Personalisierte, adaptive kooperative Systeme für automatisierte Fahrzeuge

Schlussbericht

Zuwendungsempfänger:

Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung

Förderkennzeichen:

16SV7678

Gefördert durch:

Bundesministerium für Bildung und Forschung

Betreut vom Projektträger

VDI|VDE|IT

Autoren: Dr.-Ing. Michael Voit & Manuel Martin, Fraunhofer IOSB

1 Inhalt

2 Kurzdarstellung

2.1 Aufgabenstellung

Automatisierte Fahrzeuge werden den Fahrer zukünftig nicht nur entlasten, sondern streckenweise sogar ganz von der Fahrverantwortung befreien. Dadurch entstehen vielschichtige, neue Schnittstellen zwischen Fahrer und Fahrzeug. Beispielsweise kann beim automatisierten Fahren auf der Autobahn vorgesehen sein, die Kontrolle an den Fahrer zurückzugeben, wenn die Autobahn verlassen werden soll.

Im Vorfeld der Skizzenerstellung wurden in Rücksprache mit OEMs, CarSharing-Firmen und Firmen, die große Fahrzeugpools betreiben, Szenarien identifiziert, in denen eine zukünftige Kooperation zwischen Fahrzeug und Fahrer angedacht werden kann bzw. sogar muss. Dabei wurden insbesondere drei Nutzergruppen identifiziert, die hinsichtlich ihrer Akzeptanz für die zukünftige hochautomatisierte Mobilität besondere Herausforderungen stellen. Zum einen betrifft das die „Wenig-Fahrer“, die insbesondere auf CarSharing-Angebote zurückgreifen und selbst häufig kein eigenes Fahrzeug besitzen. Die „Viel-Fahrer“, die aufgrund beruflicher Bedingungen häufig auch in verschiedenen Fahrzeugen unterwegs sind. Und die „Normal-Fahrer“, die ihr Fahrzeug mit anderen, mehreren Nutzern ggf. teilen (z.B., weil mehrere Personen im Haushalt leben). Alle drei Nutzergruppen stellen verschiedene Anforderungen an hochautomatisierte Fahrzeuge. In Folge dessen ist absehbar, dass eine Assistenz im Fahrzeuginnenraum personalisiert umgesetzt werden muss, um schlussendlich für eine hohe Akzeptanz zu sorgen.

Ziel dieses Verbundvorhabens war daher die erstmalige Entwicklung und Umsetzung eines personalisierten Kooperationsmanagers, der die Interaktion zwischen Menschen und automatisiertem Fahrzeug optimiert und ein planbares Verhalten in ausgewählten Situationen unterstützen soll. Im Vordergrund stand dabei die Entwicklung einer nicht-generischen Fahrzeug-Fahrer-Kooperation, die das Gefühl einer technischen Bevormundung verhindern und gleichzeitig die nötige Grundlage bilden sollte, um dem Fahrer mehr Freiheiten beispielsweise in der Durchführung von Nebentätigkeiten zu erlauben.

Hauptaufgabe des Fraunhofer IOSB war dabei die Koordination und Umsetzung einer Fahrzeuginnenraumbeobachtung zusammen mit den Projektpartnern, um damit die Grundlagen für ein kooperatives Übergabesystem entwickeln zu können. Das geplante System sollte in der Lage sein, verschiedene Nebentätigkeiten bei unterschiedlichem Grad der Automatisierung zu erkennen und darüber hinaus zu prädizieren, wie lange eine Übergabe an den Fahrer in der momentan vorherrschenden Situation voraussichtlich dauern wird.

Zur Umsetzung wurden 2D- und 3D-Ansätze mit entsprechenden Kameras entwickelt und integriert. Tätigkeiten wurden in Handlungsprimitive zerlegt und ihr Einfluss auf die Übernahmefähigkeit untersucht. Durch die automatische Erkennung dieser Primitive sollten auch unbekannte Handlungen in Teilen berücksichtigt werden. Dabei spielte insbesondere die personalisierte

Berücksichtigung jeweiliger Verhaltensweisen eine große Rolle. Aus diesem Grund wurde der Fahrer auch mittels Gesichtsidentifikation erkannt.

2.2 Voraussetzungen

Das Konsortium wurde mit der Absicht zusammengestellt, die wissenschaftliche Arbeiten des Fraunhofer-Instituts IOSB und den Instituten des KITs sowie der TU München zu bündeln und zur Integration und Umsetzung die in der Automobilbranche übliche Rollenverteilung der Geschäftsmodelle widerzuspiegeln.

2.2.1 Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB

Das Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB) unter der Leitung von Prof. Dr.-Ing. J. Beyerer begleitet mit seinen Kernkompetenzen der Bildgewinnung bis zur Bildauswertung die Prozesskette von der Grundlagenforschung bis zur Entwicklung. In der Abteilung ‚Interaktive Analyse und Diagnose (IAD)' werden hierzu innovative Interaktionsmethoden und Assistenzsysteme entwickelt, die Menschen bei ihren Aufgaben unterstützen. Die Anwendungsfelder reichen dabei von der aufklärenden Szenenanalyse über Fabrikprozesse in der industriellen Produktion bis zu Fahrerassistenzsystemen. Das Hauptaugenmerk liegt auf der Entwicklung videobasierter Perzeptions- und Interaktionstechniken, insbesondere auf der Erfassung von Personen und der Auswertung ihrer Aktivitäten sowie der Integration und Entwicklung multimodaler Interaktionstechniken in intelligenten und proaktiven Umgebungen. Das IOSB beschäftigt sich seit mehr als zehn Jahre mit dem Forschungsthema intelligenter Umgebungen, in denen Menschen videobasiert, aber auch akustisch wahrgenommen und durch Assistenzsysteme proaktiv unterstützt werden. Nach der Beteiligung am EU-Projekt CHIL (Computers in the Human Interaction Loop) bündelte sich die erlangte Kompetenz 2008 zunächst durch die Gründung des Themenfeldes „Maschinensehen für die Mensch-Maschine-Interaktion", welches 2012 schließlich in die Forschungsgruppe „Perceptual User Interfaces" überführt wurde. Große Teile der Forschungsarbeiten dieser Gruppe konzentrieren sich inzwischen auf Assistenzsysteme im Fahrzeuginnenraum, die mit erlangtem Kontextwissen über die Aktivitäten aller Insassen angereichert werden. Die Aktivitäten in diesem Forschungsvorhaben wurden maßgeblich durch die vorige erfolgreiche Teilnahme am BMBF-geförderten Projekt InCarIn [InCarIn.2014] motiviert, in dem das IOSB den Löwenanteil der Personen- und Innenraumerfassung für alle Insassen eines Versuchsfahrzeugs leitete, entwickelte und implementierte. Damit stellte das IOSB die idealen Voraussetzungen, um mit seiner erworbenen und einzigartigen Kompetenz der Gesamtinnenraumanalyse den notwendigen Beitrag zur Fahrerzustands- und Aktivitätserkennung beisteuern zu können.

2.2.2 Bosch

Die Robert Bosch GmbH hat sich als großes Ziel für die Zukunft sicheres, agiles und automatisiertes Fahren gesetzt. In diesem Zusammenhang wird seit langer Zeit mit Hochdruck am automatisierten Fahren gearbeitet. Das Unternehmen war bisher an der Funktionsentwicklung für die bereits verfügbaren Automatisierungsgrade stark beteiligt und liefert für diese entsprechende Sensoren. Die Robert Bosch GmbH erkennt die Relevanz der Fahrerzustandsbeobachtung und der Interaktion zwischen Fahrer und Fahrzeug im Zusammenhang mit dem automatisierten Fahren und arbeitet in verschieden Projekten zur Umsetzung der dort erforschten Lösungsansätze mit. Als Partner brachte Bosch daher insbesondere beim Aufbau von Versuchsfahrzeugen und der Integration von Funktionen seine langjährigen Erfahrungen und Kompetenzen ein und wirkte als Enabler im Verbund mit.

2.2.3 mVise

Die mVISE AG verbindet IT Know-how mit mobilen Business- und Consumer-Lösungen. Gestartet in 2000 als Pionier für mobile Mehrwertdienste implementieren mVISE-Spezialisten inzwischen komplexe Anwendungen wie die Award-Winning Bosch Toolbox, Mobile Workforce Management Produkte auf Basis von flexiblen Cloud-Architekturen oder sichere Device Connectivity Apps zur mobilen Steuerung und Verwaltung von Geräten mithilfe von Smartphones und Tablets. mVISE brachte als Projektpartner seine langjährige Erfahrung in der nutzerzentrierten Umsetzung von komplexen mobilen Anwenderprozessen mit und hat sich im Projektvorhaben hauptsächlich um die Entwicklung der Applikationen auf Basis der erstellten Referenzarchitektur konzentriert.

2.2.4 FZI

Das FZI Forschungszentrum Informatik am Karlsruher Institut für Technologie ist eine gemeinnützige Forschungstransfereinrichtung des Landes Baden-Württemberg. Auftrag des FZI ist es, Unternehmen und öffentlichen Einrichtungen dabei zu helfen, neueste Methoden und Erkenntnisse wissenschaftlicher Forschung aus Informatik, Ingenieurwissenschaften und Betriebswirtschaft in wirtschaftlichen Erfolg umzusetzen. Das FZI entwickelt für seine Geschäftspartner Organisationslösungen, Softwarelösungen und Systemlösungen für innovative Produkte, Dienstleistungen und Geschäftsprozesse. Rechtlich selbständig, ist das FZI personell eng mit der Universität verflochten. Professorinnen und Professoren des Karlsruher Instituts für Technologie (KIT) engagieren sich am FZI aktiv für den Technologietransfer. Ihnen steht ein junges Team hochqualifizierter Mitarbeiterinnen und Mitarbeiter zur Seite. Die Forscher sind Absolventen der Universität Karlsruhe oder befreundeter Partnerhochschulen. Diese enge Verbindung mit der Wissenschaft stellt sicher, dass neueste Forschungserfolge schnell in innovative Lösungen für Unternehmen fließen. Seit seiner Gründung im Jahr 1984 hat das FZI seine große Kompetenz in den verschiedensten Anwendungsbereichen in zahlreichen Projekten eindrucksvoll bewiesen. Heute beläuft sich der Anteil der kleinen und mittelständischen Unternehmen an den

Wirtschaftserträgen der Transfereinrichtung auf 70 Prozent. Im Umfeld Datenschutz, Informationsrecht für technische Systeme und Rechtsinformatik befasst sich das FZI mit der juristischen Bewertung komplexer IKT-Infrastrukturen und diese Kompetenz war es auch, die das FZI in diesem Vorhaben eingebracht hat. Die Schwerpunkte lagen dabei auf der Datensicherheit und dem Datenschutz.

2.2.5 Videmo

Seit seiner Gründung im Jahr 2008 entwickelt und vertreibt Videmo Gesichtserkennungssoftware. Im Vordergrund steht dabei die forschungsnahe Pflege der eigenentwickelten Verfahren zur Gesichtsidentifikation sowie zur Erkennung von Alter und Geschlecht. Bisherige Hauptmärkte sind der Sicherheitsbereich (Zugangskontrolle, Videoforensik) sowie die digitale Marktforschung (Kundenzählung, adaptive Werbung). In PAKoS steuerte Videmo seine Kerntechnologie bei und entwickelte diese unter den gegebenen Bedingungen im Fahrzeuginnenraum zu einer robusten Personenidentifikations- sowie Alters- und Geschlechtserkennung weiter. Damit leistete Videmo den Hauptbeitrag zur Personalisierbarkeit der Fahrzeug-Mensch-Kooperation, weil durch die Identifikation der jeweiligen Nutzer personalisierte Nutzerprofile erstellt und geladen werden können.

2.2.6 Lehrstuhl für Ergonomie (LfE), TU München

Der Lehrstuhl für Ergonomie (LfE) beschäftigt sich seit vielen Jahren mit der Entwicklung und Auslegung unterstützender Fahrerassistenzsysteme. Neben Industrieprojekten zum Thema „automatisiertes Fahren“ wurden auch haptisch-multimodale Interaktionsgestaltungen und andere Zustandsvermittlungsmethoden für die wechselseitige Kontrolle zwischen Mensch und Maschine in hochautomatisierten Fahrzeugen untersucht. Die Fragestellung, wie mit einer multimodalen Interaktionsgestaltung der Nutzerzustand in der Kommunikation zwischen Fahrer und Fahrzeug berücksichtigt werden kann, ist dabei vordergründiges Thema. Aus diesem Grund bildete das LFE die ideale Schnittstelle, um Prognosen der Nutzerablenkung vor dem Hintergrund einer Übernahmegestaltung beisteuern und integrieren zu können.

2.2.7 Karlsruher Institut für Technologie KIT

Das KIT brachte sich mit drei Lehrstühlen / Instituten in das Forschungsvorhaben ein:

1. Das **Computer Vision for Human-Computer Interaction Lab (CVHCI)** des KIT forscht seit vielen Jahren an Verfahren zur videobasierten Wahrnehmung von Menschen mit dem Ziel, die Interaktion mit technischen Systemen nutzerfreundlicher zu gestalten. Durch zahlreiche Forschungsarbeiten wurde dabei ein großes Spektrum an verschiedenen Aspekten der Menschwahrnehmung abgedeckt, das von der Verfolgung von Personen, Gestenerkennung, Personenerkennung, Analyse von Gesichtern, einschließlich der Erkennung von Alter, Geschlecht, Blick und Gesichtsausdrücken bis hin zum Erkennen von

Handlungen reicht. Diese Aspekte wurden im Rahmen verschiedener Anwendungsgebiete wie der Forschung an interaktiven Robotern, intelligenten Umgebungen, computerunterstützter Auswertung von Bildern und Bildfolgen und assistierender Systeme für Menschen mit besonderen Bedürfnissen entwickelt. In den letzten Jahren wurden diese Arbeiten in Kooperation mit dem IOSB auch zur Fahrererfassung weiterentwickelt (Erkennung von Körperhaltung, Gesten und Kopfdrehung). Vor dem Hintergrund dieser Kooperation brachte sich das CVHCI unterstützend in die Forschungsaktivitäten des IOSB mit ein, legte seinen Fokus allerdings auf die Erkennung der Handlungsprimitive bei der Akvititätserkennung sowie die personalisierte Nutzererfassung in Form der Gesichtserkennung, ergänzt um eine Alters- und Geschlechtswahrnehmung.

2. Das **Institut für Regelungs- und Steuerungssysteme (IRS)** erforscht neue Konzepte zur Regelungs- und Steuerungstechnik in Energiesystemen, Systemen mit Garantien und kooperativen Systemen. In letzterem – für dieses Vorhaben maßgeblichem – Feld wird seit einigen Jahren an einem Entwurf von Regelungskonzepten für die Kooperation von Menschen und Maschine gearbeitet. Unter dem Begriff „shared control“ werden dabei Regelungskonzepte untersucht, bei denen sowohl ein technischer Regler als auch der Mensch im Regelkreis agieren. Aufbauend auf den Grundlagenarbeiten am Lehrstuhl wurden in zahlreichen Industrie- und MWK geförderten Projekten neue Ansätze entwickelt. Mithilfe eines modellprädiktiven Regelungsansatzes (MPC) wurde dabei bereits in der Vergangenheit die Fahrzeugführungsaufgabe so bewältigt, dass die Regelungseingriffe durch das Entwurfsverfahren garantiert unterhalb der Wahrnehmungsschwelle des Fahrers blieben. Das Fahrzeug kooperierte somit im Verborgenen. Auf diesem Ansatz wurde in PAKoS aufgesetzt und durch das IRS die Frage bearbeitet, wie die Regelung des Fahrzeugs in Kooperation mit dem Fahrer unter verschiedenen Situationen umgesetzt und integriert werden kann.
3. Das **Institut für Technikfolgenabschätzung und Systemanalyse (ITAS)** erforscht wissenschaftliche und technische Entwicklungen in Bezug auf systemische Zusammenhänge und Technikfolgen. Es erarbeitet und vermittelt Wissen und Bewertungen und entwirft Handlungs- und Gestaltungsoptionen. Im Mittelpunkt der Forschung stehen ethische, ökologische, ökonomische, soziale, politisch-institutionelle und kulturelle Fragestellungen. Wesentliche Ziele sind die Beratung der Forschungs- und Technikpolitik, die Bereitstellung von Orientierungswissen zur Gestaltung sozio-technischer Systeme sowie die Durchführung diskursiver Verfahren zu offenen oder kontroversen technologiepolitischen Fragen. Die Ergebnisse der Forschung und Beratung sind öffentlich. Zu den Themenbereichen „Mensch-Technik-Interaktionen“ hat das ITAS bereits mehrere Studien durchgeführt. Auch methodisch bringt das ITAS viele Erfahrungen zur Bearbeitung von ELSI-Fragestellungen mit ein, hier sind unter anderem Reflexionsverfahren in

normativer Hinsicht wie die Angewandte Ethik zu nennen. Mit dieser umfangreichen Expertise zeigte sich das ITAS maßgeblich in der Technikfolgenabschätzung der entwickelten und integrierten Technologien verantwortlich, begleitete den Verbund aber darüber hinaus durchweg als Ansprechpartner zu allen ELSI-Fragestellungen.

