbereiche dem Auge, dem Flügel und der Vorderseite der Spielzeug-Ente richtig zugeordnet werden. In einem weiteren Fall (rot) schlug die Korrespondenzfindung fehl. Durch geometrische Optimierung kann aus den Korrespondenzen die 6D-Pose des Objekts geschätzt werden [Ka76, Ga03]. Die korrekte Pose sollte das gesuchte Objekt mit dem Bild entlang der Korrespondenzen in Übereinstimmung bringen. Erfolgt die geometrische Optimierung jedoch über alle Korrespondenzen, inklusive der stark fehlerbehafteten, wäre die geschätzte Pose nur von niedriger Qualität. Der RANSAC-Algorithmus [FB81] aus der robusten Optimierung ermöglicht genaue Schätzungen, auch wenn einige Korrespondenzen falsch sind, siehe Abb. 4 c). RANSAC wählt dazu mehrere zufällige Teilmengen von Korrespondenzen aus und erzeugt je eine Schätzung der Pose, sogenannte Hypothesen. Jede Hypothese wird dann bewertet hinsichtlich ihrer geometrischen Konsistenz mit allen übrigen Korrespondenzen. Die Hypothese mit der höchsten Konsistenz wird als finale Schätzung ausgewählt und eventuell noch durch einen iterativen Verfeinerungsprozess verbessert. Dieser Ansatz erzeugt genaue Ergebnisse und die geometrische Konsistenz des Ergebnisses lässt auf dessen Vertrauenswürdigkeit schließen. Das Verfahren kann fehlschlagen, wenn keine Korrespondenzen zwischen Objekt und Bild gefunden werden können, etwa wenn keine markanten Objekt-Strukturen für eine Zuordnung vorhanden sind.

In dieser Arbeit wird die oben genannte Strategie größtenteils beibehalten, jedoch der Schritt der Korrespondenzfindung durch ein Lernverfahren ersetzt. Zu diesem Zweck wird eine dichte, kontinuierliche Korrespondenzrepräsentation eingeführt, die sogenannten *Ob-*

jektkoordinaten. Jeder Punkt auf der Objektoberfläche hat eine eindeutige 3D-Koordinate im lokalen Koordinatensystem des Objekts, siehe auch Abb. 4 b) wo die Objektkoordinaten durch eine eindeutige Farbkodierung visualisiert werden. Das Lernverfahren entscheidet für jeden Bildbereich ob es sich um das Objekt oder den Hintergrund handelt. Falls der Bildbereich zum Objekt gehört, sagt das Lernverfahren weiterhin die korrespondierende 3D-Objektkoordinate voraus. Das Lernverfahren schätzt also eine Objektsegmentierung sowie ein dichtes, kontinuierliches Korrespondenzfeld in Form der Objektkoordinaten. Die Pose kann dann wie oben beschrieben mittels RANSAC geschätzt werden.

Als Lernverfahren für die Objektkoordinaten-Regression eignen sich Entscheidungsbäume [Br01], die mit gerenderten Ansichten des Objekts trainiert werden können. Dazu ist lediglich ein 3D-Modell des Objekts nötig, wie es in der Produktion oft verfügbar ist. Alternativ kann auch ein Datensatz mittels einer Tiefenkamera mit simultanem Posentracking erzeugt werden. Durch die Verwendung eines Lernverfahrens für die Korrespondenzfindung kann sich das technische System exakt auf das gewünschte Objekt spezialisieren. Damit wird auch die Posenschätzung von schwierigen, nicht-texturierten Objekten möglich, bei denen traditionelle Bild-Features versagen. Weiterhin können verschiedenste Umwelteinflüsse im Trainingsdatensatz simuliert werden, etwa starke Änderungen der Lichtverhältnisse. Das technische System lernt dann die entsprechende Robustheit. Einige experimentelle Ergebnisse können in Tabelle 1, links gesehen werden, wo das vorgeschlagene Verfahren mit einer nicht-gelernten Methode verglichen wird. Das vorgeschlagene Verfahren ist wesentlich robuster gegenüber teilweisen Verdeckungen und extremen Änderungen in der Beleuchtung.

