

Abschlussbericht

(Beantwortung in Stichworten genügt)

Zuwendungsempfänger: Universität Ulm	Förderkennzeichen: 16ESE0108
Vorhabenbezeichnung: Allwettertaugliches Multi-Sensorsystem für das autonome Fahren – DENSE Teilvorhaben: DENSE-Signalverbesserung	
Laufzeit des Vorhabens: 1. Juni 2016 – 29.02.2020	
Berichtszeitraum: Schlussbericht	

1. Aufgabenstellung

Die Forschungsarbeiten erfolgten im ebenfalls von der EU geförderten Projekt DENSE mit verschiedenen Partnern. Die Schwerpunkte des Teilprojektes an der Universität Ulm lagen in der Verbesserung der sensorischen Erfassung der Fahrzeugumgebung für das automatisierte Fahren unter widrigen Witterungsbedingungen.

Die drei am häufigsten verwendeten Sensoren sind Kamera-, Lidar- und Radarsensoren, die alle ihre eigenen spezifischen Eigenschaften aufweisen und deren jeweiligen Vorteile erst durch ihre Kombination vollumfänglich hervortreten. Aktive Sensoren, das bedeutet, Lidar und Radar, haben gerade hinsichtlich kompromittierter Witterungsbedingungen Vorteile in Robustheit und Datenerfassung gegenüber reinen Kamerasystemen, die wiederum als passive Sensorik eine unübertroffene Informationsdichte aufweisen können.

Nach dem Stand der Technik wurden bisher Fusionsansätze weitgehend klassisch realisiert, wobei eine sensorindividuelle Vorverarbeitung erfolgt und dann die Ergebnisse auf Objekt- bzw. Detektionsebene fusioniert werden. Diese Vorgehensweise stößt jedoch insbesondere bei schlechten Witterungsbedingungen wie Starkregen, Nebel, Schnee aber auch Dunkelheit an seine Grenzen. Im Projekt wurden daher sowohl die Daten von Kamera- und Lidarsensoren als

auch von Kamera- und Radarsensorik algorithmisch und auf verschiedene Arten neuartig miteinander in Beziehung gesetzt. Die hierzu verwendeten künstlichen neuronalen Netze sind mittlerweile in vielen technischen Bereichen sowohl in Forschung als auch Anwendung voll etabliert und liefern regelmäßig in unterschiedlichsten Ausprägungen herausragende Ergebnisse für eine Vielzahl von Aufgaben. Die Architektur der in Anlehnung an menschliche Neuronenstrukturen entwickelten Netze ist einer der Hauptschwerpunkte gegenwärtiger Forschung auf diesem Gebiet weshalb ihr folgerichtig auch innerhalb des Projekts ein besonderer Stellenwert zukam. Insbesondere ging es um die zentrale Frage, an welcher Stelle und auf welche Weise innerhalb der Struktur des künstlichen neuronalen Netzes die Daten zusammengeführt und miteinander kombiniert werden sollten.

2. Wissenschaftlich-technische und andere wesentliche Ergebnisse

2.1 Sensorfusion von Kamera- und Lidardaten am Beispiel der Semantischen Segmentierung

Eine robuste Umgebungserfassung ist eine wichtige Grundlage für intelligente Systeme, wie z.B. Hausroboter oder automatisierte Fahrzeuge. Während viele Anwendungen, wie z.B. die semantische Segmentierung der Kamerabilder, sehr gute Ergebnisse in Gutwetterbedingungen erreichen, sinkt deren Performanz stark in Schlechtwetterbedingungen, wie z.B. bei Nebel, Schnee oder Regen. Aus diesem Grund macht es Sinn, verschiedene Sensordaten miteinander zu fusionieren, da in der Regel die Sensoren unter Schlechtwetterbedingungen asynchron ausfallen. So können z.B. bei Dunkelheit Objekte mit Hilfe einer Standardkamera nur sehr schwer erkannt werden, während sie für den Lidar-Sensor gut sichtbar sind. Im Folgenden wird ein Sensordatenfusionsansatz beschrieben, der Kamera- und Lidardaten miteinander fusioniert und am Beispiel der Semantischen Segmentierung evaluiert.

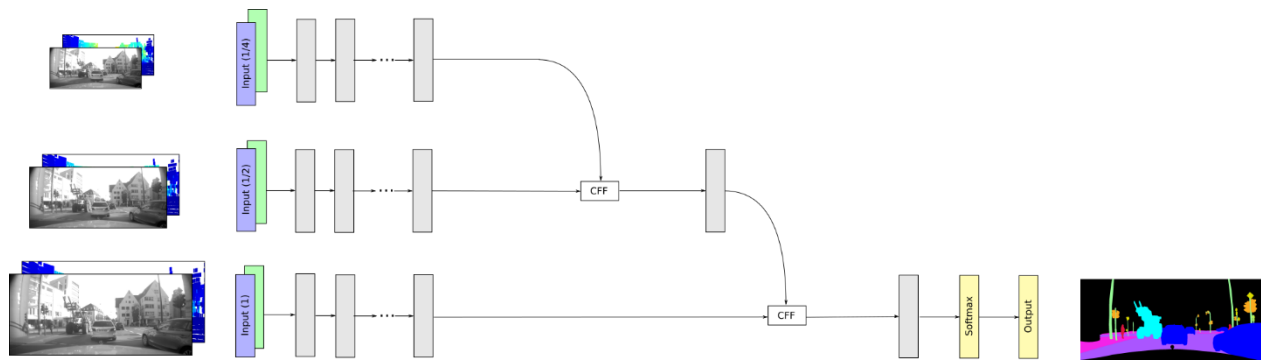


Abbildung 1 Netzwerkarchitektur des Early-Fusion Ansatzes

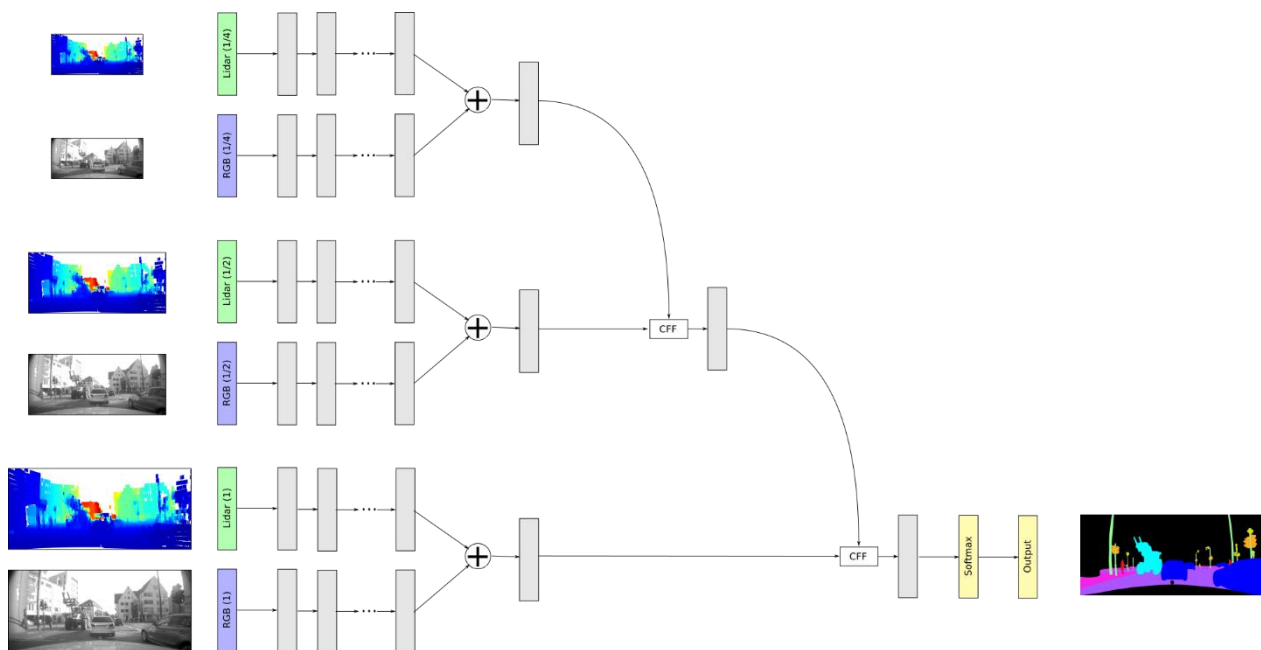


Abbildung 2 Netzwerkarchitektur des Late-Fusion Ansatzes

Bei der Semantischen Segmentierung wird jedem Bildpixel eine Klasse zugeordnet. So sollen z.B. alle Fußgänger im Bild rot markiert werden, alle Autos blau und die Straße lila. Ein Beispiel eines Verfahrens mit hoher Detektionsrate ist das ICNet, das zudem noch echtzeitfähig ist, weshalb es als Grundlage für den Segmentierungsansatz verwendet wird. In Rahmen von diesem Projekt werden zwei verschiedene Sensorfusionsansätze betrachtet, nämlich die sogenannte Early-Fusion und die Late-Fusion. Bei der Early-Fusion werden die Sensordaten zu Beginn des Neuronalen Netzes miteinander fusioniert. Dazu wird die Dimension des Eingangstensors von drei auf vier erweitert. Die ersten drei Dimensionen enthalten wieder die drei Farbkanäle des Kamerabilds, während die vierte Dimension das dazugehörige Tiefenbild enthält, das aus den