2.2.8 Spiegel Institut

Das Spiegel Institut ist ein Marktforschungs- und Beratungsinstitut mit einem weltweiten Netz an Partnerinstituten und gilt als das Gründungsinstitut der Marktpsychologie in Deutschland. Das Institut hat einen starken Automobil- und Automotivefokus und unterstützt seit Jahren Automobilhersteller und -zulieferer bei der nutzerzentrierten Produktentwicklung von Fahrzeugen, Systemen und Komponenten. Die Forschungsschwerpunkte des Spiegel Instituts liegen sowohl in der qualitativen Anforderungsanalyse (bspw. mit ethnographischen Interviews) als auch in der empirischen Konzept- und Produktabsicherung (bspw. mit Simulator- und Realfahrtstudien). In PAKoS hat das Spiegel Institut daher seine langjährige Expertise aus zahlreichen Realfahrzeug- und Simulatorstudien eingebracht, um die erarbeiteten Konzeptansätze in Fahrversuchen mit den Realfahrzeugen zu evaluieren und hinsichtlich ihrer Wirksamkeit und Akzeptanz zu bewerten.

2.2.9 BIG

Die b.i.g.-Gruppe ist Spezialist für das Beraten, Planen und Betreiben von Immobilien. Mit seinen rund 3.000 Beschäftigten und 27 Gesellschaften bietet das Familienunternehmen national wie international an 24 Standorten über 100 Dienstleistungen an. Von Bernd und Gisela Bechtold 1981 als Ingenieurbüro gegründet, hat mit Daniela Bechtold seit 2013 die nächste Generation die Führung des Unternehmens übernommen. Die b.i.g.-Gruppe belegt die Geschäftsfelder Ingenieurplanung, Facility-Management, Gebäudeservice, Sicherheits-dienstleistungen, sowie komplexe Dienstleistungen. Damit brachte sich die b.i.g.-Gruppe als Anwender in das Projekt ein, der eine große Fahrzeugflotte mit vielen verschiedenen Nutzern betreibt.

2.2.10 BMW

Die BMW Group ist einer der weltweit führenden Fahrzeughersteller des Premiumsegments mit über 68 Milliarden Euro Umsatz sowie über 1,6 Millionen verkaufter Fahrzeuge weltweit. Im Bereich Forschung, neue Technologien und Innovationen entwickelt das Team Mensch-Maschine-Interaktion Konzepte und Forschungsmethoden zur Interaktion in zukünftigen Fahrzeugen. Neben der Untersuchung von Anwenderbedürfnissen und Rahmenbedingungen im Hinblick auf neue Technologien sowie Paradigmen steht dabei auch die Entwicklung von Konzepten zur Unterstützung holistischer Reise- und Fahrerlebnisse im Vordergrund. Im Forschungsvorhaben PAKoS stellte BMW daher seine verschiedensten Werkstatteinrichtungen sowie Laboreinrichtungen zur Verfügung. Dies beinhaltet technische Gerätschaften zur Entwicklung und Untersuchung von Interaktionskonzepten mittels Blickdatenmessung, Gestensteuerung,

Augmented Reality und 3D-Darstellungen. Als Anwender lag es für BMW besonders im Fokus, praxisnahe und vorseriennahe Entwicklungen mit zu koordinieren und insbesondere aus Kunden- und Nutzersicht das Vorhaben zu steuern und zu koordinieren. BMW brachte sich damit maßgeblich in der Anforderungsanalyse und Beschreibung und Definitionen der Transitionen mit ein.

2.3 Planung und Ablauf des Vorhabens

Kern des Forschungsvorhabens bildeten dabei die wissenschaftlichen Partner KIT und Fraunhofer IOSB, die jeweils die Übergaberegelung, die im Kooperationsmanager verankert ist, und die Innenraum- und Nutzerzustandserfassung bzw. Aktivitätsklassifikation umsetzten. Vervollständigt wurden diese Inhalte durch die TU München, deren Schwerpunkt in der Prognose der Übernahmebereitschaft durch den Fahrer lag. Bosch Abstatt und BMW München zeigten sich maßgeblich in der Integration und Bereitstellung des Testfahrzeugs und Bus-Anbindung verantwortlich und agierten als Enabler und Anwender im Projektvorhaben. Das Spiegel Institut Mannheim übernahm und begleitet sämtliche Nutzerstudien und Untersuchungen zur User Experience. Letztere wurde dabei maßgeblich durch die Umsetzungen der mVise AG in Düsseldorf geprägt. Die Videmo GmbH & Co. KG aus Karlsruhe brachte ihr Kernprodukt, die Gesichtsidentifikation, ein und entwickelte die Algorithmik sukzessive an die Bedingungen im Fahrzeuginnenraum weiter, um eine robuste Personalisierung und Nutzerprofilbildung bereitstellen zu können. In Kooperation mit der BIG-Gruppe, die jedoch vorzeitig den Projektverbund verlassen musste, sollte die Identifikation zur Assistenz von Sicherheitsfahrern eingesetzt und untersucht werden. Das Forschungszentrum für Informatik (FZI) und das KIT-ITAS aus Karlsruhe begleiteten das Projektvorhaben durchgängig mit Fragen zum Datenschutz, der Datensicherheit und ELSI-Fragestellungen und trugen dabei einen grundlegenden Anteil an der Umsetzbarkeit der Personalisierung bei.

Die Bearbeitung des Vorhabens wurde auf 9 inhaltliche und 3 organisatorische und weiterführende Arbeitspakete aufgeteilt. AP 1 diente hierbei dazu, die Anforderungen an einen Übergabe-Kooperationsmanager, wie er im Projekt entwickelt werden sollte, zu erheben, konkrete Übergabeszenarien, die betrachtet werden sollen, zu definieren und einhergehende Transitionen gemeinschaftlich festzulegen. AP 2 bündelte hierzu notwendige Datengenerierungen, die als Grundlage zur Nutzererfassung in AP 3 dienten und insbesondere das Trainingsmaterial für das Einlernen der maschinellen Lernverfahren zur Fahrerzustandsbeobachtung bereitstellten. Zur Personalisierung wurden die Gesichtserkennung und Modellbildung der jeweiligen Fahrerprofile in AP 4 zusammengefasst. Die Adaption im Kooperationsmanager auf die jeweiligen Nutzerprofile wurde in AP 5 koordiniert und umgesetzt. Um im Dialog mit dem Fahrer bzw. Nutzer eine geplante Kooperation durchführen zu können, setzte sich AP 6 die Gestaltung der Interaktions- und Nutzerschnittstelle zum Ziel. Dabei wurde insbesondere zwischen zwei Transitionsphasen unterschieden: (1) der vorbereitenden Phase, wenn der Fahrer auf die kommende Kooperation

aufmerksam gemacht werden muss und (2) die Unterstützung des Fahrers bei seiner tatsächlichen Fahraufgabe als menschlicher Regler. Als das Arbeitspaket, das die Zusammenführung der zuvor genannten Inhalte steuert, wurde in AP 7 die Systemarchitektur des Kooperationsmanagers umgesetzt. Die Nutzerzustandserkennung, die Personalisierung, die Adaption und die Interaktionskonzepte wurden hier zu einem gemeinsamen Assistenzsystem verbunden und konzeptioniert. Die tatsächliche Integration in das Testfahrzeug geschah in AP 8. Und Nutzerstudien und abschließende Feldversuche schließlich in AP 9.

Über diese 9 inhaltlichen Arbeitspakete hinaus wurden in 3 weiteren die Technikfolgenabschätzung diskutiert (AP 10), das Projektmanagement koordiniert (AP 11) und die ELSI-Aspekte zum Datenschutz und der Datensicherheit gemeinschaftlich berücksichtigt (AP 12).

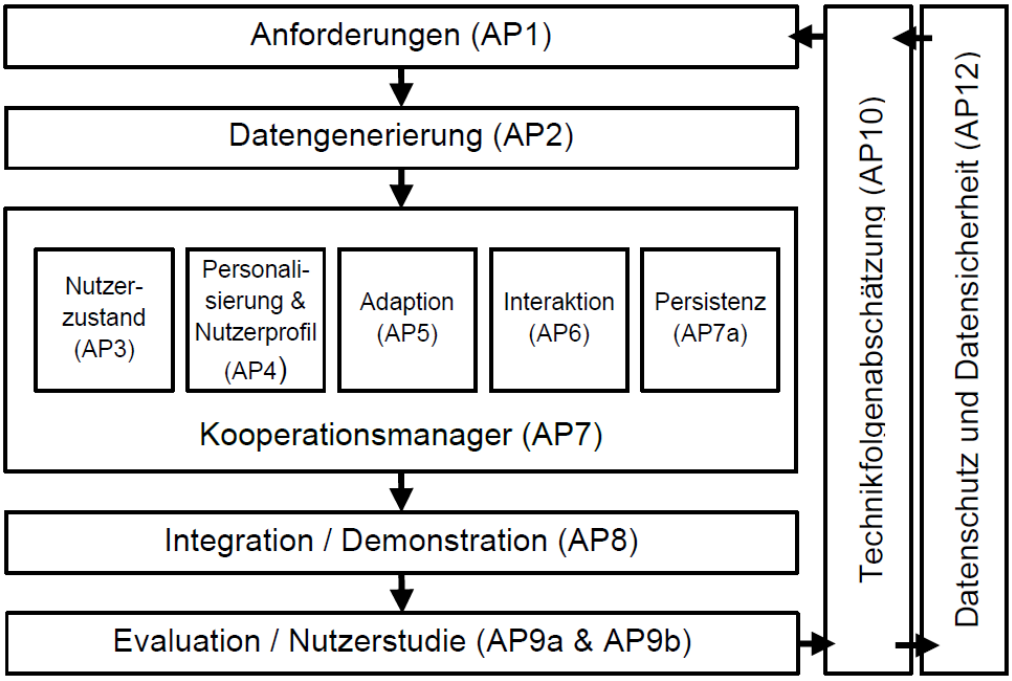


Abbildung 1: Die Arbeitspakete und das Zusammenwirken der Forschungsinhalte in PAKoS.

2.4 Wissenschaftlicher-technischer Stand zu Projektbeginn

Zu Projektbeginn zeigten die Ergebnisse aktueller Forschung, dass die Bestimmung der Kopfpose oder der Blickrichtung, sowie des visuellen Aufmerksamkeitsfokus des Fahrers bereits gut gelingen [Ji.2002; Fletcher.2005]. Durch die Erfassung der Hände konnte festgestellt werden, mit welchen Objekten im Fahrzeuginnenraum interagiert wird. Dies erlaubte sowohl Rückschlüsse auf die momentane Aufmerksamkeit als auch auf die Intention des Fahrers [Bach.2008; Pickering.2007; Ohn-Bar.2014]. Zudem konnte die gesamte Körperpose des Fahrers [Demirdjian.2009; Tran.2009] aufgezeichnet werden, um damit den momentanen Zustand [Holte.2012] oder auch die Intention [Ito.2008] zu bestimmen. Systeme zum Erkennen der Kopfpose und Blickrichtung im Auto waren schon seit einiger Zeit kommerziell verfügbar. Verfahren zur Körperposenerfassung im Auto wurden

allerdings erst vereinzelt wissenschaftlich erforscht [Demirdjian.2009; Tran.2009]. Im BMBF-geförderten Vorläuferprojekt InCarIn [InCarIn.2014], in dem auch Projektpartner dieses Projektes beteiligt waren, wurde erstmalig die Körperposenerfassung aller Fahrzeuginsassen umgesetzt, um auch Interaktionen zwischen Insassen zu erfassen. Eine umfassende Innenraumzustandserkennung ist für antizipierende Assistenzfunktionen, wie sie in InCarIn betrachtet wurden, notwendig. Eine solche Innenraumzustandserkennung ist jedoch auch erforderlich, um die vielfältigen Nebentätigkeiten während des automatisierten Fahrens zu erfassen und hierdurch eine zuverlässige Aussage über den Zustand und die Ablenkung des Fahrers zu ermitteln. Zum damaligen Stand der Technik wurde der Fahrer oftmals nur während der manuellen Fahrt modelliert. Typischerweise wurden Informationen über das Befinden des Fahrers, wie Müdigkeit, Abgelenktheit oder Überforderung, abgeleitet [Arun.2012; Kaplan.2015]. Im Gegensatz zur manuellen Fahrt, in der die Aktivitäten des Fahrers hauptsächlich auf die Fahraufgabe beschränkt sind, ist im Falle automatisierten Fahrens zu erwarten, dass der Fahrer einer Vielzahl unterschiedlicher Aktivitäten nachgeht. Im Übergabefall führt dies zu unterschiedlich langen Übergabezeiten [Petermann-Stock.2013]. Wie diese Aktivitäten robust erkannt und die nötigen Übergabezeiten geschätzt werden können, war noch wenig erforscht und PAKoS hat hier einen entscheidenden Beitrag geleistet. Die Aktivitätserkennung ist ein sehr aktives Forschungsfeld. Bei solchen Systemen erfolgt die Einordnung von Aktionen in erlernte Aktionskategorien. Hierfür werden häufig bildbasierte Merkmale als Grundlage verwendet. Bekannte Verfahren setzten zum Beispiel Raum-Zeit-Deskriptoren und ein Bag-of-Words-Modell [Rybok.2011, Wang.2011]) ein. Aktivitäten werden nicht nur als Ganzes klassifiziert, es können auch in einem Zwischenschritt zunächst Attribute gelernt und damit dann verschiedene Aktivitäten beschrieben werden [Sawhney.2013]. Weitergehende Ansätze nutzten die Körperpose als Zwischenschritt zur Aktionserkennung [Vemulapalli.2013]. Vorhandene Systeme basierten jedoch nur auf einer Auswahl von wenigen Modalitäten oder Teilaktivitäten. Ein komplexeres Aktivitätserkennungssystem, welches Übergabezeiten und Leistungsparameter des „Fahrers“ während des automatisierten Fahrens vorhersagt und eine kooperative Übergabe zum manuellen Fahren ermöglicht, wurde vor PAKoS noch nicht erarbeitet.

2.5 Zusammenarbeit mit anderen Stellen

Das Projekt PAKoS hat viele Kontakte in Forschung und Industrie eröffnet. Dies spiegelte sich unter anderem in der Teilnehmerliste der Abschlusspräsentation wieder. Auf Konferenzen wurde das Thema einer breiten Fachöffentlichkeit mehrfach präsentiert. Fraunhofer nutzte die Zwischen- und Endergebnisse auch um neue Förderprojekte mit neuen Partnern und Partnern aus dem Projekt anzustoßen. Indirekt floss die Erfahrung aus dem Projekt zudem in Kooperationen mit Industriepartnern ein.

3 Ergebnisbericht

Das Fraunhofer IOSB begleitete die Entwicklung aller im Rahmen von PAKoS entstandenen Fahrererfassungssysteme. Viele Arbeiten fanden dabei in enger Kooperation mit dem KIT-CVHCI statt. Das Fraunhofer IOSB stellte dabei einen Großteil der notwendigen Infrastruktur, wie den Fahrsimulator, und kümmerte sich um die Konzeption und Integration der Kameras und Verfahren in den Simulator und den Versuchsträger. Des Weiteren entwickelte das Fraunhofer IOSB auch mehrere Verfahren zur Fahrererfassung. In Kooperation mit dem KIT-LFR wurde ein Konzept zur kooperativen Übergabe von automatisier Fahrt zu manueller Fahrt unter Berücksichtigung des Fahrerzustands entworfen und im Laufe des Projekts prototypisch umgesetzt. Die Hauptergebnisse des Projekts wurden in mehreren Veröffentlichungen sowohl national wie auch international präsentiert. Die Aktivitäten des IOSB decken dabei die komplette Prozesskette der Innenraumbeobachtung ab. Angefangen bei dem Entwurf eines geeigneten multisensoriellen Kamerasystems, dem Sammeln und Veröffentlichen eines großen Datensatzes im Simulator zur Entwicklung maschineller Lernverfahren für die Nebentätigkeitserfassung im Fahrzeug, darauf aufbauend mehrere Verfahren zur Fahrerbeobachtung und abschließend deren Integration in den Versuchsträger.
Im Folgenden wird der Beitrag des IOSB zum Projekt detailliert dargestellt.