3 Lernen von Unsicherheit

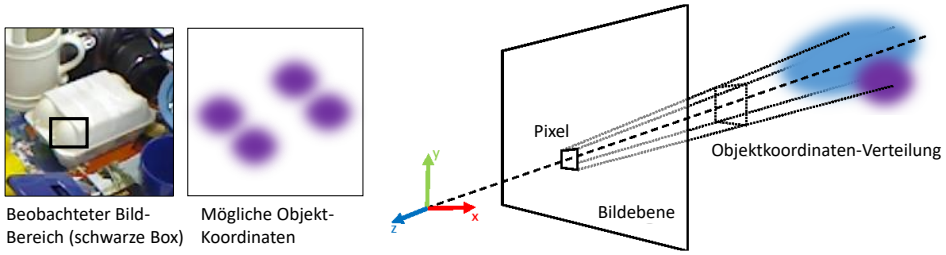


Abb. 5: **Unsichere Objektkoordinaten. Links:** Der beobachtete Bildbereich korrespondiert zu mehreren möglichen Objektkoordinaten. **Rechts:** Die optimale Pose maximiert die Likelihood der projizierten Objektkoordinaten-Verteilung.

Obwohl Lernverfahren selbst kleine Strukturmerkmale eines Objekts nutzen können, um eine Zuordnung zu Objektkoordinaten zu ermöglichen, gibt es Fälle, in denen keine eindeutige Zuordnung möglich ist, siehe Abb. 5, links. Das Objekt ist symmetrisch und der Bildbereich an der Objektecke kann nur auf vier mögliche Objektkoordinaten eingeschränkt werden. In dieser Arbeit wird die Korrespondenzvorhersage erweitert, indem *Verteilungen* von 3D-Objektkoordinaten in Form von *Gaussian Mixture Models* vorhergesagt werden. So kann das Lernverfahren Unsicherheit in der Objektkoordinaten-Schätzung

darstellen. Für eindeutige Korrespondenzen wird eine Verteilung mit einem Modus und geringer Standardabweichung vorhergesagt. Für mehrdeutige Korrespondenzen wird eine multimodale Verteilung mit gegebenenfalls hoher Standardabweichung vorhergesagt. In einer iterativen Optimierung wird die geschätzte Pose so verfeinert, dass die projizierte *Likelihood* unter den geschätzten Objektkoordinaten-Verteilungen maximiert wird, siehe Abb. 5, rechts. Die Posenverfeinerung mittels Objektkoordinaten-Unsicherheit führt insbesondere dann zu deutlich besseren Ergebnissen, wenn es sich beim Eingabebild lediglich um ein RGB-Bild handelt, siehe Tabelle 1, rechts.

RGB-D				RGB	
Methode	keine Verdeckung	mit Verdeckung	var. Lichtverhältnisse	Methode	
LINEMOD	96.6%	54.4%	70.2%	LINE2D	24.2%
Diese Arbeit	98.1%	67.3%	91.8%	Diese Arbeit	32.3%
				Diese Arbeit (mit Unsicherheit)	50.2%

Tab. 1: **Ergebnisse für Objektposenschätzung.** Die Eingabe ist entweder ein RGB-D-Bild (links) oder ein RGB-Bild (rechts). Ergebnisse werden als Prozent richtig geschätzter Posen angegeben. LINEMOD [Hi12] bzw. LINE2D [Hi11] sind nicht-gelernte Verfahren.