Lidardaten generiert wird. In Abbildung 1 ist die dazugehörige Netzarchitektur im Falle der semantischen Segmentierung abgebildet. Im Gegensatz dazu werden bei der Late-Fusion die Sensordaten separat zueinander verarbeitet, sodass für jeden Sensor und für jede Auflösung eine unabhängige Featuremap bestimmt wird. Die Featuremaps der beiden Sensoren werden dann für jede Auflösung miteinander fusioniert, bevor die verschiedenen Auflösungen miteinander vereinigt werden. Die dazugehörige Netzarchitektur ist in Abbildung 2 dargestellt.

Beide Fusionsansätze wurden auf einen Datensatz bestehend aus Kamera und Lidardaten trainiert, der im Rahmen dieses Projekts aufgebaut wurde. Um die Netzwerke robuster gegen Störungen zu machen, werden während dem Training künstliche Störungen entweder das Kamerabild oder in das Tiefenbild hinzugefügt.

Tabelle 1 Evaluierung bei optimalen Wetterbedingungen

	acc.	mIoU
ICNet (Kamera)	96.09	51.66
Early-Fusion	96.17	51.43
Late-Fusion	96.20	52.04

Tabelle 2 Evaluierung bei nichtoptimalen Wetterbedingungen (simulierter Regen)

	acc.	mIoU
ICNet (Kamera)	84.77	16.38
Early-Fusion	85.64	19.96
Late-Fusion	91.17	26.24

Der Early und der Late-Fusion Ansatz wird im Anschluss anhand zwei verschiedener Evaluierungsmetriken evaluiert, nämlich der pixelweisen Accuracy and der mean Intersection of Union (mIoU) und mit dem ursprünglichen ICNet verglichen. Die Ergebnisse (siehe Tabelle 1) zeigen, dass die beiden fusionsbasierten Ansätze bei optimalen Wetterbedingungen gleich gut oder leicht besser sind als der rein-Kamera-basierte Ansatz, wobei der Late-Fusion Ansatz die besten Ergebnisse liefert. Allerdings ist der Unterschied bei nichtoptimalen Wetterbedingungen vergleichsweise größer (siehe Tabelle 2). Dabei wurden die Kamerabilder leicht modifiziert, indem

ein regnerisches Wetter simuliert wurde. Der Unterschied wird nochmals in *Abbildung* visuell verdeutlicht. Während der klassische Segmentierungsalgorithmus die Straße kaum erkennt und noch gleichzeitig viele Geisterobjekte detektiert, kann der LateFusion-Ansatz die Straße immer noch gut erkennen.

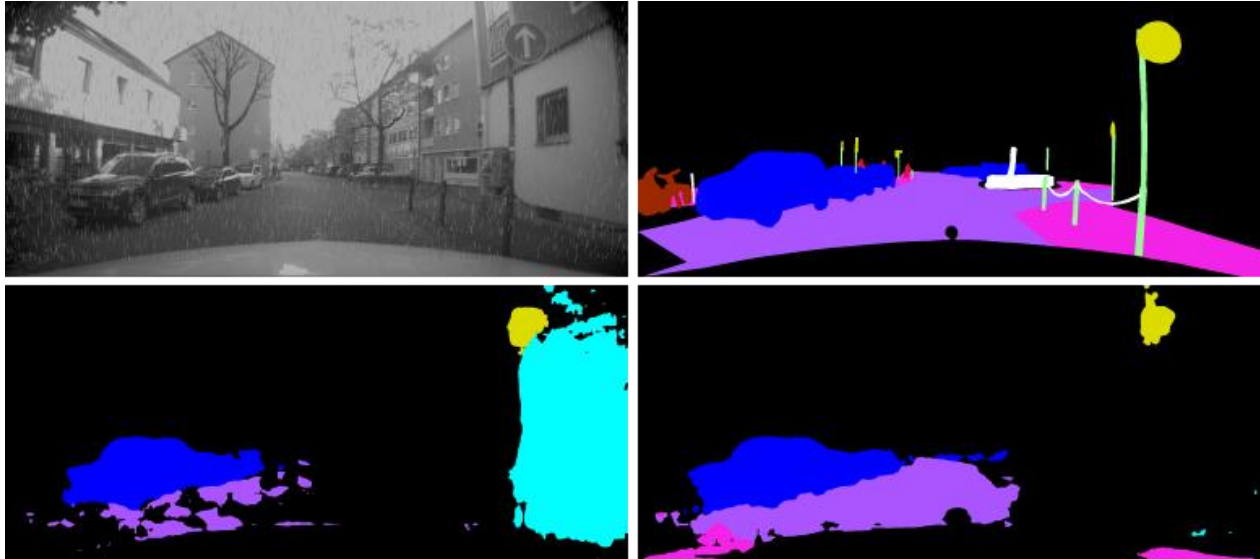


Abbildung 3: (links oben) Eingangsbild (rechts oben) dazugehörige Ground-Truth (links unten) Ergebnis eines state-of-the-art Algorithmus (rechts unten) Ergebnis des vorgestellten Late-Fusion Ansatzes

Robuste Semantische Segmentierung mit Hilfe von Videosequenz-Segmentierung

Eine zweite Möglichkeit, die Robustheit von Semantischen Segmentierungsansätze zu steigern ist neben einer geeigneten Sensordatenfusion (siehe vorherigen Abschnitt) die Videosequenz-Segmentierung. Die Motivation hierfür ist, dass die Segmentierungsergebnisse innerhalb einer Videosequenz stark schwanken. So flackern z.B. innerhalb einer Videosequenz die Objektkanten. Außerdem werden einige Teile des Objekts in einem Frame falsch klassifiziert, während sie im nächsten Frame korrekt erkannt werden. In den meisten Fällen treten die beschriebenen Fehler nur in einem oder wenigen Frames der Videosequenz auf, während sie in den nächsten Zeitschritten korrekt klassifiziert werden. Eine Möglichkeit die Robustheit der Videosegmentierung zu erhöhen, ist neben den Bildinformationen aus dem aktuellen Zeitschritt die Bildinformationen der vorherigen Zeitschritte zu verwenden. Dazu können rekurrente Neuronale Netze, wie z.B. convolutional Long-Short-Term-Memories (convLSTMs) verwendet werden. Aus diesem Grund wurde ein echtzeitfähiges state-of-the-art Segmentierungsverfahren, das ICNet, an geeigneten

Stellen mit convLSTM-Zellen erweitert. Das erweiterte ICNet wird im Folgenden LSTM-ICNet genannt. Insgesamt wurden sechs verschiedene Netzwerkarchitekturen entwickelt, wobei im Folgenden nur auf die besten drei Versionen, d.h. auf Version 2, Version 5 und Version 6, eingegangen wird. Bei der LSTM-ICNet Version 2 wird eine convLSTM-Zelle direkt vor dem Softmax-Layer hinzugefügt, was einer zeitlichen Filterung der Ausgabe entspricht. Bei der Version 5 wird jeweils eine convLSTM-Zelle am Ende jedes Resolution-Branches platziert, sodass die bestimmten Bild-Features über die Zeit gespeichert werden können. Version 6 ist schließlich eine Kombination aus Version 2 und Version 5. Die betrachteten Netzarchitekturen sind in Abbildung 4 dargestellt. Dabei veranschaulichen die grauen Boxen die ICNet-Struktur, während die farbigen Boxen den möglichen Positionen der convLSTM-Zellen entsprechen.