3.1 Multi-Kamerasystem für die Erfassung des Fahrers

Die Erfassung des Fahrerzustands kann über unterschiedliche Sensorik stattfinden. In manuellen Fahrzeugen, insbesondere Lastwagen, werden bereits in Serie die Lenkeingaben überwacht, um so auf die Müdigkeit des Fahrers zu schließen. Diese Methode ist in einem automatisierten Fahrzeug allerdings nicht zweckdienlich sobald der Fahrer die Hände vom Lenkrad nehmen darf. Andere Ansätze nutzen Eyetracker, um auf Abgelenktheit bzw. Müdigkeit zu schließen, aber auch hier ergeben sich Schwierigkeiten bei automatisierter Fahrt, da viele Tätigkeiten die Hände und Objekte involvieren, die mit einem Eyetracker nicht erfasst werden können.
In PAKoS wurde deshalb untersucht wie mit weitwinkeligen Kameras die ganze vordere Fahrzeugkabine erfasst werden kann. Hierdurch sind alle Körperbewegungen des Fahrers sowie auch alle zur Interaktion verfügbaren Objekte und Innenraumelemente wahrnehmbar. Hauptziel der Innenraumerfassung von PAKoS war die Bestimmung von Nebentätigkeiten des Fahrers. Das Kamerasystem sollte dabei sowohl praxisnahe Funktionalität bieten, aber auch die Möglichkeit unterschiedliche Kameraperspektiven und Kameratypen zu evaluieren. Das Fraunhofer IOSB entwarf deshalb für die Aufzeichnung von Trainingsdaten einen komplexen Kameraaufbau aus insgesamt 6 Kameras. Dies ermöglichte im Laufe des Projekts eine Evaluation der erreichbaren Güte jeder Kameraperspektive sowie die Verwendung mehrerer Kameras, um 3D-Informationen zu rekonstruieren. Das vollständige Kamerasystem wäre zu aufwändig für eine Umsetzung in Serie.

Durch das Vorgehen im Projekt wurde aber bereits eine zweckdienliche Teilmenge Identifiziert, die dann am Ende auch im Versuchsträger von Bosch verbaut wurde.

Die optischen Anforderungen an eine Innenraumkamera sind durchaus herausfordernd. Sie muss robust gegenüber starken Beleuchtungsunterschieden sein – von strahlendem Sonnenschein bis zu absoluter Dunkelheit bei Nacht. Des Weiteren ist ein großer Öffnungswinkel bei geringer Verzeichnung notwendig, um trotz der Nähe der Kamera zum Fahrer eine ausreichende Abdeckung zu erreichen. Das Kamerasystem soll dabei möglichst klein sein, um platzsparend verbaut werden zu können. Gerade die Kameragröße kann aber bei einem Prototypensystem nicht vollständig minimiert werden. Um die Anforderungen möglichst gut zu erfüllen wurde im Projekt hauptsächlich auf Nahinfrarotkameras mit aktiver Beleuchtung für Nachtfahrten gesetzt. Zusätzlich wurde mit der Kinect V2 noch eine Tiefenkamera verbaut. Im Folgenden werden die Kameras und deren Position im Innenraum detailliert beschrieben.

3.1.1 Nahinfrarotkamerasystem

Es gibt insbesondere im Überwachungsbereich bereits viele Kameras im Nahinfrarotbereich mit aktiver Beleuchtung für den Nachtbetrieb. Leider sind diese Kameras häufig bereits mit einem festen Objektiv ausgestattet, das nicht weitwinkelig genug für die Anwendung im Fahrzeuginnenraum ist. Viele sind auch zu groß für einen Verbau im Fahrzeug und deshalb nicht geeignet. Das Fraunhofer IOSB entwarf deshalb auf Basis einer kleinen Industriekamera ein eigenes System für den prototypischen Einsatz.

Das Paket besteht aus einer NIR-Kamera (IDS UI-3241LE-NIR-GL) mit Bandpassfilter und 4mm S-Mount Objektiv und einem Infrarot LED-Ring (850nm). Alle Komponenten wurden ohne Gehäuse beschafft und in einem selbstentwickelten 3D-gedruckten Gehäuse zu einer möglichst kompakten Einheit zusammengefügt (siehe Abbildung 2). Die Verwendung aktiver infraroten Beleuchtungsquellen, insbesondere in Augennähe, ist kritisch. Entsprechend wurde die Augensicherheit des Gesamtsystems im Auftragslabor geprüft und in der hiesigen Ausprägung für unkritisch befunden.

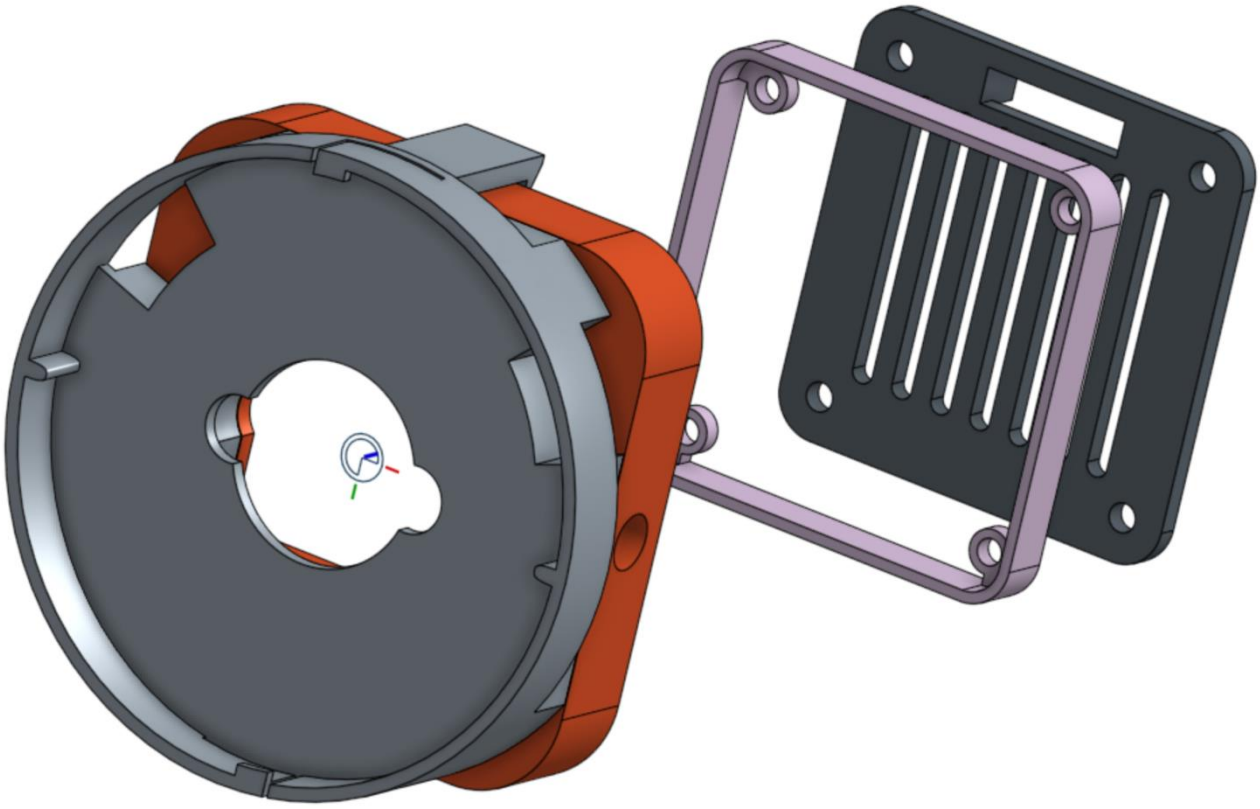


Abbildung 2: 3D-Modell des Gehäuses des NIR-Kamerasystems

3.1.2 Tiefenkamera

Tiefenkameras auf Basis des Time of Flight (ToF)-Prinzips oder des Structured Light-Prinzips nutzen auch aktive Beleuchtung für die Bildgebung und sind deshalb in Dunkelheit ähnlich geeignet wie Nahinfrarotkameras mit Aktivbeleuchtung. Sie liefern allerdings direkt 3D-Daten, was für einige Anwendungen von Vorteil ist. Entsprechend wurde auch dieser Sensortyp berücksichtigt. Viele Tiefenkameras haben allerdings Schwierigkeiten bei direktem Sonnenlicht, da dann die aktive Lichtquelle überstrahlt wird und keine Bildgebung möglich ist. Aufgrund der Erfahrung des IOSB mit solchen Kameras wurde trotz der Größe deshalb die Kinect V2 für den Aufbau ausgewählt. Ein weiterer Nachteil dieser Sensorik ist das oftmals fest verbaute Objektiv, was die möglichen Verbaupositionen im Fahrzeuginnenraum stark einschränkte. Zusätzlich zum Tiefenbild liefert die Kinect V2 ähnlich zu den Nahinfrarotkameras ein Infrarotbild. Dieses ist allerdings auf Grund des Messprinzips noch weit unempfindlicher für Störlichtquellen. Zuletzt hat die Kinect V2 auch noch eine Webcam verbaut, die ein Farbbild liefert. Dieses ist nicht gut geeignet für die Verwendung im Fahrzeuginnenraum, da zwangsweise eine Lichtquelle im sichtbaren Bereich notwendig ist, was nachts nicht der Fall ist. Da Farbbilder aber für das Training maschineller Lernverfahren üblich sind, wurde dieser Bildstrom trotzdem aufgezeichnet.

3.1.3 Positionierung im Fahrzeuginnenraum

Insgesamt wurden im Fahrzeuginneren für die Aufzeichnung von Trainingsdaten für die im Projekt entwickelten maschinellen Lernverfahre 6 Kameras verbaut. Davon 5 der selbstentwickelten

Infrarotsysteme und eine Kinect V2. Abbildung 3 zeigt eine Übersicht der Sensorpositionen im Fahrsimulator des Fraunhofer IOSB. Die Positionen wurden so gewählt, dass der vordere Teil der Kabine möglichst aus mehreren Perspektiven vollständig abgedeckt wird. Zusätzlich bietet die Kamera auf der Lenksäule einen detaillierten Blick auf das Gesicht des Fahrers. Diese Ansicht entspricht dem, was in vielen anderen Fahrerbeobachtungssystemen zur Bestimmung des Lidschluss, Kopforientierung oder Eyetracking genutzt wird. Letztlich wurde die Kinect V2 an der Beifahrer-A-Säule verbaut, da nur diese Position auf Grund des verbauten Objektivs einen ausreichenden Blick auf den Fahrer gewährt. Abbildung 3 zeigt die Verbaupositionen aller Sensoren im Simulator des Fraunhofer IOSB.

Mit diesem Kamerasystem können insgesamt 8 Bildströme (6 Nahinfrarot, 1 Tiefenbild, 1 Farbbild) aus 5 verschiedenen Perspektiven aufgezeichnet werden. Abbildung 4 zeigt Beispielaufnahmen jeder Ansicht.

Abbildung 3: Übersicht des im Fahrsimulator verbauten Gesamtsystems mit allen Kameras. NIR-Kameras in blau. Kinect V2-Tiefenkamera in rot.

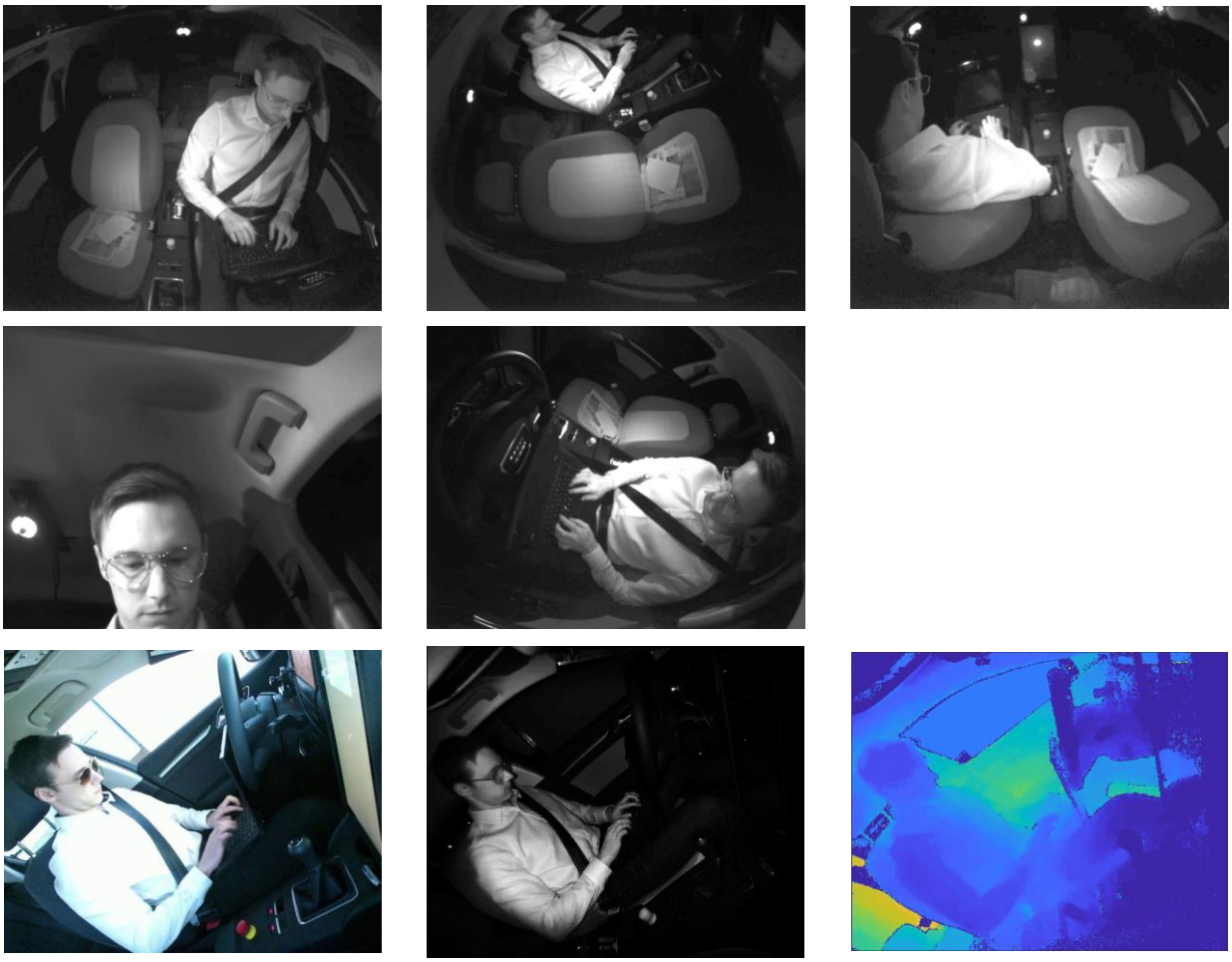


Abbildung 4: Beispielbilder aller Ansichten des Kamerasystems. NIR-Sensoren in Reihe eins und zwei. Kinect V2 Sensor in Reihe drei.

3.2 Aufzeichnung des Drive&Act-Datensatzes

Für die Entwicklung aktueller maschineller Lernverfahren sind große Datenmengen nötig. Für viele Anwendungen gibt es bereits entsprechende öffentliche Datenquellen. Dies war für die Bestimmung von Nebentätigkeiten des Fahrers nicht der Fall. In PAKoS wurde deshalb in einer großen Simulatorstudie eine entsprechende Datenbasis geschaffen und diese später auch auf einer internationalen Konferenz unter dem Namen „Drive&Act“ veröffentlicht. Die Planung und Durchführung fand dabei in enger Zusammenarbeit mit dem KIT-CVHCI statt.

Bestenfalls würde man die Daten in Realfahrten im Straßenverkehr sammeln, um mit den Kameras auch alle auftretenden Umwelteinflüsse aufzuzeichnen. Dies ist aber aus verschiedenen Gründen schwer umsetzbar. Hauptziel war die Sammlung von Nebentätigkeiten während automatisierter Fahrt. Diese können deutlich komplexer und ablenkender sein als bei momentan üblichen Nebentätigkeiten. Aktuell ist der Betrieb eines automatisierten Fahrzeugs aber nur unter Aufsicht durch einen Fahrer möglich, was den Fahrer als Probanden entsprechend ausschließen würde. Ein entsprechendes Experiment wäre nur mit einem Wizard of Oz-Fahrzeug möglich, das im Projekt allerdings nicht zur Verfügung stand. Der Vorteil einer Simulatorstudie ist allerdings die gute

Kontrollierbarkeit. Es treten keine unvorhergesehenen Ereignisse auf und es ist leichter, die gewünschten Nebentätigkeiten in kurzer Zeit hervorzurufen. Hauptnachteil ist die unrealistische Beleuchtung. Auf Grund der Wahl der Kameras, die schon versuchen die Abhängigkeit von der Beleuchtung von außen zu minimieren, ist dieser Effekt aber zumindest abgeschwächt. Es konnte im Projektverlauf auch gezeigt werden, dass Verfahren, die auf den Daten aus dem Simulator trainiert wurden, auch im Versuchsträger von Bosch unter realen Bedingungen funktionieren.