4 Differenzierbare Robuste Optimierung

Ein wesentlicher Faktor für den Erfolg von *Deep Learning* ist das Ende-zu-Ende-Training. Alle Schichten eines neuronalen Netzes, von den Bild-Features bis zu den semantischen Repräsentationen, können sich optimal aneinander anpassen um eine maximale Genauigkeit zu erreichen. Gleichermaßen soll die Objektkoordinaten-Regression in dieser Arbeit möglichst so gelernt werden, dass sich ihre Voraussagen gut für die robuste Posenoptimierung eigenen. Leider ist ein Trainieren des Gesamtsystems ende-zu-ende nicht möglich, da der RANSAC-Algorithmus nicht differenzierbar ist. Das bedeutet, dass die Korrekturrichtungen für eine Fehlerminimierung nicht durch RANSAC hindurch an die Objektkoordinaten-Regression weitergeleitet werden können.

Diese Arbeit stellt eine Variante von RANSAC vor, welche differenzierbar ist und daher im Rahmen von *Deep Learning* verwendet werden kann. Im Zentrum des differenzierbaren RANSAC, kurz *DSAC*, steht die Minimierung des folgenden Erwartungswerts:

$$\mathbb{E}_{j\sim p(j)}\left[\ell(\mathbf{h}_j)\right]$$

Dabei bezeichnet $p(j)$ eine Wahrscheinlichkeitsverteilung über alle RANSAC-Hypothesen $\mathbf{h}_j$, die sich aus den individuellen Konsistenzbewertungen ergibt, und $\ell(\cdot)$ ist eine Fehlerfunktion, die die Genauigkeit einer Pose angibt. Durch die Minimierung dieses Ausdrucks lernt das System gute Hypothesen hoch zu bewerten und ihre Genauigkeit weiter zu steigern und schlechte Hypothesen schlechter zu bewerten wobei ihre Genauigkeit keine Rolle spielt. Die Genauigkeitssteigerung durch ein Trainieren des Systems ende-zu-ende mittels DSAC kann in Tabelle 2 nachvollzogen werden. Im Moment

des Schreibens dieses Textes ist das hier vorgestellte System weltweit führend hinsichtlich der Genauigkeit im Kamera-Relokalisierungsproblem auf Basis eines einzelnen RGB-Bildes [BR18b]. Der vorgestellte DSAC-Algorithmus ist nicht auf die Anwendung in der Posenschätzung beschränkt, sondern kann, ähnlich wie RANSAC, für viele Probleme der Computer Vision und anderen Wissenschaftszweigen verwendet werden.

	Methode	Bilder korrekt	Genauigkeit (Median)
Verwandte Arbeiten	PoseNet [KGC15]	-	44.6cm, 9.8
	ORB-Feat. + RANSAC [Sh13]	38.6%	-
	Active Search [SLK12]	-	5.1cm, 2.5
Diese Arbeit	Entscheidungswald + RANSAC	55.2%	4.5cm, 2.0
	Neuronales Netz + RANSAC	61.0%	4.0cm, 1.6
	Neuronales Netz + DSAC	66.2%	3.5cm, 1.6
	Neuronales Netz + DSAC v.2	76.1%	3.6cm, 1.1

Tab. 2: **Ergebnisse für Kamera-Relokalisierung.** Von allen aktuellen Verfahren erzielt der Ansatz dieser Arbeit die höchste Genauigkeit. DSAC v.2 wurde auf Grundlage dieser Arbeit in [BR18b] veröffentlicht.

Fazit

Naturgemäß kann diese Kurzfassung nur einen oberflächlichen Überblick über die zugrundeliegende Dissertation geben. Insbesondere auf methodische, technische und experimentelle Details musste verzichtet werden. Beispielsweise kann das vorgestellte Verfahren die Posen von mehreren Dutzend Objekten gleichzeitig schätzen und weist damit eine gute Skalierbarkeit auf. Weiterhin kann das System über die Konsistenzprüfung einer Posenschätzung entscheiden, ob ein gesuchtes Objekt im Bild zu sehen ist oder nicht. Es wurde für artikulierte, also deformierbare, Objekte erweitert [Mi15] und in einen Echtzeittracker eingebettet [Kr14]. Die Dissertation gibt am Beispiel der 6D-Posenschätzung eine Antwort auf die Frage, wie die neuen Werkzeuge des *Deep Learning* genutzt werden können, ohne frühere Forschungsergebnisse, etwa aus der Computer Vision, komplett zu verwerfen. Insbesondere mit dem DSAC-Algorithmus stellt diese Arbeit ein wichtiges Werkzeug zur Verfügung, mit dem die mächtigen neuronalen Netze in etablierte technische Systeme eingebettet werden können.