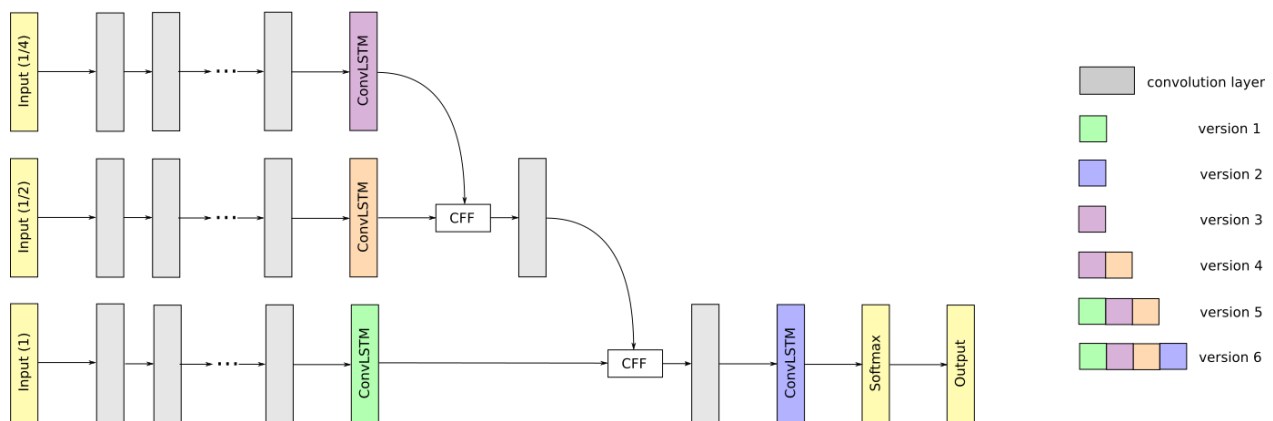


Abbildung 4 Netzarchitektur des LSTM-ICNets

Die ursprüngliche ICNet und die verschiedenen LSTM-ICNet Versionen wurden anhand des Cityscapes Datensatz evaluiert, wobei eine Zeitsequenz von vier Bildern verwendet wurde. Jeder der beschriebenen Ansätze wurde unter identischen Trainingsbedingungen trainiert und analog zum vorherigen Kapitel anhand von „pixelwise accuracy (acc.)“ und „mean Intersection over Union“ evaluiert. Die Ergebnisse (siehe Tabelle 3) zeigen, dass die Performanz des Segmentierungsalgorithmus durch die Verwendung von zeitlichen Bildinformationen verbessert wird. So konnte z.B. die Performanz bis zu 0,25% in Bezug auf „pixelwise accuracy“ und bis zu 1% in Bezug auf „mean Intersection over Union“ gesteigert werden.

Tabelle 3 Evaluierung auf dem Cityscapes-Datensatz

	acc.	mIoU
ICNet	92.92%	61.70%
LSTM-ICNet version 2	93.04%	62.36%
LSTM-ICNet version 5	93.15%	62.68%
LSTM-ICNet version 6	93.17%	62.76%

Um die Robustheit der verschiedenen LSTM-ICNet Versionen zu testen, wurde das letzte Bild der Videosequenz durch unbekannte Störungen gestört, z.B. durch simulierten Regen. Dabei wurde der Regen simuliert, indem der Bildkontrast um 30% reduziert wurde, da Regentage gewöhnlich dunkler sind als sonnige Tage. Außerdem wurden zufällig N kleine graue Linien mit einer Länge von l Pixeln in das Bild gezeichnet. Zur Evaluierung wurden unterschiedliche Regenintensitäten simuliert, wie z.B. leichter Regen ($N=500, l=10$), mittelstarker Regen ($N=1500, l=30$) und starker Regen ($N=2500, l=60$). Tabelle 4 zeigt die Evaluierungsergebnisse bei starkem Regen, die zeigen, dass die LSTM-ICNet Versionen robuster gegen diese Störungen sind als das normale ICNet. Beispielsweise übertrifft die LSTM-ICNet Version 5 das ICNet um 32% in Bezug auf Accuracy und um 23% in Bezug auf mIoU.

Tabelle 4 Evaluierung bei simuliertem starkem Regen

	acc.	mIoU
ICNet	35.85%	14.29%
LSTM-ICNet version 2	78.20%	37.16%
LSTM-ICNet version 5	78.87%	37.79%
LSTM-ICNet version 6	77.92%	37.58%

Abbildung zeigt die mIoU-Kurve der betrachteten Ansätze in Abhängigkeit der Regenintensität. Der Graf zeigt, dass die Performanz von allen Verfahren bei zunehmender Regenstärke abnimmt, wobei die LSTM-ICNet Version 5 am besten abschneidet. Außerdem wird der Abstand zwischen dem ICNet und den LSTM-ICNet immer größer, je stärker der Regen wird. Die LSTM-ICNet Versionen sind daher robuster gegen diese Störungen als das ursprüngliche ICNet.

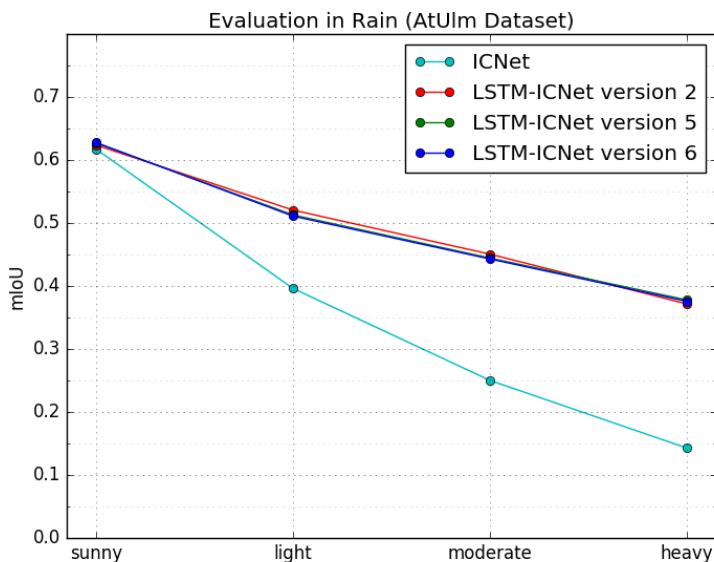


Abbildung 5 mIoU-Kurve in Abhängigkeit der Regenintensität

In Abbildung wird das ICNet und die LSTM-ICNet Version 5 qualitativ am Beispiel von starkem Regen verglichen. Die qualitative Analyse zeigt, dass die LSTM-ICNet Version 5 robuster als das ICNet ist. So flackert beispielsweise in dieser Videosequenz die Straßenkante auf der linken Seite weniger und der rechte Bürgersteig wird zuverlässiger. Außerdem treten weniger Störungen im Bild auf, wie es z.B. beim ICNet in Abbildung zu sehen ist.