3.2.1 Studienvorbereitung

Zur Vorbereitung des Versuchs wurde durch das KIT-CVHCI eine Literaturrecherche durchgeführt mit dem Ziel Nebentätigkeiten zu identifizieren, die voraussichtlich häufig während zukünftiger automatisierter Fahrten durchgeführt werden. Zusätzlich wurden auch die Industriepartner im Projekt befragt, um deren Expertise und Wünsche soweit möglich zu berücksichtigen. Die dadurch entstandene Liste wurde priorisiert und jede Nebentätigkeit auf Machbarkeit in der Simulatorstudie bewertet. Dabei wurde zum Beispiel die Aktivität „Schlafen“ ausgeschlossen, da eine gespielte Situation nicht erwünscht war und Schlaf im Simulator zu erreichen zwar möglich ist, aber den Studienverlauf zu stark beeinträchtigt hätte. Abschließend wurden 30 Aktivitäten ausgewählt.

Probanden direkt zu instruieren, Nebentätigkeiten durchzuführen wäre zwar eine einfache Möglichkeit, aber ergibt weder einen natürlichen Bewegungsablauf noch eine natürliche Abfolge verschiedener Aktivitäten. Stattdessen wurden 11 Aufgaben entworfen, die jeweils mehrere Nebentätigkeiten umfassen. Eine Aufgabe war zum Beispiel, auf dem Laptop das Wetter nachzuschlagen und das Ergebnis auf dem Smartphone als Nachricht zu verschicken. Diese Aufgabe lässt sich in eine Vielzahl kleiner Aktivitäten unterteilen, die sich in natürlicher Weise ergeben. Für einige Nebentätigkeiten waren Objekte notwendig (z.B. Wasserflasche, Essen, Smartphone) die den Probanden bereitgestellt oder von ihnen mitgebracht wurden.

Zur Durchführung des Versuchs wurde der Fahrsimulator des Fraunhofer IOSB entsprechend vorbereitet. Der Fahrsimulator besteht aus einem umgebauten Audi A3, umgeben von mehreren Leinwänden zur Darstellung der Simulation. Da ein echtes Fahrzeug verwendet wird, steht auch ein realistischer Innenraum zur Verfügung, was für die Aufzeichnung der Kameradaten essentiell war. Der Simulator wurde für den Versuch mit dem beschriebenen Kamerasystem und mit entsprechender Aufzeichnungshardware ausgestattet. In der Simulationssoftware SILAB von WIVW wurde eine Autobahnstrecke umgesetzt, die für den Großteil des Versuchs automatisiert befahren wurde. Die entworfenen Aufgaben wurden in zufälliger Reihenfolge auf einem Display in der Fahrzeugmitte angezeigt und mussten nach Bearbeitung durch den Probanden abgehakt werden.

3.2.2 Studiendurchführung

Am Versuch nahmen 15 Probanden teil. Jeder Proband wurde zwei Mal aufgezeichnet. Beim zweiten Durchlauf war den Probanden der Versuchsablauf entsprechend bereits bekannt. Dies führt

zu anderen Aktivitätsabfolgen und Aktivitätsdauern, was in diesem Fall einen Vorteil darstellt, da es die Varianz der Trainingsdaten erhöht.

Der Ablauf jeder Sequenz war wie folgt: Zuerst wurden die für die Aufgaben notwendigen Objekte vom Versuchsleiter im Fahrzeug möglichst zufällig, aber sinnvoll verteilt. Anschließend startete der Durchlauf mit dem Probanden außerhalb des Fahrzeugs. Der erste Schritt bestand im Einsteigen, Einstellen des Sitzes und Anschnallen. Anschließend fuhren die Probanden mehrere Minuten manuell auf der Autobahn, um sich an die Simulation zu gewöhnen. Anschließend wurden sie auf dem Mitteldisplay aufgefordert den automatisierten Fahrmodus zu aktivieren. Nach und nach wurden Sie dann in zufälliger Reihenfolge aufgefordert die einzelnen Aufgaben zu lösen. Nach dem Lösen einer Aufgabe verging ein zufälliger Zeitraum von bis zu mehreren Minuten bis die nächste Aufgabe eingeblendet wurde. Vor dem Versuch wurde jeder Proband aufgefordert auch außerhalb der instruierten Aufgaben sich mit den im Innenraum zur Verfügung stehenden Objekten zu beschäftigen um noch natürlichere und abwechslungsreichere Aktivitätsabfolgen zu erzeugen. Viele nutzten die Pausen entsprechend, um zu essen oder zu trinken. Jeder Durchlauf endete mit dem Abstellen des Fahrzeugs am Straßenrand und dem Aussteigen aus dem Fahrzeug.

3.2.3 Manuelle Annotation der Daten

Insgesamt entstanden bei dem Versuch etwa 12 Stunden Videomaterial in 30 Sequenzen. Da die einzelnen Nebentätigkeiten nur lose durch die Aufgaben instruiert wurden, mussten die Daten manuell annotiert werden. Durch die lose Annotation ergab sich aber auch die Gelegenheit vor der Annotation die Videos nochmals genauer zu analysieren und weitere Label festzulegen. Hieraus entstand eine Hierarchie von Labels mit drei Leveln. Als erstes Hierarchielevel dienen die instruierten Aufgaben. Das zweite Hierarchielevel zerlegt diese Aufgaben in semantisch abgeschlossene Aktivitäten. Diese entsprechen Großteils den durch die Literaturrecherche identifizierten Nebentätigkeiten. Das dritte Hierarchielevel abstrahiert diese Aktivitäten weiter in allgemeine Interaktionen mit der Umgebung. Diese werden als Triplet aus Aktivität, Ort und Objekt abgebildet. Die vollständige Hierarchie ist in Tabelle 1 dargestellt. Der hierdurch erzeugte Datensatz wurde im Projektverlauf vom Fraunhofer IOSB und vom KIT-CVHCI zur Entwicklung maschineller Lernverfahren verwendet und gemeinsam auf einer internationalen Konferenz veröffentlicht [Martin.ICCV.2019].

Tabelle 1: Annotationshierarchie des Datensatzes zur Aktivitätserfassung

Level 1: Instruierte Aufgaben (Anzahl: 12)
Essen/Trinken, Fahrvorbereitung, Fahrzeug parken und aussteigen, Übernahme aus automatisierter Fahrt, Jacke anziehen, Sonnenbrille aufsetzen, Magazin lesen, Zeitung lesen, Jacke ausziehen, Sonnenbrille absetzen, Video schauen, Laptop arbeiten
Level 2: Feingranulare Aktivitäten (Anzahl: 34)

Objektmanipulationen: Laptop aus Rucksack nehmen, Laptop in Rucksack verstauen, Rucksack öffnen, Laptop öffnen, Laptop schließen, Objekt verstauen, Objekt holen **Fahrvorbereitung:** Tür von außen schließen, Tür von innen schließen, Tür von außen öffnen, Tür von innen öffnen, Fahrzeug betreten, Fahrzeug verlassen **Essen/Trinken:** Essen vorbereiten, Flasche schließen, Flasche öffnen, trinken, essen **Kleidung:** Sonnenbrille aufsetzen, Sonnenbrille absetzen, Jacke anziehen, Jacke ausziehen **Bewegung:** Etwas suchen, stillsitzen **Bedienung integrierter Geräte**: Automation aktivieren, Multimedia Display bedienen **Arbeiten**: Auf Papier schreiben, Am Laptop arbeiten **Unterhaltung**: Telefonieren, Smartphone bedienen, Zeitung lesen, Zeitschrift lesen
Level 3: Atomare Aktivitäten (Triplets aus Aktivität, Ort und Objekt) **Aktivitäten (Anzahl: 5):** öffnen, schließen, etwas ablegen, nach etwas greifen, von etwas zurückziehen, interagieren **Orte (Anzahl: 14)**: Hosentasche, Beifahrertür, Rücksitz links, Fahrertür, Rücksitz rechts, Mittelkonsole vorn, Beifahrer Fußraum, Mittekonsole hinten, Kopf, Schoß, Lenkrad, Beifahrersitz, vor dem Fahrer, kein Ort **Objekte (Anzahl: 17)**: Brillenetui, Stift, Schalthebel, Rucksack, Sonnenbrille, Gurt, Jacke, Automationstaste, Schreibblock, Flasche, Laptop, Zeitung, Magazin, Multimedia Display, Smartphone, Essen, kein Objekt

3.3 Verfahren zur 3D-Körperposenschätzung mit einem Multi-Kamera-System

Ein der Hauptaufgaben des Fraunhofer IOSB im Projekt war die Weiterentwicklung der vorhandenen Verfahren zur Körperposenschätzung auf Tiefendaten. Dabei sollten vor allem andere Kameramodalitäten, wie zum Beispiel Nahinfrarotaufnahmen anstelle von Tiefendaten verwendet werden, um so eine größere Flexibilität und Robustheit zu erreichen. Um dieses Ziel zu erreichen wurde das öffentlich verfügbare Openpose-Framework verwendet [Cao.2018], um in mehreren Kameraansichten die Position der Körpergelenke im Bild zu bestimmen. Diese wurde anschließend durch Einmessen der Kameras und Triangulation mehrerer Ansichten in ein 3D-Skelett überführt. Im Folgenden wird zuerst das Openpose-Framework zur 2D-Körperposenschätzung kurz vorgestellt und anschließend die Triangulation beschrieben.

3.3.1 2D-Körperposenschätzung auf Nahinfrarot und Farbbildern

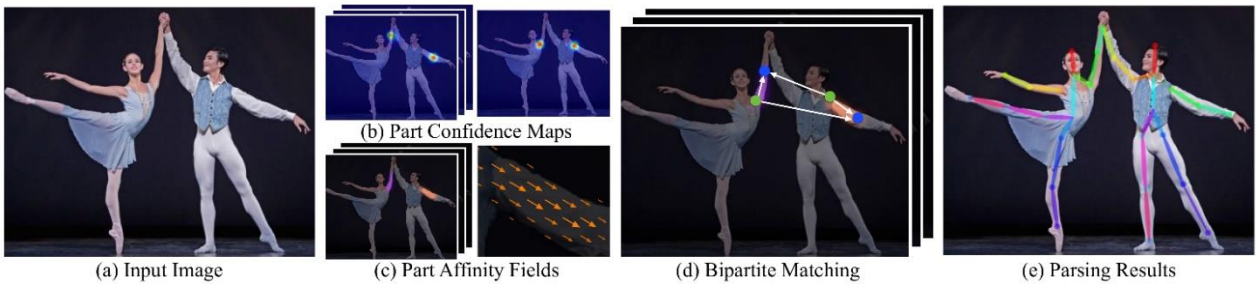


Abbildung 5: Übersicht der Einzelschritte zur Körperposenerfassung im OpenPose Framework (Bildquelle: [Cao.2018]).

Openpose basiert auf einer Veröffentlichung von 2018 [Cao.2018]. Das Verfahren bestimmt die Position der Körpergelenke auch von mehreren Menschen im Bild. Insbesondere bei mehreren Personen, aber auch bei Fehldetektion einzelner Körperteile, stellt sich das Problem der Zuordnung der Körperteile zu Personen. Verfahren lassen sich hierbei in zwei Kategorien unterteilen:

- Top-Down-Verfahren bestimmen zuerst mit einem Detektor die Bounding Box von Personen und im zweiten Schritt für jede Bounding Box getrennt die Gelenkpositionen der Person.
- Buttom-Up-Verfahren bestimmen zuerst die Gelenkpositionen aller Personen im Bild und lösen danach das Zuordnungsproblem einzelner Gelenkknoten zu sinnvollen Körperskeletten.

Openpose ist eines der ersten Verfahren, das den Bottom-Up-Ansatz erfolgreich umsetzt. Im Gegensatz zum Top-Down-Ansatz sind diese Verfahren besser für eine Echtzeitauswertung geeignet, da nur ein Detektor notwendig ist und im Allgemeinen die Rechenzeit nicht stark von der Anzahl der Personen im Bild abhängt.

Openpose basiert im Kern auf einem Neuronalen Netz und einer anschließenden Nachbearbeitung der Ergebnisse. Das gesamte Verfahren lässt sich in mehrere Teilschritte aufteilen, die im Folgenden kurz erläutert werden (siehe Abbildung 5). Das Neuronale Netz analysiert jedes Einzelbild einer Videosequenz separat. Es werden zuerst im Backbone-Netzwerk komplexere Merkmale aus den Bilddaten abgeleitet. Diese werden dann parallel weiter untersucht, um sowohl auf die Position von Gelenkknoten (part confidence maps) als auch auf die paarweise Zusammengehörigkeit einzelner Gelenkknoten (part affinity fields) zu schließen. Abschließend wird mit einem analytischen Verfahren aus der Position und den Zugehörigkeitspaaren das Skelett der Personen im Bild abgeleitet. Im Folgenden werden die Schritte genauer erläutert und abschließend die Veränderungen für den Fahrzeuginnenraum dargestellt.

3.3.1.1 Backbone Network

Ein Bild stellt für ein Neuronales Netz eine hochdimensionale Eingabe mit häufig hohem Anteil an für die eigentliche Aufgabe irrelevanter Information dar. Hierzu gehören Sensoreffekte wie Bildrauschen oder Verzerrung, aber auch Umwelteffekte wie Beleuchtung oder ähnliches. Für

Neuronale Netze, die die Körperpose erfassen sollen, ist aber beispielsweise auch der Hintergrund des Bildes oder die Farbe und Textur der Kleidung irrelevante Information. Die Aufgabe des Backbone-Netzes ist die Verarbeitung des Rohbildes und die Filterung und Abstraktion des Bildes zu Merkmalen, die für die folgende Aufgabe besser geeignet sind. Diese Netze stammen üblicherweise aus dem Bereich der Bildklassifikation und werden auf dem Datensatz Imagenet trainiert und evaluiert. Die Backbone-Architektur kann in vielen Fällen ausgetauscht werden, um unterschiedliche Ziele zu erreichen. Sei es eine möglichst gute Erkennungsrate (z.B. VGG [Simonyan.2015], ResNet [He.2016], DenseNet [Huang.2017]) oder einen möglichst kleinen Speicherbedarf (z.B. MobileNet V3 [Howard.2019]). Im Falle von Openpose wird üblicherweise VGG als Backbone verwendet.

3.3.1.2 Part Confidence Maps

Ziel dieser Stufe ist das Ableiten der Position aller Körpergelenke im Bild. Die wichtigste Designentscheidung ist hier die Repräsentation des Ergebnisses bzw. des Trainingsziels. Ein naiver Ansatz wäre eine direkte Regression der Bildkoordinaten jedes Gelenks. In der Praxis ist dies allerdings schwierig und für nicht zu guten Ergebnissen. Stattdessen verwenden die meisten aktuellen Ansätze Confidence Maps. Dabei handelt es sich um eine zweidimensionale Matrix in der jeder Pixel die Konfidenz beschreibt, dass diese Position im Eingangsbild einem Gelenk entspricht. Sowohl zum Training, als auch später zur Auswertung ist eine Matrix mit sehr wenigen diskreten Werten aber nicht stabil. Stattdessen wird eine zweidimensionale Gaussglocke mit kleiner Varianz an der Gelenkposition positioniert. Das Ergebnis ist in Abbildung 5b dargestellt. Im Falle von Openpose wird eine Confidence Map pro Gelenktyp bestimmt. Diese Confidence Maps werden anschließend in mehreren Stufen verbessert um zum Beispiel initiale Verwechslungen zwischen Rechter und Linker Seite zu korrigieren.

3.3.1.3 Part Affinity Fields

Betrachtet man nur die Confidence Maps einer Szene mit mehreren Personen so ergeben sich in jeder Matrix mehrere Maxima. Dies führt zu einem Zuordnungsproblem, da bei Personen die dicht beieinander sind die Zuordnung nur auf Basis der Gelenkpositionen mehrdeutig ist. Dies wird durch die Part Affinity Fields und die anschließende Nachbearbeitung gelöst.

Part Affinity Fields repräsentieren die Abhängigkeit von immer zwei benachbarten Gelenken in der kinematischen Kette des menschlichen Körpers. Ein Part Affinity Field wird wieder durch eine zweidimensionale Matrix repräsentiert. Jeder Wert der Matrix stellt einen Einheitsvektor dar. Bei Positionen in der Matrix die zwischen zusammengehörigen Gelenken liegen zeigt der Vektor die Richtung von einem zum anderen Knoten an. Abbildung 5c zeigt ein Beispiel.