Literaturverzeichnis

[Br01] Breiman, Leo: Random Forests. Machine Learning, 2001.

[Br18a] Brachmann, Eric: Learning to Predict Dense Correspondences for 6D Pose Estimation. Dissertation, Dresden University of Technology, Germany, 2018.

[BR18b] Brachmann, Eric; Rother, Carsten: Learning Less Is More - 6D Camera Localization via 3D Surface Regression. In: CVPR. 2018.

[Br18c] Brahmbhatt, Samarth; Gu, Jinwei; Kim, Kihwan; Hays, James; Kautz, Jan: MapNet: Geometry-Aware Learning of Maps for Camera Localization. In: CVPR. 2018.

[FB81] Fischler, M. A.; Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM, 1981.

[Ga03] Gao, Xiao-Shan; Hou, Xiao-Rong; Tang, Jianliang; Cheng, Hang-Fei: Complete solution classification for the perspective-three-point problem. TPAMI, 2003.

[Hi11] Hinterstoisser, S. Holzer, S.; Cagniart, C.; Ilic, S.; Konolige, K.; Navab, N.; Lepetit, V.: Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In: ICCV. 2011.

[Hi12] Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N.: Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In: ACCV. 2012.

[Ka76] Kabsch, Wolfgang: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1976.

[KC17] Kendall, Alex; Cipolla, Roberto: Geometric loss functions for camera pose regression with deep learning. In: CVPR. 2017.

[KGC15] Kendall, A.; Grimes, M.; Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Relocalization. In: ICCV. 2015.

[Kr14] Krull, Alexander; Michel, Frank; Brachmann, Eric; Gumhold, Stefan; Ihrke, Stephan; Rother, Carsten: 6-DoF Model Based Tracking via Object Coordinate Regression. In: ACCV. 2014.

[Lo04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. IJCV, 2004.

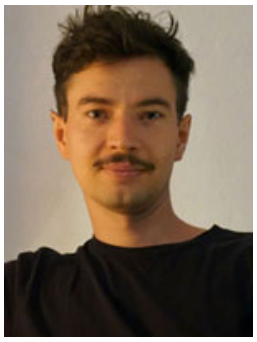
[Mi15] Michel, F.; Krull, A.; Brachmann, E.; Yang, M. Y.; Gumhold, S.; Rother, C.: Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression. In: BMVC. 2015.

[Sh13] Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: CVPR. 2013.

[SLK12] Sattler, Torsten; Leibe, Bastian; Kobbelt, Leif: Improving Image-Based Localization by Active Correspondence Search. In: ECCV. 2012.

[St12] Stallkamp, Johannes; Schlipsing, Marc; Salmen, Jan; Igel, Christian: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks, 2012.

[Wa17] Walch, Florian; Hazirbas, Caner; Leal-Taixé, Laura; Sattler, Torsten; Hilsenbeck, Sebastian; Cremers, Daniel: Image-based Localization with Spatial LSTMs. In: ICCV. 2017.



Eric Brachmann, geboren 1987, studierte Medieninformatik an der TU Dresden von 2006-2012 und schloss das Diplom mit Auszeichnung ab. Unmittelbar im Anschluss promovierte er am Lehrstuhl für Computergraphik und Visualisierung von Prof. Gumhold an der TU Dresden. Von 2015-2017 war er zusätzlich Mitarbeiter des Computer Vision Lab Dresden von Prof. Rother. Seit dem Abschluss der Promotion im Jahr 2018, arbeitet Eric Brachmann als PostDoc am Visual Learning Lab von Prof. Rother an der Universität Heidelberg. Sein Forschungsinteresse gilt der Verbindung traditioneller Verfahren der Computer Vision mit den Möglichkeiten des Deep Learnings.