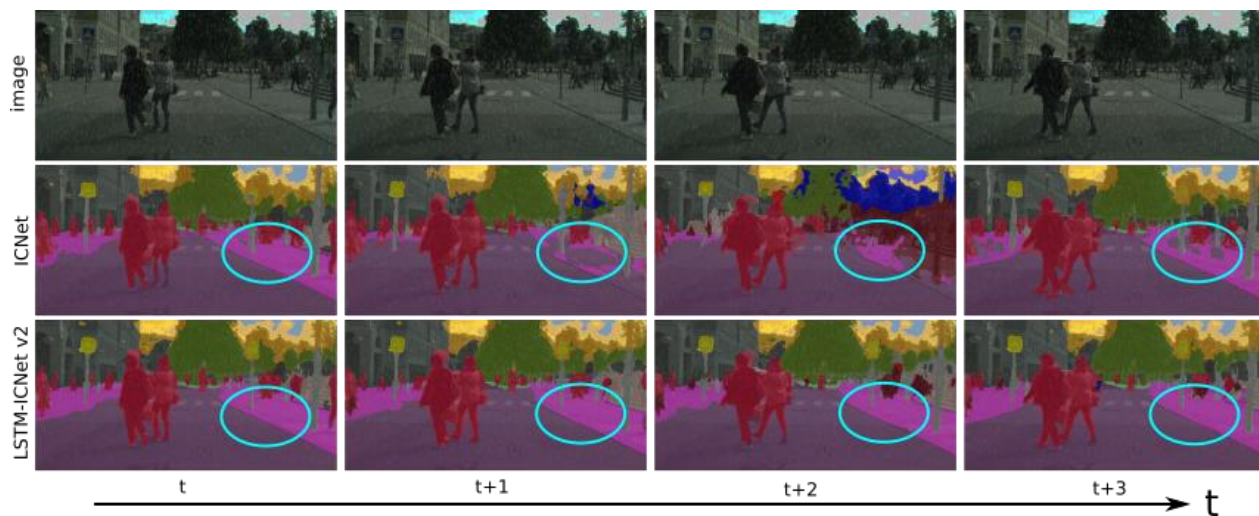


Abbildung 6 qualitativer Vergleich der LSTM-ICNet Version 5 mit dem ICNet

Die oben beschriebene Evaluierung zeigt, dass die Performanz und die Robustheit des LSTM-ICNets im Vergleich zum ursprünglichen ICNet steigt, allerdings auf Kosten der Rechenzeit. Beispielsweise erhöht sich die Inferenzzeit der LSTM-ICNet Version 2 um ca. 35% von 48ms auf 65ms, obwohl nur eine convLSTM-Zelle verwendet wird. Die anderen LSTM-ICNet Versionen, die mehr convLSTM-Zellen enthalten, brauchen noch länger. Das Problem ist, dass die convLSTM-Zellen rechenintensiv sind, da innerhalb einer Zelle acht Faltungen berechnet werden müssen. In der Literatur gibt es mehrere Möglichkeiten, wie man diese standard convolution-layer beschleunigen kann, z.B. durch spatial separable convolutions, depthwise convolution und depthwise separable convolutions.

Diese Ansätze wurden daher im Folgenden verwendet, um die Berechnung der convLSTM-Zelle zu beschleunigen. Im Folgenden werden drei convLSTM-Versionen betrachtet, dessen Inferenzzeit kürzer ist als die der standard convLSTM, nämlich die spatial-convLSTM, bei der jede Faltung innerhalb der convLSTM-Zelle durch spatial-separable convolutions ersetzt wird, die depth-convLSTM, bei der jede Faltung durch depthwise convolutions ersetzt wird und die sep-convLSTM, bei der jede Faltung durch eine depthwise separable convolution ersetzt wird. Mit Hilfe dieser Änderungen können die Rechenzeit und der Speicherverbrauch der LSTM-ICNet Versionen stark reduziert werden, wie es in Tabelle 5 am Beispiel der LSTM-ICNet Version 2 gezeigt wird. So kann z.B. die Anzahl der Floating-Point Operations (FLOPs) bei den spatial-convLSTMs um ca. 19% und bei den depthwise-convLSTMs um ca. 57% reduziert werden.

Tabelle 5 Vergleich des Rechenaufwands der verschiedenen convLSTM-Versionen

	standard	spatial	depthwise	separable
GFLOPs	135.74	109.97	59.03	67.62
	100%	81.01%	43.49%	49.82%
Inference time (CPU)	2503ms	2265ms	1616ms	1734ms
	100%	90.47%	64.56%	69.26%
	65.01ms	68.15ms	60.62ms	62.18ms

Inference time (GPU)	100%	104.83%	93.24%	95.65%
----------------------	------	---------	---------------	--------

Außerdem kann die Inferenzzeit auf der CPU um bis zu 35% verringert werden. Die Beschleunigung auf der GPU ist jedoch nicht ganz so groß wie auf der CPU. Der Grund dafür liegt darin, dass die Standard-3x3-Faltungen in Deeplearning Frameworks wie Tensorflow oder Pytorch im Gegensatz zu den spatial-convolutions und depthwise-convolutions stark optimiert sind. Durch die Optimierung dieser Operationen kann die Inferenzzeit auch auf der GPU weiter reduziert werden.

Die schnellere Inferenzzeit könnte auf Kosten der Performanz gehen. Daher werden die vorgestellten Ansätze auf dem Cityscapes Datensatz hinsichtlich der pixelwise accuracy und mIoU evaluiert. Die entsprechenden Ergebnisse sind in Tabelle 6 dargestellt, wobei sich herausstellte, dass das LSTM-ICNet eine ähnliche Performanz hat, wenn die standard convLSTM Zelle durch eine spatial-convLSTM Zelle ersetzt wird. Im Gegensatz dazu sinkt die Performanz, wenn die depthwise-convLSTMs anstatt der standard convLSTMs verwendet werden. Sie übertreffen jedoch immer noch die Performanz des ursprünglichen ICNets, wie man der Tabelle 6 entnehmen kann.

Tabelle 6 Evaluierung der verschiedenen convLSTM-Versionen auf dem Cityscapes-Datensatz

	acc.	mIoU
ICNet	92.66%	61.55%
standard convLSTM	93.05%	62.71%
spatial convLSTM	93.05%	62.88%
depth convLSTM	92.92%	61.65%
sep convLSTM	92.81%	62.22%

2.2 Verbesserte Fusion von monokularen Kameradaten mit Radar Signal Signaturen

Während vereinzelt Anbieter automatisierter Fahr- und Sicherheitsfunktionen auf rein Kamera-basierte Systeme zur Gewährleistung der hohen Sicherheitsanforderungen setzen, hat sich mittlerweile bei den meisten Herstellern zumeist die Ansicht durchgesetzt nur durch eine Kombination und entsprechende Ausnutzung der unterschiedlichen Vorteile eines breiten Spektrums von Sensorik, allen relevanten Hindernissen zuverlässig begegnen zu können. Gerade die Hinzunahme von Radarsystemen hat sich hier in der jüngsten Vergangenheit als unerlässlich herausgestellt um insbesondere bei unvorteilhaften Witterungsbedingungen weiterhin robust und zuverlässig die Fahrzeugumgebung erfassen zu können. Die hierfür notwendige Rada-signalverarbeitung ist wie schon zuvor beschrieben im Vergleich zu Laser oder Kamerasystemen ungleich komplexer und im Gegensatz zu diesen muss zunächst eine Reihe von Schritten zur Datenaufbereitung vollzogen werden, um die im Rohsignal verborgenen Informationen einer weiteren Analyse zielgerichtet zugänglich zu machen.

Dafür bieten die resultieren bildgebenden Radarsysteme dann aber charakteristische Eigenschaften wie sie kein anderer Sensor heutzutage vorweisen kann. Neben einer direkten und simultanen Erfassung der Distanz und der relativen Geschwindigkeit zwischen Sensor und Hindernissen sind Radarsensoren ebenfalls in der Lage auch kleinste Abstände zu Objekten mit großer Genauigkeit zu erfassen. Aus diesem Grund finden Radare neben Adaptive Cruise Control Systemen (ACC) heutzutage unter anderem bei Einpark- oder Notbremsassistentensystemen Verwendung. Zukünftige Anwendungsgebiete sind eine intelligente, datengestützte, echtzeitfähige Umgebungserfassung als weitere Stütze zur Absicherung autonomer Fahrfunktionen.

Radarsensorik funktioniert wie bereits angedeutet auch bei widrigen Witterungsbedingungen zuverlässig und ist robust gegenüber schwankenden Lichtverhältnissen, weshalb sie sich insbesondere für den Einsatz bei Nacht-, Nebel, oder Schneefahrten eignen, was auch unmittelbar zum zentralen Punkt des gegenwärtigen Forschungsprojektes überleitet.