3.3.1.4 Bipartite Graph Matching

Aus den durch das neuronale Netz bestimmten Part Confidence Maps und Part Affinity Fields wird abschließend ein Graph erzeugt, der alle Gelenke, die im Bild gefunden wurden, abbildet und dessen Kantengewichte die Wahrscheinlichkeit der Zusammengehörigkeit einzelner Gelenke anzeigt. Um die Knoten des Graphen zu bestimmen werden dabei alle Confidence Maps auf lokale Maxima untersucht. Um die Kanten des Graphen zu erzeugen werden anschließend alle Gelenkpositionen mit allen Nachbarn in der kinematischen Kette verbunden (z.B. alle rechten Schultern mit allen rechten Ellbogen). Jeder dieser Kanten wird dann ein Gewicht zugeordnet, indem entlang der Kante die Part Affinity Fields untersucht werden. Der so gewichtete Graph wird anschließend in energieoptimale Körperskelette unterteilt und somit die Zuordnung gelöst. Dieser Prozess ist NP-vollständig und deshalb global nur sehr aufwändig zu lösen. Im OpenPose Framework kommt deshalb ein Heuristik zum Einsatz die sehr effizient eine gute Lösung findet, die aber nicht garantiert das globale Optimum darstellt.

3.3.1.5 Adaption für den Fahrzeuginnenraum

Das Openpose-Framework funktioniert bereits sehr gut im Fahrzeug. Hauptfokus der Anpassungen lag auf der Verarbeitungsgeschwindigkeit da die Rechenkapazitäten im Versuchsträger begrenzt sind und zur Triangulation die Auswertung mehrere Kameraströme notwendig ist. Im ersten Schritt wurden die Trainingsdaten von Openpose mit den Daten der Kameras aus PAKoS verglichen (siehe Tabelle 2).

Tabelle 2: Vergleich der Trainingsdaten von Openpose mit den NIR Daten aus PAKoS.

	OpenPose Trainingsdaten	**PAKoS**
Bildursprung	Internet	Fahrsimulator, Versuchsträger
Bildtyp	Farbbild	Nahinfrarotbild
Perspektive	Nicht festgelegt	5 Ansichten
Szene	Hauptsächlich Outdoor-Szenen vor allem von Innenstädten	Fahrzeuginnenraum
Sichtbarkeit der Personen	Meist ganze Person sichtbar	Hauptsächlich Oberkörper
Personengröße	Starke Varianz, auch kleine Personen	Fahrer und Beifahrer füllen das halbe Bild. Rückbank in Teilen sichtbar
Personenanzahl	Hauptsächlich 1-2 aber nach oben unbegrenzt	Hauptsächlich 2, maximal 5

Es zeigen sich starke Unterschiede in fast allen Punkten. Trotzdem funktioniert das Verfahren bereits sehr gut. Ein erneutes Training speziell auf Daten des Fahrzeuginnenraums würde die Qualität voraussichtlich weiter steigern. Da dies aber einen hohen Annotationsaufwand bedeutet

hätte und die Qualität schon zu Anfang hoch genug erschien, wurde davon Abstand genommen und mehr Zeit auf die weitere Analyse der Körperpose für die Schätzung der Nebentätigkeit verwendet. Weitere Anpassungen konnten deshalb aber auch nur empirisch überprüft werden.

Selbst ohne weitere Trainingsdaten sind einige Anpassungen für den Fahrzeuginnenraum möglich. Der Hauptvorteil ist die feste Position der Kamera und die im Vergleich zum MS Coco-Datensatz hohe Größe der Personen im Bild. Hierdurch konnte das Eingabebild des Detektors ohne große Qualitätseinbußen auf 320x240 Pixel halbiert werden, um die Geschwindigkeit zu steigern. Des Weiteren wurde MobileNet V2 als Backbone-Netzwerk getestet, um die Geschwindigkeit noch weiter zu steigern, allerdings war der Qualitätsverlust dadurch zu groß. Das Originalverfahren verbessert sowohl Confidence Maps als auch Part Affinity Fields in 6 Schritten. Dies konnte auf 4 reduziert werden ohne sichtbar Qualität zu verlieren. Durch diese Anpassungen war es am Ende möglich das Verfahren auf drei Kameraströmen mit je 20Hz im Versuchsträger auszuführen.

3.3.2 3D-Daten durch Triangulation monokularer Messungen

Die Verwendung von 3D-Daten statt 2D-Bildmerkmalen für darauf aufbauende Analysen bietet einige Vorteile. Im Gegensatz zu 2D-Bildmerkmalen sind sie weniger stark abhängig von der Perspektive und können frei im 3D-Raum transformiert werden, um so beispielsweise eine geänderte Kameraposition zu kompensieren. Das Arbeiten mit 3D-Daten ermöglicht auch direkt über Position und Abstände von Körperteilen, Objekten oder ähnlichem auf Interaktion oder Nähe zu schließen.

Abhängig vom Kameratyp ergeben sich verschiedene Möglichkeiten 3D-Daten und insbesondere die 3D-Körperpose zu bestimmen. Im Projekt InCarIn und als Vorarbeit für PAKoS entwickelte das Fraunhofer IOSB bereits ein System, das die Körperpose auf Basis einer Tiefenkamera erfassen kann. Dieses System war zu Anfang auch Teil von PAKoS und diente unter anderem als Rückfallebene für neue Verfahren. Es funktionierte rein auf dem Tiefenbild der im Projekt vorhandenen Kinect V2. Nachteil des Systems waren hauptsächlich die geringe Robustheit gegenüber Störobjekten und Verdeckungen. In PAKoS wurde das System deshalb durch ein Multi-Kamerasystem erweitert bzw. am Ende ersetzt. Der klassische Ansatz aus mehreren monokularen Kameras 3D-Daten zu erzeugen ist die Triangulation, wie sie auch hier zum Einsatz kommt. Zusätzlich zur Körperpose wurde im Projekt auch der Einfluss von Objekten auf die 3D-merkmalsbasierte Nebentätigkeitsschätzung untersucht. Hierfür musste auch die 3D-Position dieser Objekte im Drive&Act-Datensatz bestimmt werden, was eine Triangulation von Objektboundingboxen erforderte. Im Folgenden werden beide Triangulationsverfahren kurz vorgestellt.

3.3.2.1 Triangulation der 3D-Körperpose

Wie bereits in Kapitel 3.1.3 beschrieben waren im Fahrsimulator des Fraunhofer IOSB zur Aufzeichnung des Drive&Act-Datensatzes 6 Kameras verbaut. Es sind jedoch nicht alle Ansichten

für die Triangulation der Körperpose geeignet. Die Kamera auf der Lenksäule sieht hauptsächlich den Kopf des Fahrers. Für Gesichtsmerkmale ist diese Kamera gut geeignet, aber da sie zu wenig Kontext zeigt war es nicht zielführend diese Kamera für die Triangulation zu nutzen. Die Kamera im Dachhimmel zeigt den Fahrer von hinten aus der Vogelperspektive. Die linke Körperhälfte ist dadurch häufig durch den Kopf und den Torso verdeckt. Damit verbleiben noch drei NIR-Kameras (A-Säule Fahrer/Beifahrer, Innenspiegel) und die Kinect V2 (A-Säule Beifahrer). Da sich die Perspektive von der Kinect zur NIR-Kamera an ähnlicher Position kaum unterscheidet wurden nur die drei NIR-Kameras ausgewertet. Abbildung 6 zeigt den schematischen Aufbau und die Ansichten aller Kameras. Nach Schätzung der 2D-Körperpose mit OpenPose und Einmessung des Kamerasystems mit intrinsischen und extrinsischen Parameters konnte dann durch klassische Triangulation die 3D-Position aller Gelenkknoten bestimmt werden.

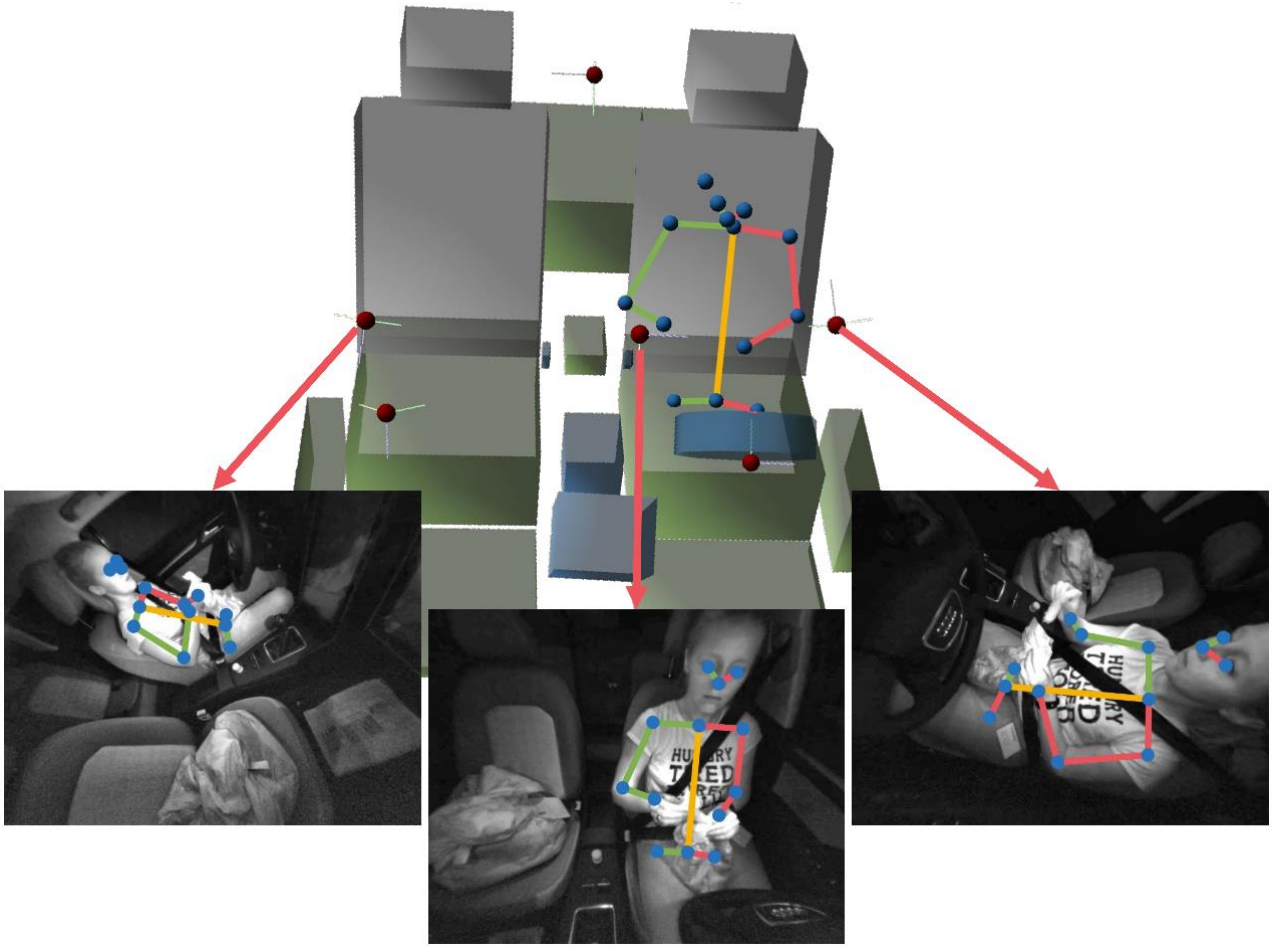


Abbildung 6: Schematische Darstellung des 3D-Innenraums mit 3D-Körperpose und den zur Triangulation verwendeten Kameraansichten

3.3.2.2 Triangulation der 3D-Objektpositionen

Die im Drive&Act Datensatz verwendeten Objekte stellen für die Objektdetektion aber insbesondere auch für die Triangulation eine Herausforderung dar. Schwierigkeiten für die Detektion umfassen die starke Varianz der Größe und die starken Verdeckungen. Zum Beispiel füllt eine Jacke teilweise das gesamte Kamerabild, während ein Smartphone während der Nutzung fast vollständig von der

Hand verdeckt wird. Um trotz dieser Herausforderungen den Einfluss der Objekte auf die Nebentätigkeitserfassung evaluieren zu können wurden alle Objektpositionen manuell annotiert und auf diesen Daten gearbeitet. Bei der manuellen Annotation wurden dabei Bounding Boxen jedes Objekts im gesamten Video annotiert. Bounding Boxen sind aber nicht kompatibel mit dem im letzten Abschnitt beschriebenen Verfahren zur Triangulation, da dieses korrespondierende Punkte in den einzelnen Kameraansichten voraussetzt. Die Ecken von Bounding Boxen erfüllen diese Bedingung aber nicht. Stattdessen wurde im Projekt eine volumetrische Triangulationsmethode entwickelt. Das Verfahren basiert auf der Betrachtung, dass eine Bounding Box im Bild einer Pyramide im 3D-Raum entspricht. Diese Pyramiden der Bounding Boxen aus unterschiedlichen Ansichten schneiden sich. Der Schnitt bildet eine Hülle um den Raum, den das Objekt belegt. Die Mitte dieser Hülle diente im Projekt als Position des Objekts. Bei kleinen Objekten wie dem Smartphone ist diese Position sehr genau während bei großen Objekten wie Zeitungen oder Jacken die berechnete Position ungenauer ist. Deren Zentrum ist im allgemeinen aber auch nicht genau festzulegen.

Um dieses Verfahren zu implementieren kamen verschiedene Verfahren aus dem Bereich CAD insbesondere constructive solid geometry (CSG) zum Einsatz. Diese Methoden ermöglichen es, den Schnitt der Pyramiden zu berechnen was im Ergebnis ein 3D-Mesh des umschlossenen Raums ergibt (siehe Abbildung 7). Der Mittelwert aller Ecken des Meshs diente als Objektposition.

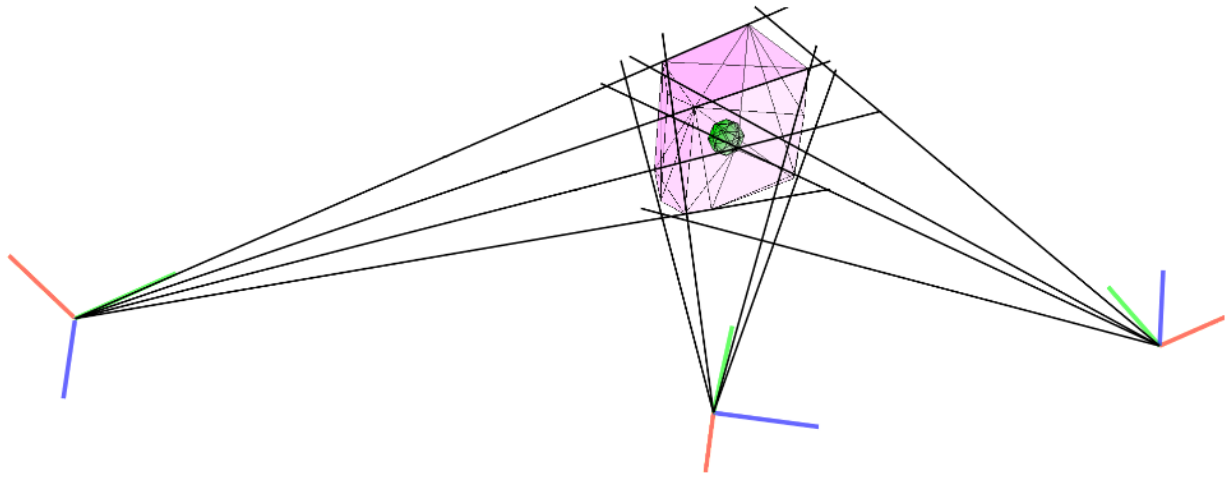


Abbildung 7: Darstellung der Triangulation von Bounding Boxen aus drei Ansichten durch Schnitt der belegten Raumvolumen.

3.4 Verfahren zur Schätzung der Nebentätigkeit des Fahrers auf 3D-Innenraumdaten

Im Projekt wurden durch das KIT-CVHCI und durch das Fraunhofer IOSB verschiedene Methoden untersucht, um auf die Nebentätigkeit des Fahrers zu schließen. Das KIT-CVHCI fokussierte dabei hauptsächlich auch die direkte Auswertung der aufgezeichneten Videodaten. Einer der Vorteile dieses Ansatzes ist, dass dem Verfahren alle im Video vorhandenen Informationen zur Verfügung

stehen. Dies stellt aber auch einen Nachteil dar, da dadurch das Verfahren auch allen im Video vorhandenen Störeinflüssen ausgesetzt ist. Am Fraunhofer IOSB wurde ein anderer Ansatz verfolgt. Statt die Videodaten direkt auszuwerten wurde mit den bereits vorgestellten Verfahren die 3D-Position von Objekten und das 3D-Skelett des Fahrers bestimmt. Diese Daten wurden durch die Position von Innenraumelementen weiter angereichert. Diese abstrahierte Darstellung wurde dann durch weitere maschinelle Lernverfahren ausgewertet, um Nebentätigkeiten zu bestimmen. Durch dieses Vorgehen werden viele Störeinflüsse eliminiert. Angefangen von Beleuchtungsunterschieden und Kameraperspektive bis hin zur Innenraumausstattung. Nachteilig für diesem Ansatz ist, dass nur die Informationen zur Verfügung steht, die explizit modelliert wurden, was sich auch auf die Qualität der Erfassung auswirken kann.

Im Folgenden wird das im Projekt entwickelte Verfahren zur Auswertung des abstrahierten Innenraummodells im Detail beschrieben. Die Eingabe wird dabei als Graph modelliert (siehe Abbildung 8). Jeder Knoten im Graph beschreibt eine 3D-Position im Fahrzeuginnenraum. Körpergelenke, Objekte und Innenraumelemente stellen Knoten im Graph dar. Die Kanten des Graphen beschreiben den Zusammenhang zwischen einzelnen Knoten. Verbindungen zwischen Körpergelenken bilden dabei das Skelett des menschlichen Körpers ab. Verbindungen zwischen Körpergelenken und Objekten bzw. Innenraumelementen stellen die Interaktion bzw. Beteiligung dieser Knoten bei der aktuell durchgeführten Nebentätigkeit dar. Zur Auswertung eines solchen Graphen eignen sich Neuronale Netze, die Graph-Convolutions verwenden. Die Folgenden Abschnitte beschreiben zuerst die Grundlagen des Neuronalen Netzes, gefolgt von der detaillierten Modellierung der Eingabedaten als Graph und dem Aufbau des gesamten Neuronalen Netzes. Abschließen wird die Güte des Verfahrens auf dem gesammelten Drive&Act-Datensatz evaluiert.

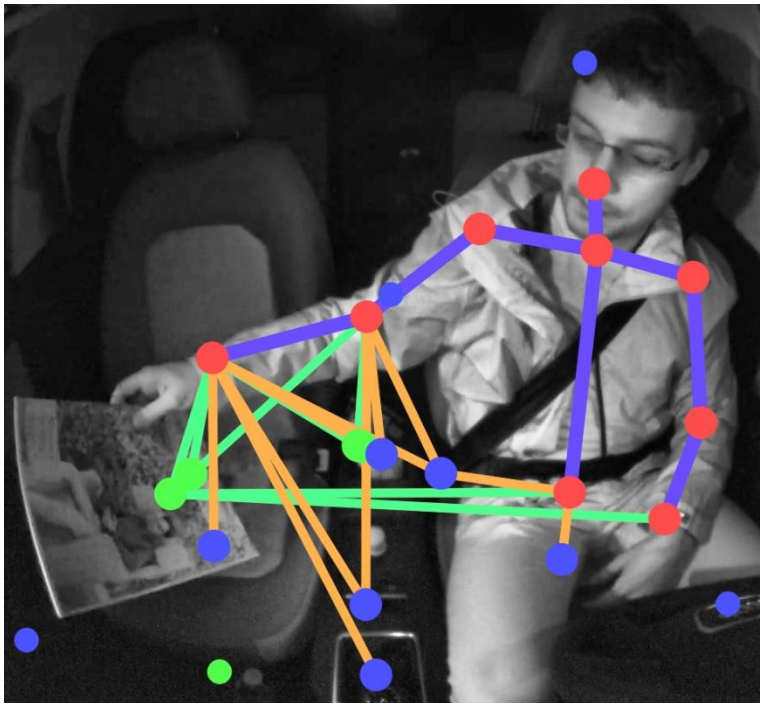


Abbildung 8: Darstellung des Interaktionsgraphen für die Aktivität "Ein Objekt greifen".

3.4.1 Aufbau des temporalen Graph Convolution-Layers

Insbesondere Netze die Bilder verarbeiten bestehen häufig aus vielen Schichten die auf Faltung (Convolution) mit einer gelernten Gewichtsmatrix basieren. Ein Bild kann auch als Graph mit fester Topologie aufgefasst werden. Die allgemein übliche Faltung auf Bildern stellt somit einen Spezialfall von Faltungen auf Graphen dar. Die dem hier vorgestellten Verfahren zugrundeliegende Formalisierung einer Faltung auf Graphen wurde erst 2017 von Kipf et al. als Methode zur Klassifizierung von Knoten in Graphen vorgestellt [Kipf.2017]. Um diese Formalisierung für die Schätzung von Aktivitäten auf Basis der Körperpose zu verwenden sind allerdings Erweiterungen notwendig, da diese Graphen sowohl eine räumliche Komponente in Form der Position der einzelnen Gelenke, als auch eine temporale Komponente in Form der Bewegung einzelner Gelenke über die Zeit haben. In der Literatur gibt es hierfür verschiedene Ansätze. Im vorgestellten Verfahren wird auf den Arbeiten von Yan et al. aufgebaut [Yan.2018]. In den folgenden Abschnitten wird zuerst die Faltung über Graphen im Allgemeinen präsentiert und anschließend die für die Aktivitätserkennung notwendigen Erweiterungen.

Ein Graph kann als eine Menge von Knoten *V* und Kanten *E* definiert werden, die diese Knoten verbinden. Ein solcher Graph kann auch als Adjazenzmatrix *A* repräsentiert werden. Jeder Eintrag der Matrix stellt hierbei eine Kante dar. Der Wert des Eintrags kann als Gewicht der Kante aufgefasst werden. Die Adjazenzmatrix beschreibt die Kanten des Graphen und deren Gewichte zusätzlich kann auch jedem Knoten noch ein Merkmal zugeordnet werden. Diese können in der Merkmalsmatrix *H* zusammengefasst werden. Ziel des Graph Convolution Layers ist die Überführung der Merkmalsmatrix H^l zu H^{l+1} basierend auf den Merkmalen der Nachbarknoten, die durch die Adjazenzmatrix $\tilde{A}$ definiert ist unter Berücksichtigung einer lernbaren Gewichtsmatrix W^l. Daraus ergibt sich folgende Formel:

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l W^l)$$

σ stellt die Aktivierungsfunktion des Layers dar. $\tilde{D}$ ist eine Normalisierungsmatrix mit der Zeilensumme der Adjazenzmatrix auf der Diagonalen. Hierdurch wird verhindert, dass Knoten mit einer größeren Summe von Kantengewichten ein größeres Gewicht im Vergleich zu anderen Knoten bekommen.

Ein Nachteil dieser Definition ist die fehlende Ordnung der Nachbarknoten daraus folgt, dass nur ein unspezifisches Gewicht für alle Nachbarn gelernt werden kann. Im Gegensatz dazu werden beispielsweise bei einer 3x3-Faltung auf Bildern 9 Gewichte gelernt, da hier die Nachbarschaft eindeutig definiert ist. Auf generischen Graphen ist die Ordnung der Nachbarschaft im Allgemeinen nicht definiert. Eine Ordnung lässt sich nur anwendungsspezifisch festlegen. Für die Erfassung von Tätigkeiten aus der Körperpose gibt es hierfür verschiedene Ansätze. Für das hier vorgestellte Verfahren wird die Nachbarschaft über die Distanz zu einem Wurzelknoten definiert. Hierdurch lassen sich Nachbarn in drei Gruppen trennen – Knoten die näher an der Wurzel liegen, Knoten mit

gleichem Abstand und Knoten, die weiter weg liegen. Dies kann in der Formel durch eine Aufteilung der Adjazenzmatrix in 3 Teilmatrizen abgebildet werden.

Die bis jetzt beschriebene Definition betrachtet alle Kanten gleich. Man könnte zwar die räumlichen und temporalen Daten der Skelettbewegung über die Zeit mit dieser Definition abbilden, allerdings würde der Graph, und dadurch die Adjazenzmatrix, bei vielen Zeitschritten sehr groß. Die temporale Dimension wird deshalb separat behandelt. Für den räumlich aufgespannten Graphen jedes Zeitschritts kommt die bereits präsentierte Definition zum Einsatz. Temporal werden alle Knoten mit ihren direkten zeitlichen Nachbarn verknüpft. Diese Kanten haben eine eindeutige Ordnung und kann deshalb mit einer eindimensionalen Faltung über die Zeit modelliert werden.

3.4.2 Erstellung des Interaktionsgraphen

Nachdem im letzten Abschnitt der grundsätzliche Aufbau des Graph Convolution Layer definiert wurde wird hier der Aufbau des Graphen beschrieben den das Neuronale Netz als Eingabe nutzt und auswertet. Im Gegensatz zu den meisten Vorarbeiten integriert der aktuelle Ansatz nicht nur die Körperpose, sondern auch Innenraumelemente und Objekte. Da insbesondere Objekte nicht ständig vorhanden sein müssen, ergibt sich daraus ein dynamischer Graph mit unterschiedlicher Anzahl an Knoten. Des Weiteren sind die Kanten des Graphs zwischen Gelenkknoten durch den Aufbau des menschlichen Körpers bereits gut definiert. Es ist aber nicht offensichtlich was die beste Methode ist, um Objekte und Körperpose mit Kanten zu verbinden. Das hier vorgestellte Verfahren folgt der Intuition, dass Objekte mit denen interagiert wird im Allgemeinen auch nahe am Körper sein müssen. Entsprechend werden Objekt und Innenraumknoten in allen Zeitschritten des Auswertezeitraums mit allen Knoten der Körperpose verbunden, wenn die Distanz zu einem Zeitpunkt in diesem Intervall unter eine definierte Schwelle fällt. Um den Graph vollständig zu definieren müssen die Knoten, das Merkmal jedes Knotens und die Kanten im Graph festgelegt werden.

Das **Merkmal jedes Knotens** ist seine Position zum Zeitschritt t im Auswerteintervall. Als Besonderheit sind hier noch die Innenraumelemente zu nennen, da diese eine feste Position haben wird deren Position für jeden Zeitschritt dupliziert. Falls Objekte oder Teile der Körperpose in manchen Zeitschritten nicht erfasst werden konnten wird deren Position zu diesem Zeitschritt genullt.

Alle Knoten des Graphen sind entweder Körperteile, Objekte oder Innenraumelemente. Der Graph beinhaltet aber nur Knoten die auch mit Kanten verbunden sind. Alle nicht verbundenen Knoten werden ignoriert, selbst wenn sie erfasst wurden.

Die Kanten des Graphen bestehen aus zwei Teilen. Es gibt fest definierte Kanten. Diese bilden das menschliche Oberkörperskelett ab. Des Weiteren gibt es noch dynamische Interaktionskanten. Diese bilden die Interaktion zwischen Körperknoten und Objekt bzw. Innenraumknoten ab.

Interaktionskanten werden nur erzeugt, falls die euklidische Distanz zwischen den Endknoten unter einen definierten Schwellwert fällt.

3.4.3 Struktur des Neuronalen Netzes

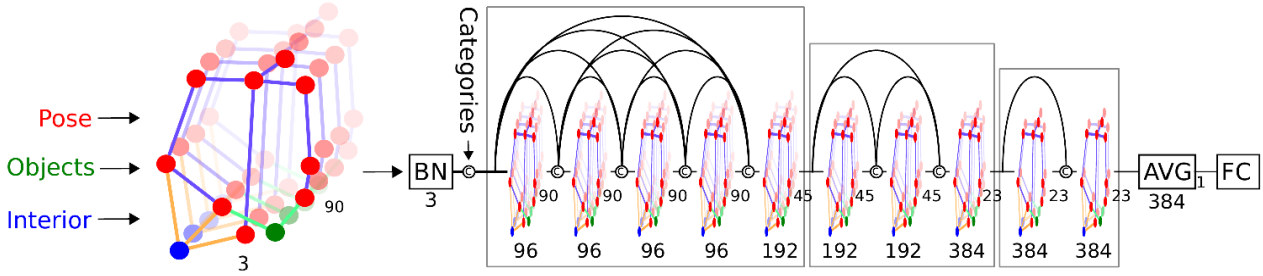


Abbildung 9: Überblick über das graphbasierte Neuronale Netz zur Aktivitätserfassung.

Abbildung 9 zeigt einen Überblick der Struktur des Neuronalen Netzes. Zuerst wird Batch-Normalisierung auf die Eingabe angewendet, um die Eingabevarianz und den Versatz der Knoten im Weltkoordinatenraum zu normalisieren. Der Rest des Netzwerks folgt der Idee von DenseNet [Huang.2017]: Es wird die Ausgabe aller vorhergehenden Schichten eines Blocks zur Eingabe der aktuellen Schicht verknüpfen. Insgesamt besteht das Netzwerk aus drei Blöcken. Am Ende jedes Blocks wird die Faltung über die Zeit mit einer Spreizung von zwei angewendet, um so mehrere Zeitschritte im folgenden Block zu aggregieren. Dadurch verringert sich die Eingabe am Ende des Neuronalen Netzwerks auf ein Viertel in der temporalen Domäne. Um die Merkmale aller verbleibenden Zeitschritte zu kombinieren, wird dann der Durchschnitt über alle Zeitschritte in der temporalen Achse gebildet.

Die kombinierten Merkmale aller Knoten werden am Ende durch eine vollverbundene Schicht mit Softmax-Aktivierung klassifiziert. Jeder raum-zeitliche Baustein besteht aus einer raum-zeitlichen Faltung, gefolgt von Batch-Normalisierung, Dropout und Rectified Linear Units (Relu) als Aktivierungsfunktion. Das Fenster des zeitlichen Faltungskerns beträgt 13 für alle raum-zeitlichen Faltungsschichten.

3.4.4 Evaluation

Der in PAKoS aufgezeichneten Drive&Act-Datensatz wurde mit unterschiedlichen Ansätzen sowohl vom Fraunhofer IOSB als auch vom CVHCI-Lab verwendet. Im Folgenden wird zur besseren Vergleichbarkeit mit dem Stand der Technik das beste Ergebnis des CVHCI-Lab (I3D Net) und ein Vorgängerverfahren des Fraunhofer IOSB (Three-Stream) mit aufgeführt.

Zur Evaluation wurden die Videoströme des Datensatzes in Drei-Sekunden-Blöcke unterteilt und von den Verfahren klassifiziert. Zum Training und zur Evaluation wurden die 15 Probanden des Datensatzes in 10 für das Training, 2 für die Validierung und 3 für das Testing aufgeteilt. Um den Datensatz möglichst effizient zu nutzen fand diese Aufteilung drei Mal statt mit jeweils anderen Probanden im Validierungs- und Testset. Das Gesamtergebnis wurde dann durch Aggregation der

Ergebnisse aller drei Teile und anschließender Berechnung von Evaluationsmetriken erzeugt. Als Metrik diente hier die durchschnittliche Top 1 Erkennungsrate pro Klasse (Mean Average Precision). Es wird für jede Hierarchieebene der Annotation ein neues Modell trainiert und evaluiert: 12 instruierte Aufgaben (Level 1), 34 feingranulare Aktivitäten (Level 2) und atomare Aktivitäten mit 372 möglichen Kombinationen der [Aktion, Objekt, Ort] Tripletts (Level 3). Da die Anzahl der Triplett-Kombinationen sehr hoch ist, werden auch getrennt die Ergebnisse für korrekt klassifizierte Aktivitäten, Objekte und Orte (6, 17 bzw. 14 Klassen) bestimmt.

Tabelle 3, Tabelle 4 und Tabelle 5 zeigen die Ergebnisse aller Modelle auf allen Annotationsleveln. Um den Einfluss der einzelnen 3D-Eingabedaten (Pose, Objektpositionen, Innenraumpositionen) einzeln evaluieren zu können wurden Modelle mit unterschiedlichen Kombinationen überprüft. Es lassen sich ähnliche Trends in allen Leveln erkennen. Bereits die Verwendung der Körperpose ohne weitere Eingabedaten führt in den meisten Fällen zu besseren Ergebnissen als frühere Verfahren die auf der Körperpose aufbauen. Dies zeigt, dass Graph basierte Verfahren besser die Abhängigkeit unterschiedlicher 3D-Elemente sowohl räumlich als auch zeitlich modellieren können. Das Hinzufügen der Innenraumpositionen führt allerdings nicht zu einer weiteren Verbesserung, sondern zu einer leichten Verschlechterung. Eine mögliche Erklärung liegt in der Duplikation der statischen Innenraumposition über alle Zeitschritte. Dadurch entsteht viel redundante Information, die durch das Netz erst wieder aggregiert werden muss. Das Hinzufügen von Objektpositionen zu den Gelenkpositionen verbessert das Ergebnis in allen Annotationslevel stark. Insbesondere die Objektkategorie der Triplets atomarer Aktivitäten profitier mit einer Verbeserung von 11% am meisten.

Abschließend zeigt Abbildung 11 noch einige Beispielergebnisse des Gesamtsystems für verschiedene feingranulare Aktivitäten (Level 2). Es sind neben der Aktivitätserkennung auch die 3D-Körperpose, 3D-Objektpositionen, die 3D-Innenraumelemente sowie der Interaktionsgraph dargestellt. Zur Darstellung wurden die 3D-Daten auf die Ansicht der Innenspiegelkamera projiziert. Die Verbindungen zwischen diesen Positionen bilden den Interaktionsgraphen der durch die Aktivitätserkennung ausgewertet wird. Lila markierte Verbindungen stellen dabei die Körperpose dar, wie sie durch die Triangulation von OpenPose entsteht. Orangene Kanten markieren die Interaktion zwischen Körperpose und Objekten, und grün markierte Kanten zeigen die Interation zwischen Körperpose und Innenraumelementen. Es ist gut zu erkennen, dass das Skelett selbst in schwierigeren Situationen mit großen Verdeckungen noch gut erkannt wird. Die Aktivitätserkennung funktioniert in vielen Fällen bereits gut. Oftmals treten Verwechslungen zwischen Verwandten Aktivitäten auf wie in den letzten zwei Bildern dargestellt.