Die Fortschritte im Bereich autonomen Fahrens der letzten Jahre sind zu großen Teilen auf Szenarien beschränkt gewesen, in denen vorteilhaftes Wetter herrschte und nahezu in allen Fällen wurden überwachte Lernverfahren verwendet, die entsprechend gekennzeichnete Datensätze der jeweiligen Sensoren voraussetzen. Und obwohl es zahlreiche Datensätze zur Erfassung nahezu jedes Verkehrsszenarios mittels Lidar- und Kamerasensorik gibt, so ist Vergleichbares bei Radarsystemen bis zum heutigen Tage praktisch nicht existent. Hier war also eine alternative Möglichkeit zu erdenken, welche eine manuelle Annotation durch visuelle Inspektion (Human-in-the-loop approach) umgeht unter gleichzeitiger Nutzung eines möglichst hohen Grads des Sensorpotentials. In anderen Worten, wie kann die volle Ausschöpfung der Sensorkapazität eines Radars unter Einbeziehung möglichst aller wertvollen Informationen gewährleistet werden, was implizit das Manipulieren durch vorgelagerte Signalverarbeitungsalgorithmik ausschließt und gleichzeitig der Mehrwert des Radars zu seiner maximalen Entfaltung gebracht werden. Und genau hier setzt das beschriebene Vorgehen an, bei dem versucht wird, ein Maximum an Informationsgehalt für eine spätere Entscheidungsfindung zu erhalten, worauf der Fusionsansatz, realisiert durch ein künstliches neuronales Netz, dann selbstständig entscheiden soll, welche Informationen relevant sind. Es gilt sich bewusst zu machen, dass grundsätzlich alle durchgeführten Bearbeitungsschritte Informationen ändern, verfälschen oder sogar vernichten. Das Ziel ist deshalb, sämtliche Daten möglichst nah am Sensor abzugreifen, um einem entsprechenden Abfluss vorzubeugen. Aus gleichem Grund hinderlich und durch die Komplexität und unintuitive Natur von Radardaten nahezu unmöglich ist weiterhin eine Annotation assoziierter Daten, wie dies nach wie vor bei Kamera und zunehmend bei Lidardaten geschieht. Hierdurch ergeben sich ebenfalls vollkommen neue Herausforderungen die in den Bereich des sogenannten selbstüberwachten Lernens überführen, wobei versucht wird, Korrespondenzen zwischen multimodalen Sensordatenflüssen zu erkennen. Die Informationen eines Sensors werden hierbei zur Extraktion und zum Erkenntnisgewinn innerhalb des jeweils anderen herangezogen. Krude Annahmen, Heuristiken oder das Setzen von Schwellwerten wird somit umgangen und das neuronale Netz selbst in die Lage versetzt, die für sich essentiellen Informationen zu identifizieren. Ein weiterer Vorteil ist hierbei, dass es sich gleichwohl um Informationen handeln kann, die von einem menschlichen Annotator unerkannt oder missachtet worden wären. Ermöglicht wird also ein theoretisch breiteres Spektrum von Erkennungsmöglichkeiten und ein bislang verkanntes riesiges Informationspotential innerhalb der Datenströme. Im Gegensatz hierzu, ist es bei überwachten Lernverfahren naturgegeben nur möglich, Objekte und Situationen in der Fahrzeugumgebung wahrzunehmen, die zuvor in den Trainingsdaten manuell und händisch als solche gekennzeichnet worden sind. Dies verdeutlicht

einmal mehr, den wesentlichen Unterschied zwischen beiden Zugängen und hebt die Vorteile selbstüberwachter Lernverfahren hervor, die auch deshalb in der letzten Zeit einen maßgeblichen Aufschwung in der Forschungsgemeinschaft erlebt haben.

Um, wie oben erwähnt, überwachende Signale der einen Modalität zur Kontrolle der jeweils anderen heranziehen zu können, sind zeitlich genau aufeinander abgestimmte Datenströme aus Radar und Kamera notwendig. Danach werden zunächst intermodal-unabhängige hochdimensionale Merkmalsrepräsentationen, sogenannte Feature Vektoren für sowohl Kamera als auch Radaraufnahmen erzeugt, indem einem sogenannten künstlichen neuronalen Zwillingnetzwerk im Wechsel zueinander passende, sowie zeitlich nicht passende Beispiele präsentiert werden. Die Annotation der Daten, die für den Lernprozess, das heißt, die Anpassung der Netzwerkgewichte mittels des Backpropagation Algorithmus, notwendig ist, wird hier durch eine sogenannte proxy-task in Form einer binären Klassifikationsentscheidung erfüllt. Sie ergibt sich hier eleganterweise direkt aus der Problemstellung und ist a-priori bekannt, d.h. es werden hier lediglich Übereinstimmung und Diskrepanz binär encodiert. Der manuelle Eingriff in den Lernprozess des Netzes bleibt somit minimal und das Netzwerk ist in der Lage semantisch wertvolle Informationen in hochdimensionalen Feature Vektoren zu erlernen, welche für die im folgenden beschriebene Zielsetzung der Lokalisierung elektromagnetische Wellen reflektierender Objekten im Kamerabild verwendet werden.

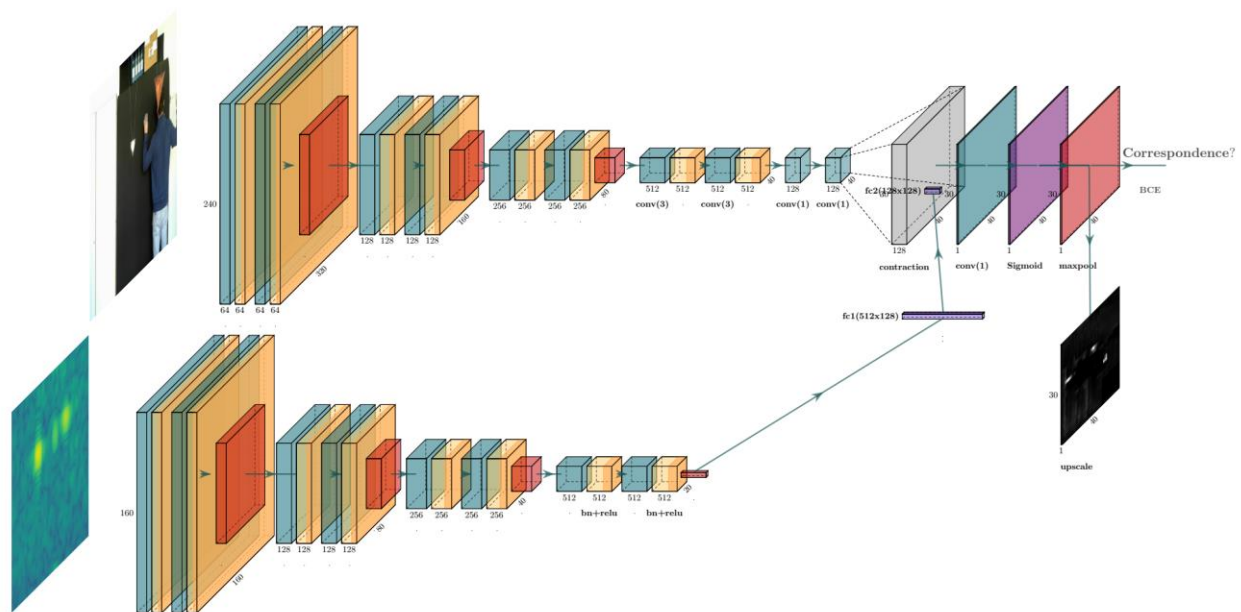


Abbildung 7 Erweiterte neuronale Architektur zur Lokalisierung von EM-Reflexionen

Die Grundidee ist, dass der 512-dimensionale Merkmalsvektor aus dem Radarzweig des Netzwerks als Suchmaske für die hochdimensionale Kamera Repräsentation verwendet und somit implizit eine Metrik vorgibt, die bei Übereinstimmungen an den jeweiligen Positionen im Merkmalsraum einen vergleichsweise hohen Wert, bei merklichen Unterschieden einen relativ geringen Wert ausgibt. Die Projektion dieser Informationen auf das räumliche Gitter des transformierten Kamerabildes (grau in Abbildung 7) erfolgt anschließend durch eine Sequenz von Faltungen mit einem eindimensionalen Kernel (grüne Schicht in Abbildung 7) bevor eine Sigmoid Funktion die resultierende Ähnlichkeitsdarstellung auf das Einheitsintervall transformiert (lila Schicht in Abbildung 7). Diese Abfolge ermöglicht es dem Netzwerk genau die Regionen innerhalb des Kamerabildes mit hohen Werten hervorzuheben, an denen es die Quelle von Radarreflektionen vermutet und welche folglich hell auf der Korrespondenzkarte erscheinen. Eine abschließende Anwendung der Maximum Funktion auf die gesamte Lokalisierungsfläche ergibt einen skalaren Wert, der für die unterlegte binäre Klassifikationshilfsaufgabe (proxy task) benötigt wird.