Tabelle 3: Mean Average Precision der instruierte Aufgaben (Level 1) in Prozent.

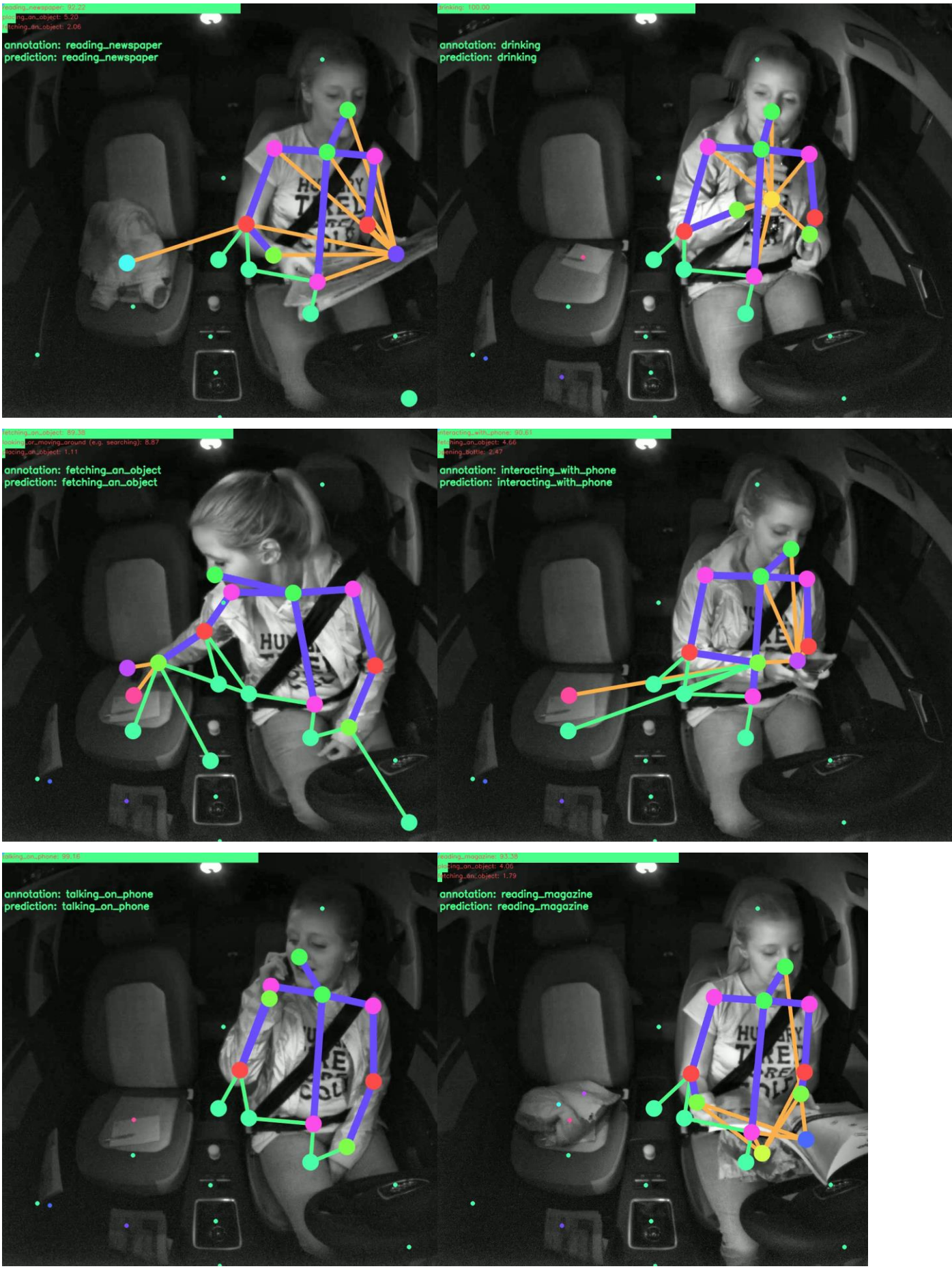
Model	Validation	Test
Zufall (Baseline)	2,94	2,94
I3D Net [31]	69,57	63,64
Two-Stream [10]	53,76	45,39
Three-Stream [24]	55,67	46,95
Pose	55,8	51,26
Pose+Innenraum	53,4	48,07
Pose+Objekte	65,87	58,8
Pose+Objekte+Innenraum	63,15	59,02

Tabelle 4: Mean Average Precision der feingranulare Aktivitäten (Level 2) in Prozent.

Model	Aktivitäten		Objekte		Ort		Kombiniert	
	val	test	val	test	val	test	val	test
Zufall (Baseline)	16,67	16,67	5,88	5,88	7,14	7,14	0,39	0,31
I3D Net [31]	62,81	56,07	56,07	56,15	47,7	51,12	15,56	12,12
Two-Stream [10]	57,86	48,83	48,83	42,79	53,99	54,73	10,31	7,11
Three-Stream [24]	59,29	50,65	50,65	45,25	59,54	56,5	11,57	8,09
Pose	56,85	51,63	51,63	45,57	49,83	52,35	11,5	8,85
Pose+Innenraum	56,00	48,24	48,24	44,5	47,25	55,01	8,88	7,43
Pose+Objekte	54,55	50,65	50,65	56,23	44,93	44,72	15,5	10,75
Pose+Objekte+ Innenraum	57,76	58,12	58,12	56,04	47,71	50,11	15,99	13,38

Tabelle 5: Mean Average Precision der atomaren Aktivitäten (Level 3) in Prozent.

Model	Validation	Test
Zufall (Baseline)	8,33	8,33
I3D Net (End-to-end)	44,66	31,80
Two-Stream	39,37	34,81
Three-Stream	41,70	35,45
Pose+Objekte+Innenraum	42,82	37,84


annotation: reading_newspaper
prediction: reading_newspaper
annotation: drinking
prediction: drinking
annotation: fetching_an_object
prediction: fetching_an_object
annotation: interacting_with_phone
prediction: interacting_with_phone
annotation: talking_on_phone
prediction: talking_on_phone
annotation: reading_magazine
prediction: reading_magazine

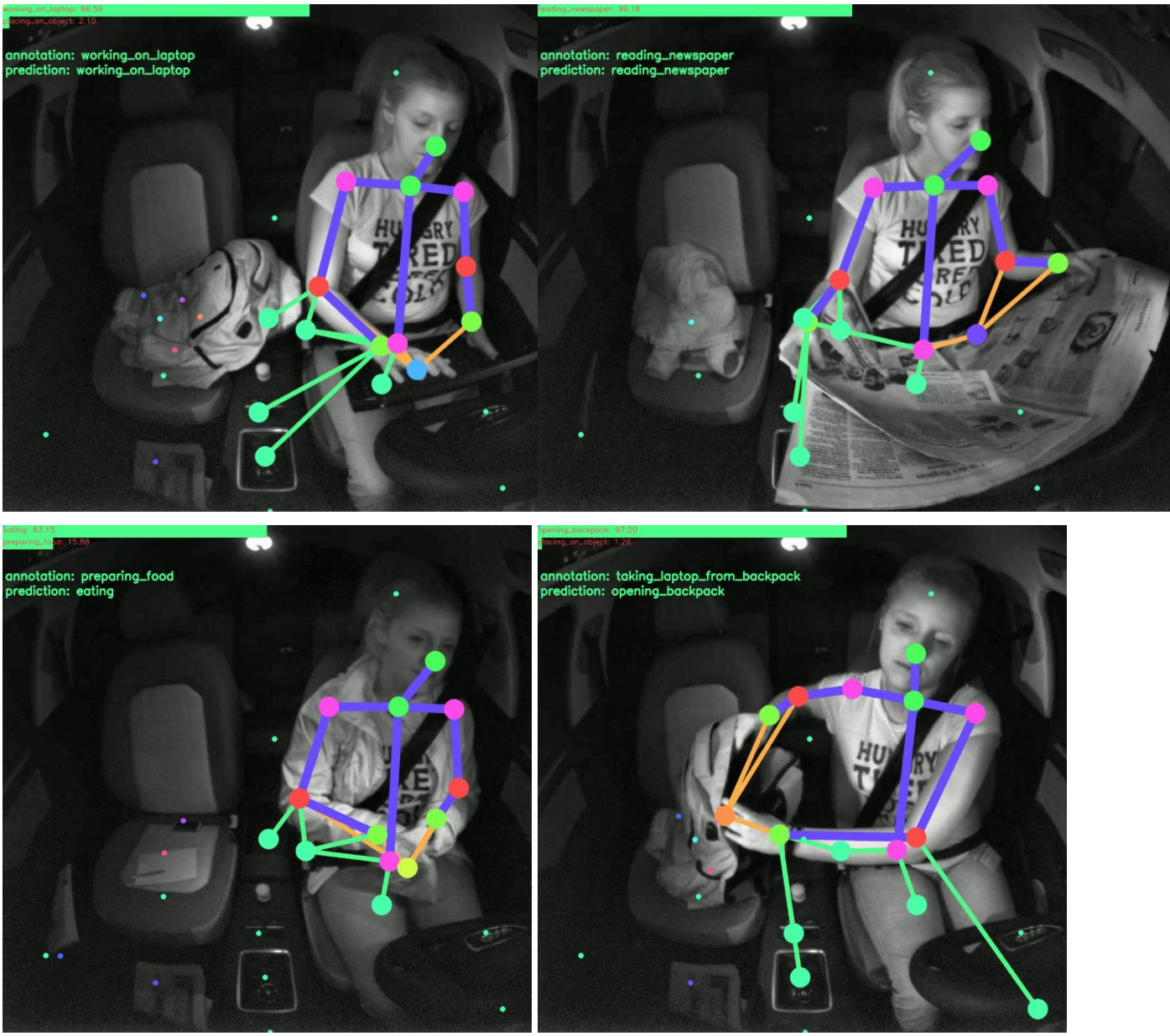


Abbildung 10: Einige qualitative Ergebnisse der gesamten Innenraumerfassung für die Annotation feingranulare Aktivitäten (Level 2). Annotiert ist der Interaktionsgraph als Projektion der 3D-Informationen ins Bild und die Aktivitätserkennung als Text. Das durch Triangulation der Openpose Ergebnisse erzeugte 3D-Skelett ist Lila markiert. Die letzten zwei Bilder zeigen typische Fehlerfälle, die durch Verwechslung verwandter Aktivitäten entstehen.

3.5 Gesamtsystementwurf und Integration in den Simulator und den Versuchsträger

Auf Grund der Vielzahl der Kameras und mehrerer Algorithmen unterschiedlicher Projektpartner, die auf die Videodaten angewiesen waren, war eine Integration aller Komponenten in einer flexiblen Struktur notwendig. Als Framework diente hier das Quelloffene Robot Operating System (ROS). Es ermöglicht die lose Kopplung einzelner Verarbeitungsknoten durch Kommunikation zwischen Knoten über eine Netzwerkschnittstelle. Das System bietet dabei auch die Möglichkeit die Verarbeitungsknoten auf unterschiedliche Rechner zu verteilen. Des Weiteren kann jegliche Netzwerkkommunikation zwischen Knoten aufgezeichnet und wieder abgespielt werden, was auch

zur Aufzeichnung der Kameraströme für den Drive&Act-Datensatz genutzt wurde. Alle Algorithmen von Projektpartnern wurden durch das Fraunhofer IOSB in diese Architektur integriert.

Das gesamte Sensorsystem und auch zwei Verarbeitungsrechner mit nVidia-Grafikkarten für die Beschleunigung maschineller Lernverfahren wurden sowohl im Fahrsimulator als auch im Versuchsträger von Bosch installiert. Der Fahrsimulator diente dabei zur Aufzeichnung des Drive&Act-Datensatzes und als Entwicklungsplatform. Anschließend wurde das System unverändert im Versuchsträger bei Bosch installiert und zur Kommunikation mit dem Fahrzeug eine zusätzliche CAN-Schnittstelle implementiert.

Abbildung 11 zeigt eine Übersicht aller im Laufe des Projekts entstandenen Verarbeitungsknoten und deren Kommunikationsfluss. Da die Visualisierung mit fast allen anderen Knoten verbunden ist wurden deren Verbindungen auf Grund der Übersichtlichkeit nicht dargestellt. Sowohl im Simulator als auch im Versuchsträger von Bosch wurden bei Bedarf zwei Rechner für die Verarbeitung verwendet. Blau markierte Knoten wurden dabei auf dem Rechner ausgeführt an die auch die Kameras angeschlossen waren. Rot markierte Knoten waren auf einen zweiten Rechner ausgelagert, der nur die Verarbeitung durchführte.

Auf Grund des flexiblen Systems konnten problemlos im Laufe des Projekts einzelne Knoten oder sogar Kameras hinzugefügt beziehungsweise entfernt werden. Was zu verschiedenen Zeitpunkten im Projekt auch aktiv genutzt wurde:

Aufzeichnung des Drive&Act-Datensatzes: Da eine Verarbeitung der Daten hier nicht notwendig war und das Aufzeichnen aller Ströme den Rechner bereits sehr belastete, wurden hier nur alle Kameraknoten und der Recorder gestartet.

KIT- Jahresfeier (siehe Kapitel 3.6): Hier wurde die Innenraumerfassung im Stand demonstriert. Da die Kinect und die Dachhimmel Kamera für die Verarbeitung am Ende des Projekts keine Rolle spielten wurden sie entsprechend ausgebaut und die Knoten deaktiviert. Alle anderen Knoten mit der Ausnahme von Recorder und CAN-Schnittstelle waren aktiv.

PAKoS-Abschlussveranstaltung: Für die Demofahrten war nur die Aktivitätserkennung notwendig. Entsprechend waren nur drei NIR-Kameras (Innenspiegel, Beifahrer und Fahrerseite) verbaut und aktiv sowie alle an der Aktivitätserkennung beteiligten Knoten und die CAN-Schnittstelle zur Weitergabe der Daten an das Fahrzeug.

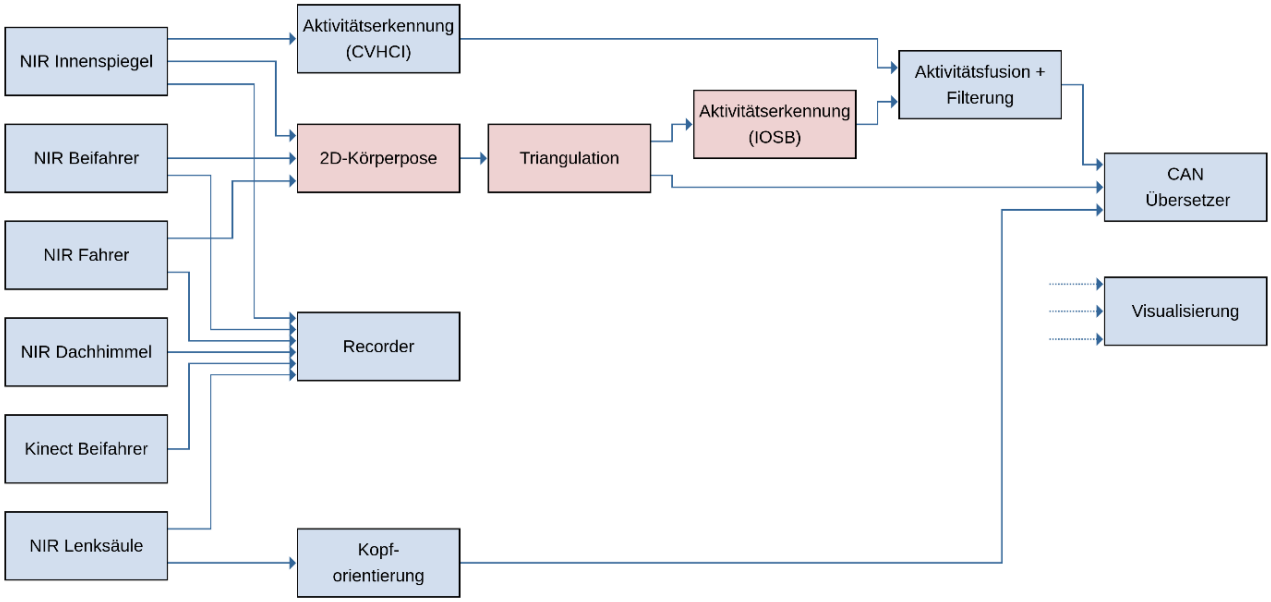


Abbildung 11: Gesamtsystemarchitektur der Innenraumerfassung. Rot markierte Knoten werden auf einem zweiten Rechner ausgeführt.

3.6 Verbreitung der Ergebnisse

Die PAKoS-Homepage wurde durch das KIT erstellt. Die vollständige Homepage ist unter folgender URL verfügbar: http://www.projekt-pakos.de/. Durch das Fraunhofer IOSB wurden die Ergebnisse auf mehreren nationalen und internationalen Konferenzen vorgestellt (siehe Kapitel 6). Teile des Systems wurden schon während der Projektlaufzeit regelmäßig im Fahrsimulator des Fraunhofer IOSB verschiedenen Industriekunden demonstriert. Des Weiteren wurde die Innenraumerfassung im Versuchsträger von Bosch an der Jahresfeier des KIT 2019 und der Abschlusspräsentation von PAKoS in Pferdsfeld präsentiert.

4 Voraussichtlicher Nutzen & Verwertbarkeit der Ergebnisse

Die im Projekt erzielten Ergebnisse flossen schon zur Projektlaufzeit in verschiedene Industrieprojekte ein und es ist zu erwarten, dass sich dieser Trend fortsetzt, da durch die Verabschiedung von Regeln für automatisierte Fahrfunktionen absehbar ist, dass eine Innenraumerfassung für den Betrieb eines automatisierten Fahrzeugs voraussichtlich verpflichtend wird. Hier helfen die Fortschritte, sowohl bei der robusteren Erfassung der Körperpose, als auch der detaillierten Erfassung von Nebentätigkeiten, die Expertise des Fraunhofer IOSB am Markt zu platzieren.

Wissenschaftlich wurden die Ergebnisse schon zur Projektlaufzeit in Form von Veröffentlichungen und Vorträgen verwertet. Die Ergebnisse fließen auch direkt in Dissertationen ein, die im Projektrahmen fortgeführt wurden. Mit dem veröffentlichten Drive&Act-Datensatz steht auch erstmals eine große öffentliche Datenbasis für die Entwicklung robuster maschineller Lernverfahren für den Fahrzeuginnenraum zur Verfügung. Das Fraunhofer IOSB wird in Form von Dissertationen

und Abschlussarbeiten auch über das Projektende hinaus mit diesem Datensatz arbeiten. Gleiches kann durch die Veröffentlichung auch die übrige weltweite Forschungslandschaft tun was schon in anderen Bereichen zu schnellem Fortschritt führte. Die Verfahren zur Körperposenerfassung und Aktivitätserkennung können aber nicht nur im Fahrzeuginnenraum eingesetzt werden. Am Fraunhofer IOSB werden ähnliche Systeme auch zum Beispiel bei der manuellen Montage oder der Assistenz in der Pflege eingesetzt. Hier flossen Teilergebnisse schon zur Projektlaufzeit in das BMBF-geförderte Projekt ASARob ein.

Abschließend dienen die gewonnenen Erkenntnisse auch als Grundlage für weitere Anträge öffentlich geförderte Projekte im Bereich Mobilität, aber auch anderen Anwendungsgebieten die ein Verständnis menschlicher Aktivitäten bzw. moderne Mensch-Maschine-Interaktion erfordern.

5 Fortschritt auf diesem Gebiet bei anderen Stellen

Bereits in den Jahren vor Projektbeginn begann im Bereich maschinelles Lernen und maschinelles Sehen mit der Popularität und dem Erfolg von Deep Learning ein großer Umbruch. Dieser breitete sich nach und nach über verschiedene Aufgaben und Anwendungsfelder aus. Haupttreiber für diese Ausbreitung ist die Verfügbarkeit großer annotierter Datenmengen, die es oft erst möglich machen moderne neuronale Netze zu entwickeln und zu trainieren. Die Anwendung im Fahrzeuginnenraum profitiert hier stark vom Fortschritt in anderen Domänen und folgt der Entwicklung oft mit etwas Verzögerung. Im Folgenden werden deshalb zuerst die allgemeine Datenlage und der Fortschritt der Algorithmen dargestellt bevor auf die fahrzeugspezifische Anwendung eingegangen wird. Im Falle der für das Projekt und die Aufgaben des Fraunhofer IOSB relevanten Körperposenerfassung und Aktivitätserkennung fiel der Umbruch zu tiefen neuronalen Netzen auf Grund der Veröffentlichung entsprechender Datensätze in etwa auf den Projektbeginn.

Für die Entwicklung der Körperposenerfassung auf monokularen Bildern stand bereits ab 2014 ein Datensatz des Max-Plank-Instituts [Andriluka.2014] zur Verfügung. Die Popularität dieses Forschungsfeldes nahm aber erst mit der Veröffentlichung des COCO Keypoint Challenge 2016 stark zu [Lin.2014]. Seitdem werden die Verfahren stetig besser und erreichen inzwischen eine Qualität, die es erlaubt die Verfahren ohne domänenspezifische Trainingsdaten erfolgreich einzusetzen. Ein solches Verfahren ist das OpenPose-Framework [Cao.2018] das 2018 veröffentlich wurde. Es dient auch in PAKoS als Grundlage der Körperposenerfassung und wird auch in vielen anderen Anwendungen unter anderem im Fahrzeug Innen- und Außenraum eingesetzt. Es gibt aktuell allerdings immer noch keinen öffentlichen Datensatz für die Körperposenerfassung speziell im Fahrzeug. Entsprechend sind auch keine fahrzeugspezifischen neuen Verfahren bekannt.

Die Entwicklung von Verfahren zur Aktivitätserkennung verläuft ähnlich. Hier wurde 2016 mit dem NTU RGB+D Datensatz [Shahroudy.2016] eine Datenbasis für die Aktivitätserkennung speziell auf

der Körperpose veröffentlicht. Im darauffolgenden Jahr folgte dann mit dem Kinetics-Datensatz [Carreira.2017] eine sehr großer Datenbasis zur videobasierten Aktivitätserkennung. In beiden Bereichen entwickelten sich die Verfahren anschließend schnell weiter. Die videobasierten Aktivitätserkennungssysteme setzen dabei anfänglich noch auf bewegungsbasierte Merkmale wie Optischen Fluss und nutzen Neuronale Netze, die zuerst jedes Einzelbild mit einem Convolutional Neuronal Network und anschließend die Sequenz mit einem Rekurrenten Netz auswerten. Erfolgreicher war die direkte Betrachtung eines Videos als Block von gestapelten Einzelbildern und die Auswertung mit Dreidimensionalen Convolutional Neuronal Networks. Verfahren, die auf der Körperpose aufbauen verzichteten auf Grund der Struktur der Eingabedaten häufig auf Convolutional Neuronal Networks und setzten direkt auf Rekurrente Netze. Schwierigkeiten bereitete dabei die erfolgreiche Modellierung der Zusammenhänge zwischen einzelnen Gelenken. Mit der späteren Einführung graphbasierter Neuronaler Netze konnten aber auch hier weitere Fortschritte erzielt werden. Ähnliche Verfahren werden auch im System des Fraunhofer IOSB zur Schätzung der Nebentätigkeiten des Fahrers genutzt. Bei der Erfassung von Nebentätigkeiten im Fahrzeug gab es zur Projektlaufzeit weniger Fortschritt. Hauptgrund hierfür ist vermutlich die kaum vorhandene Datenbasis. Vereinzelt wurden aktuelle Verfahren im Fahrzeug angewendet, aber deren Qualität ließ sich auf Grund der Datenlage schwer beurteilen. Hier leistet PAKoS mit der Veröffentlichung des Drive&Act-Datensatzes einen großen Beitrag sowohl das CVHCI-Lab als auch das Fraunhofer IOSB konnten mit diesem Datensatz zeigen, dass aktuelle Verfahren erfolgreich für den Fahrzeuginnenraum entwickelt und evaluiert werden können.

6 Veröffentlichung der Ergebnisse

- Veröffentlichungen:
  - AT Automatisierung [Ludwig.2018]
  - VDI Mensch Maschine Mobilität [Martin.VDI.2019]
  - International Conference on Computer Vision (ICCV) [Martin.ICCV.2019]
  - International Conference on intelligent Transportation Systems (ITSC) [Martin.2020]
- Abschlusspräsentation in Pfedsfeld
- Vortrag: Autonomous Vehicle Interior Design & Technology Symposium 2019

7 Literaturangaben

[Ji.2002] Q. Ji and X. Yang, “Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance,” Real-Time Imaging, vol. 8, no. 5, pp. 357–377, 2002.

[Fletcher.2005] L. Fletcher, G. Loy, N. Barnes, and A. Zelinsky, “Correlating driver gaze with the road scene for driver assistance systems,” *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 71–84, 2005.

[Bach.2008] K. M. Bach, M. G. Jäger, M. B. Skov, and N. G. Thomassen, “You can touch, but you can’t look: interacting with in-vehicle systems,” in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pp. 1139 – 1148, 2008.

[Pickering.2007] C. A. Pickering, K. J. Burnham, and M. J. Richardson, “A Research Study of Hand Gesture Recognition Technologies and Applications for Human Vehicle Interaction,” in *Proceedings of the Institution of Engineering and Technology Conference on Automotive Electronics*, pp. 1–15, 2007.

[Ohn-Bar.2014] E. Ohn-Bar, S. Martin, A. Tawari and M. Trivedi, “Head, Eye, and Hand Patterns for Driver Activity Recognition”, Proceedings of IEEE International Conference on Pattern Recognition, 2014.

[Demirdjian.2009] D. Demirdjian and C. Varri, “Driver pose estimation with 3D Time-of-Flight sensor,” in *Proceedings of the IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems*, pp. 16–22, 2009.

[Tran.2009] Cuong Tran and M. M. Trivedi, “Introducing ’XMOB’: Extremity Movement Observation Framework for Upper Body Pose Tracking in 3D,” in *Proceedings of IEEE International Symposium on Multimedia*, 2009.

[Holte.2012] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, “Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, 2012.

[Ito.2008] T. Ito and T. Kanade, “Predicting driver operations inside vehicles,” in *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–6, 2008.

[InCarIn.2014] https://www.technik-zum-menschen-bringen.de/projekte/incarin

[Arun.2012] S. Arun, K. Sundaraj, and M. Murugappan, “Driver inattention detection methods: A review,” in *Proceedings of the IEEE Conference on Sustainable Utilization and Development in Engineering and Technology*, pp. 1–6, 2012.

[Kaplan.2015] S. Kaplan, M. A. Guvensan, A. G. Yavuz and Y. Karalurt, “Driver Behavior Analysis for Safe Driving: A Survey”, IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, 2015.

[Petermann-Stock.2013] I. Petermann-Stock, L. Hackenberg, T. Muhr and C. Mergl, “Wie lange braucht der Fahrer? – Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt,” in *6. Tagung Fahrerassistenz*, München 2013.

[Rybok.2011] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen, “The kit robo-kitchen data set for the evaluation of view-based activity recognition systems,” in *Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 128–133, 2011.

[Wang.2011] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176, 2011.

[Sawhney.2013] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos, “Recognizing activities via bag of words for attribute dynamics,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2587–2594, 2013.

[Vemulapalli.2013] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595, 2013.

[Ludwig.2018] J. Ludwig, M. Martin et al., “Driver observation and shared vehicle control: supporting the driver on the way back into the control loop,” *at - Automatisierungstechnik*, vol. 66, no. 2, pp. 146–159, 2018

[Martin.VDI.2019] M. Martin et al., “Innenraumbeobachtung für die kooperative Übergabe zwischen hochautomatisierten Fahrzeugen und Fahrer,” *Der (Mit-)Fahrer im 21. Jahrhundert? 10. VDI Fachtagung Mensch-Maschine-Mobilität*, pp. 67–78, 2019.

[Martin.ICCV.2019] M. Martin et al., “Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2801–2810, 2019

[Martin.2020] M.Martin et al., “Dynamic Interaction Graphs for Driver Activity Recognition”, in *Proceedings of the International Conference on Intelligent Transportation Systems (ITSC)*, 2020

[Cao.2018] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *arXiv:1812.08008 [cs]*, 2018

[Lin.2014] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV*, pp. 740–755, 2014

[Andriluka.2014] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D Human Pose Estimation: New Benchmark and State of the Art Analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014

[Carreira.2017] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, 2017

[Shahroudy.2016] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016

[Kipf.2017] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” *arXiv:1609.02907 [cs, stat]*, 2017

[Yan.2018]S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” *presented at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018

[Huang.2017] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017

[Howard.2019] A. Howard et al., “Searching for MobileNetV3,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019

[Simonyan.2015] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556 [cs]*, 2015

[He.2016] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Abschlussbericht

3. Titel
Forschungsprojekt PAKoS: Personalisierte, adaptive kooperative Systeme für automatisierte Fahrzeuge

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens 31. Dezember 2019
Dipl. Inform. Manuel Martin Dr.-Ing. Michael Voit	6. Veröffentlichungsdatum Juni 2020
	7. Form der Publikation Bericht
8. Durchführende Institution(en) (Name, Adresse) Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB Fraunhoferstraße 1 76131 Karlsruhe	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 16SV7678
	11. Seitenzahl 39
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 32
	14. Tabellen 5
	15. Abbildungen 11

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

18. Kurzfassung
Dies ist der Abschlussbericht zu den Arbeiten des Fraunhofer IOSB im BMBF-geförderten Forschungsprojekt PAKoS - Personalisierte, adaptive kooperative Systeme für automatisierte Fahrzeuge.
Hauptaufgabe des Fraunhofer IOSB war die Entwicklung und Integration von videobasierten Innenraumbeobachtungssystemen, um damit den von Projektpartnern entwickelten Kooperationsmanager und die kooperativen Übergabesysteme zu unterstützen. Hierzu wurde im ersten Schritt ein komplexes Mehrkamerasystem für den Fahrzeuginnenraum konzipiert, was auch die Entwicklung eines Nahinfrarotsystems mit Aktivbeleuchtung mit einschloss. Anschließend wurde mit diesem System im Fahrsimulator des Fraunhofer IOSB eine Studie durchgeführt, um Trainingsdaten für die Entwicklung von Algorithmen zur Nebentätigkeitserfassung bei automatisierter Fahrt zu sammeln. Dieser Datensatz wurde auch veröffentlicht und international vorgestellt. Auf Basis dieser Daten fand im Anschluss die Entwicklung von Verfahren zur Körperposenerfassung, Objekterkennung und Nebentätigkeitserkennung statt. Hierbei zeichnen sich die Systeme des Fraunhofer IOSB durch einen hierarchischen Aufbau aus. Im ersten Schritt werden die Körperbewegungen und Objekte in der Umgebung erfasst und im zweiten Schritt diese Daten genutzt um auf Nebentätigkeiten zu schließen. Zur Integration dieser Systeme, sowohl im Fahrsimulator des Fraunhofer IOSB, als auch im Versuchsträger des Projekts, entwarf das Fraunhofer IOSB ein verteiltes System auf Basis des Open Source Frameworks „Robot Operating System“ (ROS). Hier wurden durch das Fraunhofer IOSB auch weitere Verfahren von Projektpartnern eingebunden.

19. Schlagwörter

20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final Report
3. title Forschungsprojekt PAKoS: Personalisierte, adaptive kooperative Systeme für automatisierte Fahrzeuge	
4. author(s) (family name, first name(s)) Dipl. Inform. Manuel Martin Dr.-Ing. Michael Voit	5. end of project December 31, 2019
	6. publication date June 2020
	7. form of publication Report
8. performing organization(s) (name, address) Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB Fraunhoferstraße 1 76131 Karlsruhe	9. originator's report no.
	10. reference no. 16SV7678
	11. no. of pages 39
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references 32
	14. no. of tables 5
	15. no. of figures 11
16. supplementary notes	
17. presented at (title, place, date)	

18. abstract

This is the final report on the work of the Fraunhofer IOSB in the BMBF-funded research project PAKoS – personalized, adaptive cooperative systems for automated vehicles.

Main focus of the Fraunhofer IOSB was the development and integration of video based interior monitoring systems to support the cooperation manager and cooperative transition systems developed by project partners. The first step in this process was the development of a multi camera system for the interior of the car. This included the development of a NIR-camera system with active illumination. Afterwards this system was used to conduct a study in the driving simulator of the Fraunhofer IOSB to collect training data for the development of algorithms for secondary activity detection in automated cars. This dataset was also published and presented internationally. Based on the data the Fraunhofer IOSB developed algorithms for body pose estimation, object recognition and secondary activity detection. Compared to other approaches the methods of the Fraunhofer IOSB distinguish themselves by their hierarchical nature. First the driver's body movement and objects in the surroundings are determined and on top of this secondary activities are inferred.

To integrate these systems, both in the driving simulator of the Fraunhofer IOSB and the test platform of the project, the Fraunhofer IOSB developed a distributed system based on the open source framework "Robot Operating System "(ROS). Algorithms of project partners were also integrated into this system by the Fraunhofer IOSB.

19. keywords	
20. publisher	21. price