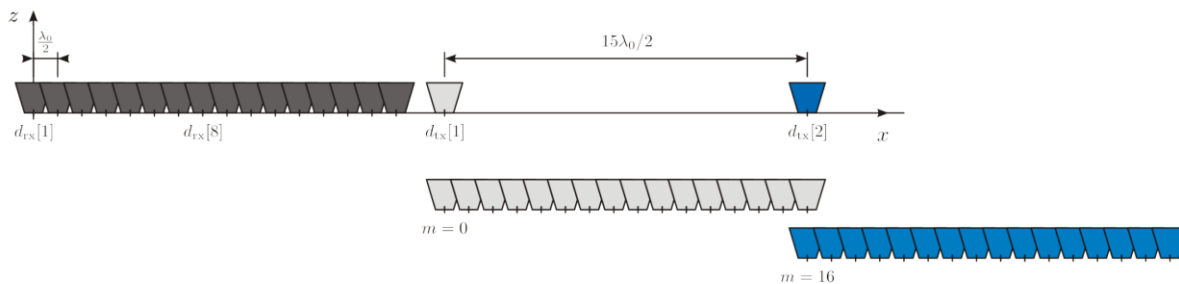


Abbildung 8 Räumliche Faltung von Sende- und Empfangsantennen (Virtuelle Apertur)

Zuvor wurden zu diesem Zweck sogenannte Range-Doppler Darstellung der Radarsignale verwendet, welche bereits wertvolle Informationen zum radialen Abstand sowie Relativgeschwindigkeit zwischen Zielen und Radarsensoren bereitstellen und damit die zweidimensionale Projektion der Kamerabilder vorteilhaft ergänzen. Es besteht aber auch die Möglichkeit, die Radardaten ebenfalls ohne aufwendige Signalverarbeitungsschritte um eine weitere laterale räumliche Dimension zu ergänzen und somit dem Netzwerk Winkelinformationen bereitzustellen. Mithilfe mehrerer Empfangsantennen eines Radarsensors, die in einem definierten räumlichen Abstand zueinander angebracht sind, ist es möglich, durch kleinste Laufzeitunterschiede beim Empfang der reflektierten Wellen, eine Winkelschätzung, also eine Bestimmung der Empfangsrichtung in azimuthaler Richtung vorzunehmen. Die Winkelauflösung ist

hierbei durch die Größe der Apertur, also der Breite des Empfangszweigs gegeben, und wird demnach implizit durch die Anzahl der Empfangsantennen bestimmt. Durch das räumliche Falten von Sende- mit Empfangsantennen bei Multiple Input, Multiple Output (MiMo) Radaren ist es weiterhin möglich, die Apertur Größe um beinahe das Zweifache zu steigern, was entsprechend eine doppelte Winkelauflösung bedeutet. Abbildung 8 stellt diesen Sachverhalt anschaulich dar. Datenbezogene Ergebnisse dieser Vorgehensweise sind in Abbildung 9 aufgezeigt, wobei die Abszisse den Winkel und die Ordinate die radiale Entfernung bedeuten. Auf der linken Abbildung ist das Ergebnis bei 8 Empfangskanälen und einer Sendeantenne dargestellt, während die mittlere Darstellung aus der Verwendung von 16 Empfangsantennen resultiert. Die rechte, und offensichtlich lokalisierteste Winkelbestimmung des Ziels im Sichtfeld des Radars erfolgt bei der zusätzlichen Hinzunahme einer weiteren Sendeantenne und der Bildung einer sogenannten virtuellen Apertur, wodurch in diesem Falle eine Winkelauflösung von knapp 3 Grad erzielt werden kann.

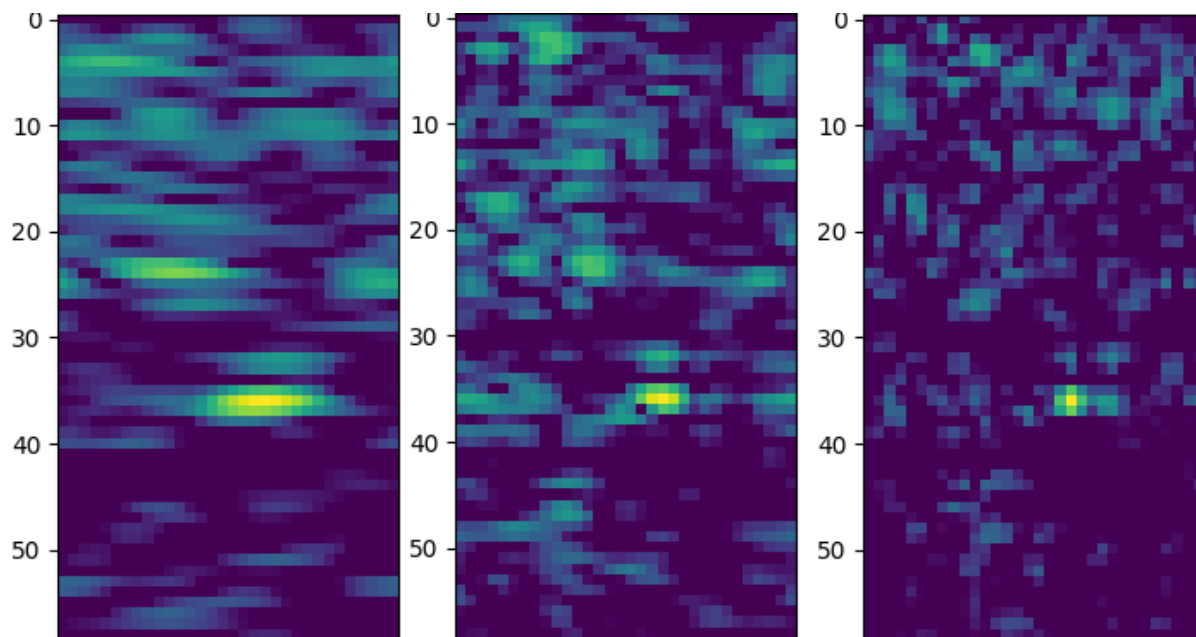


Abbildung 9 Verbesserung der räumlichen Trennfähigkeit durch virtuelle Apertur

Kombiniert man Range-Doppler Darstellungen mit der hinzugewonnenen Winkelauftragung, erhält man eine radarspezifische Kubusdarstellung der Daten, die nur minimal prozessiert sind und in der jede Dimension die jeweils zuvor gezeigten Informationen von Entfernung, Geschwindigkeit und Winkel kodieren und die zuvor beschriebenen Darstellungen als Orthogonal Projektion entlang jeweils einer Achse beinhalten. Dieser Radarwürfel in Abbildung 10 beinhaltet alle Informationen des Sensors, die innerhalb eines Frames extrahierbar sind, eben auch die

Zieldarstellung nur sehr schwach reflektierender Objekte getrennt nach Abstand, Winkel und Geschwindigkeit. Mit Blick auf diese Abbildung treten die beiden Ziele deutlich aus dem Grundrauschen hervor und sind trotz annähernd gleicher Entfernung zum Sensor doch über ihren unterschiedlichen Winkel und Geschwindigkeiten sehr gut trennbar. Bei realistischen, komplexeren Szenarien, wie sie im Straßenverkehr auftreten, ist die Darstellung des Radarwürfels ungleich komplizierter und oftmals sind Signale überlagert, was eine unmittelbare Zuordnung zwischen Erscheinung im Kamerabild und Darstellung im Würfel erschwert und erneut die Sinnlosigkeit einer manuellen Annotation dieser Daten unterstreicht.

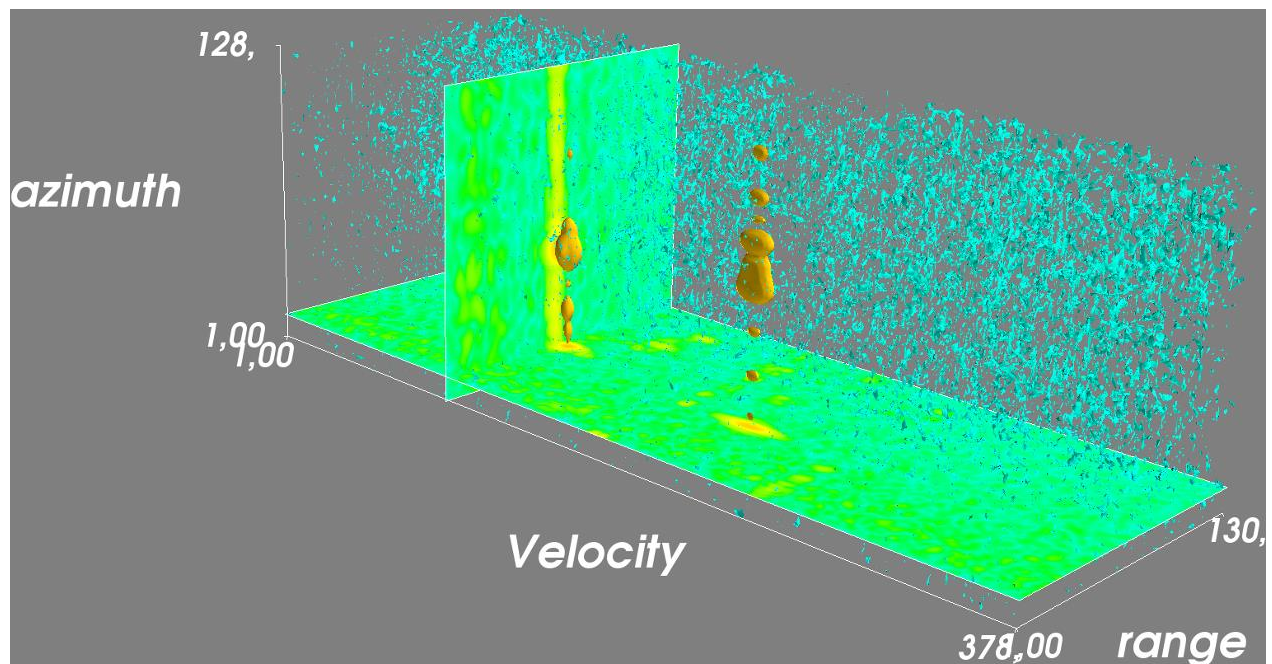


Abbildung 10 Darstellung des Radarwürfels durch wiederholte Fourier Transformationen

Es besteht jedoch die Vermutung, dass dennoch jede noch so kleine enthaltene Information positiv zum Lokalisierungsergebnis des neuronalen Netzes beitragen kann. Hierzu musste jedoch zunächst der Radarzweig des Netzwerks entsprechend der Eingabedaten angepasst werden. Insbesondere kommen im Folgenden nun dreidimensionale Faltungsmatrizen anstelle der bisher bei dreikanaligen Kamerafarbbildern bzw. graustufigen Range-Doppler Bildern üblichen zweidimensionalen Kernel zum Einsatz.

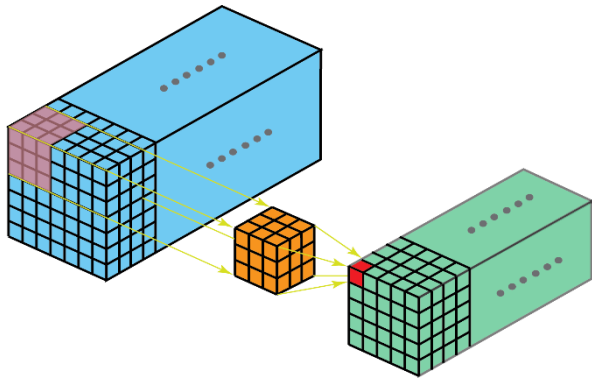


Abbildung 11 Dreidimensionale Faltungsoperationen auf räumlich ausgedehnten Daten

Auch mit Grafikkarten der neusten Generation (ca. 10 GB GPU Speicher) ist dieser Rechen- und Speicheraufwand nur durch diverse Kompromisse und Annahmen vertretbar zu bewältigen. Zunächst wurden sämtliche Operationen und Daten nur noch im half-precision Format abgelegt und innerhalb des Algorithmus verarbeitet. Während float32 noch 4 Byte Speicher pro Würfelzelle/Voxel belegt, kann mit dem float16 Format eine Speicherreduktion um die Hälfte erreicht werden. Dem dargelegten Sachverhalt zufolge besteht die begründete Annahme, dass durch das Hinzunehmen der Winkelinformation sowie das Zurückgreifen auf diese datengewaltigen jedoch informationsreichen strukturierten Radarrohdaten eine erhebliche Verbesserung bei der selbstüberwachten Lokalisierung von Objekten im Kamerabild erzielt werden kann.

Eine in den letzten Monaten weiterverfolgte Evaluierung des Ansatzes scheint die grundsätzliche Korrektheit der Annahmen zu großen Teilen zu bestätigen, da das Netzwerk im Falle korrespondierender Signale aus Kamera und Radarsensoren (grüne Kennzeichnung oben links in Abbildung 12) eine systematische Lokalisierung reflektierender Objekte vornimmt. Hierbei ist sich noch einmal vor Augen zu führen, dass dem neuronalen Netz weder temporale Informationen über einen entsprechenden zeitlichen Horizont während des Trainingsprozesses zur Verfügung gestellt wurden, noch jemals eine semantische Annotation präsentiert worden ist. Vor diesem Hintergrund sind die erzielten Ergebnisse durchaus beachtlich.

Ein regelrechter Mehrwert ist mit diesem Ansatz bei einem Szenario zu erkennen, bei dem ein gestörtes oder nur anderweitig unzureichendes Kamerabild (bspw. gestört durch Schnee, Nebel,

usw.) zur Verfügung steht, da auch in diesem Fall eine zuverlässige Erkennung von eventuellen Gefahrenquellen für nachgelagerte autonome Fahrfunktionen zur Verfügung stünde.

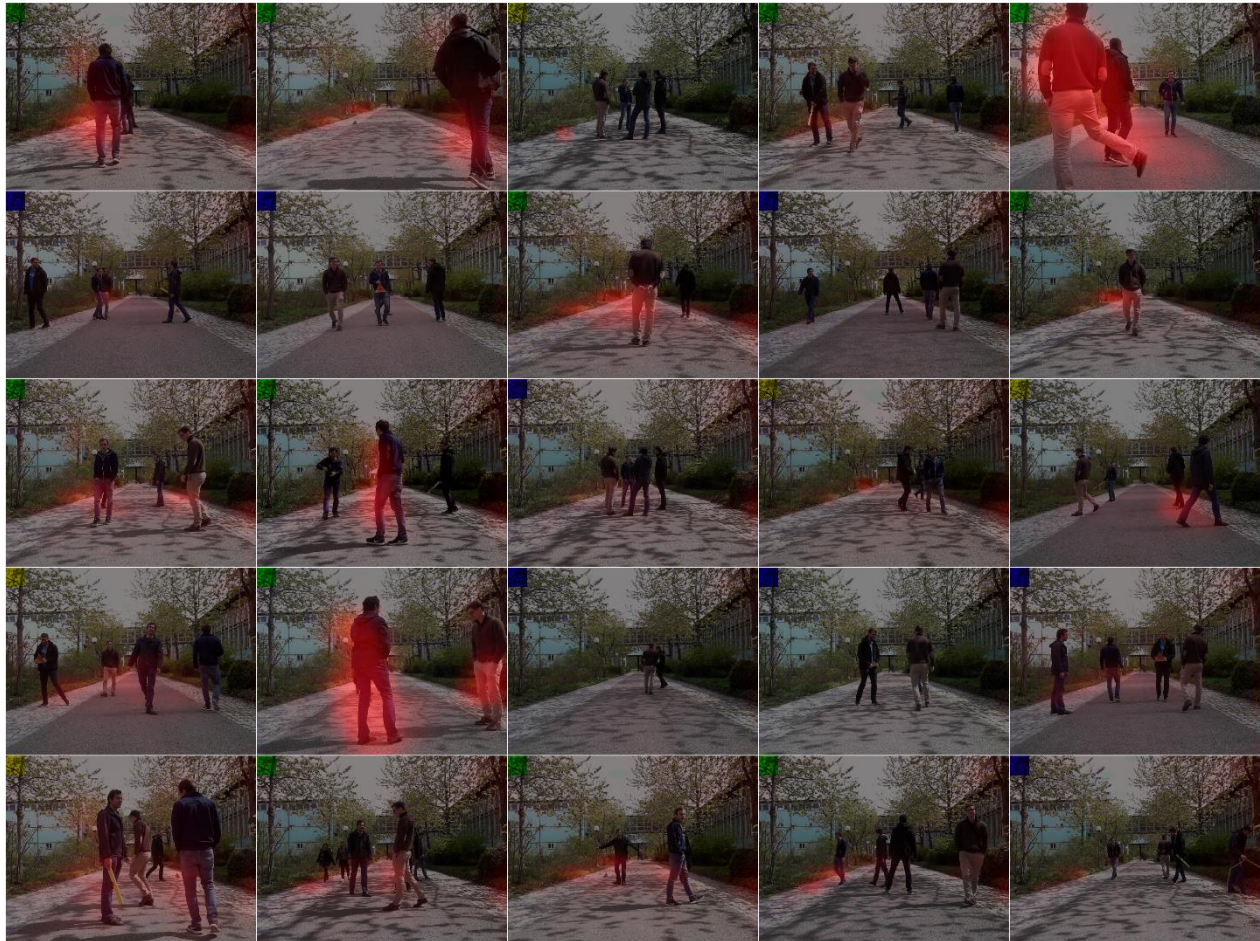


Abbildung 12 Lokalisierungsergebnisse elektromagnetische Wellen reflektierender Objekte im Kamerabild

3. Veröffentlichungen

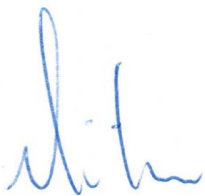
Im Rahmen des Projekts erfolgten folgende Veröffentlichungen.

- Pfeuffer, Andreas and Dietmayer, Klaus: „Optimal Sensor Data Fusion Architecture for Object Detection in Adverse Weather Conditions“; 21st International Conference on Information Fusion (FUSION 2018), Seite 2592 - 2599. Cambridge, United Kingdom (Great Britain); Juli 2018
- Pfeuffer, Andreas and Schulz Karina, and Dietmayer, Klaus: “Semantic Segmentation of Video Sequences with Convolutional LSTMs”; 2019 IEEE Intelligent Vehicles Symposium (IV), Seite 1441-1447; Juni 2019
- Pfeuffer, Andreas and Dietmayer, Klaus: “Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion”, 22st International Conference on Information Fusion (FUSION 2029). Ottawa, Canada; Juli 2019
- Pfeuffer, Andreas and Dietmayer, Klaus: “Separable Convolutional LSTMs for Faster Video Segmentation”; 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Seite 1072-1078; Oktober 2019

4. Verwertung

Die Projektergebnisse fließen unmittelbar in weitere Forschungsprojekte ein, da die Frage der robusten Umgebungserfassung unter allen Witterungsbedingungen noch nicht umfassend gelöst werden konnte aber für die flächendeckende Einführung automatisierter Fahrzeuge eine Schlüsselfrage sein wird. Die erzielten Ansätze im Projekt unter Nutzung künstlicher neuronaler Netze sind aber sehr vielversprechend. Ferner fließen die Ergebnisse unmittelbar in die universitäre Lehre, d.h. Studierendenausbildung sowie Promotionsverfahren ein, so dass ein Wissenstransfer über Köpfe auch in die Industrie erfolgt.

Ulm, den 16.09.2020



Prof. Dr.-Ing. Klaus Dietmayer

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel Abschlussbericht Verbundprojekt DENSE Teilvorhaben: DENSE-Signalverbesserung	
4. Autor(en) [Name(n), Vorname(n)] Ditzel, Carsten Pfeuffer, Andreas Dietmayer, Klaus	5. Abschlussdatum des Vorhabens 31.05.2018
	6. Veröffentlichungsdatum 30.11.2018
	7. Form der Publikation Bericht
8. Durchführende Institution(en) (Name, Adresse) Institut für Mess-, Regel- und Mikrotechnik Universität Ulm Albert-Einstein-Allee 41 89081 Ulm	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 16ESE0108
	11. Seitenzahl 19
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 4
	14. Tabellen 6
	15. Abbildungen 12
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung Die übergeordnete Zielsetzung im Projekt war eine verbesserte Umfelderkennung automatisierter Fahrzeuge bei unvorteilhaften äußeren Wetterbedingungen durch intelligente Fusion von Datenströmen mehrerer Sensortypen. Die drei zu diesem Zweck heutzutage am häufigsten verwendeten Messgeber sind Kamera, Lidar- und Radarsensoren, die alle ihre eigenen spezifischen Eigenschaften aufweisen und deren jeweiligen Vorteile erst durch ihre Kombination vollumfänglich hervortreten. Aktive Sensoren, das bedeutet, Lidar und Radar haben gerade hinsichtlich kompromittierter Witterungsbedingungen Vorteile in Robustheit und Datenerfassung gegenüber reinen Kamerasystemen, die wiederum als passive Sensorik eine unübertroffene Informationsdichte aufweisen können. Im beschriebenen Projekt wurden daher sowohl die Daten von Kamera- und Lidarsensoren als auch von Kamera- und Radarsensorik algorithmisch und auf verschiedene Arten miteinander in Beziehung gesetzt. Die hierzu verwendeten künstlichen neuronalen Netze sind mittlerweile in vielen technischen Bereichen sowohl in Forschung als auch Anwendung voll etabliert und liefern regelmäßig in unterschiedlichsten Ausprägungen herausragende Ergebnisse für eine Vielzahl von Aufgaben. Die Architektur der in Anlehnung an menschliche Neuronenstrukturen entwickelten Netze ist einer der Hauptschwerpunkte gegenwärtiger Forschung auf diesem Gebiet weshalb ihr folgerichtig auch innerhalb des Projekts ein besonderer Stellenwert zukam. Insbesondere ging es um die zentrale Frage, an welcher Stelle und auf welche Weise innerhalb der Struktur die Daten zusammengeführt und miteinander kombiniert werden sollten.	
19. Schlagwörter Automatisiertes Fahren, Robuste Sensorik, Schlechte Witterung, Sensorfusion	
20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final report
3. title Final report of the project: DENSE Subproject: DENSE-Signalverbesserung	
4. author(s) (family name, first name(s)) Ditzel, Carsten Pfeuffer, Andreas Dietmayer, Klaus	5. end of project February, 29th, 2020
	6. publication date August, 31st 2020
	7. form of publication report
8. performing organization(s) (name, address) Institute for Measurement, Control and Microtechnology Ulm University Albert-Einstein-Allee 41 89081 Ulm Germany	9. originator's report no.
	10. reference no. 16ESE0108
	11. no. of pages 19
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references 4
	14. no. of tables 6
	15. no. of figures 12
16. supplementary notes	
17. presented at (title, place, date)	
18. abstract The objective of the project was to improve the environment detection of automated vehicles at adverse weather conditions by intelligent fusion of data streams of several sensor types. The three sensors most commonly used for this purpose today are camera, lidar and radar sensors, each of which has its own specific characteristics and whose respective advantages only become fully apparent when they are combined. Active sensors, i.e. lidar and radar have advantages in robustness and data acquisition, especially with regard to adverse weather conditions, compared to pure camera systems, which in turn as passive sensors can provide an unsurpassed information density. In the project, therefore, the data of camera and lidar sensors as well as of camera and radar sensors were algorithmically and in different ways related to each other. The artificial neural networks used for this purpose are now fully established in many technical areas, both in research and application, and regularly deliver outstanding results in a wide variety of forms for a multitude of tasks. The architecture of the networks, which are based on human neuron structures, is one of the main focuses of current research in this field. In particular, the central question of where and how the data should be merged and combined within the structure was addressed.	
19. keywords Environment Perception, Autonomous Driving, Adverse Weather Conditions, Sensor Fusion.	
20. publisher	21. price