Abschlussbericht zum Vorhaben

„PräVISION - Methodenentwicklung zur präventiven Steigerung der Arbeitssicherheit an Flurförderzeugen mit Umsetzung eines Assistenzsystems durch Fusion und Analyse von 2D- und 3D-Bilddaten“

(FP-0379)

Laufzeit

01.03.2015 – 31.03.2018

Bericht vom 28.11.2018

Axel Börold, Prof. Dr.-Ing. Michael Freitag, BIBA
Armin Lang, Prof. Dr.-Ing. Willibald A. Günthner, TU München, fml
Dr. Anatoly Sherman, SICK AG
Dr. Joachim Tödter, STILL GmbH
Dr. Hans-Peter Kany, BGHW

Inhaltsverzeichnis

1. Kurzfassung deutsch

Jeder arbeitsbedingte Ausfall von Mitarbeitern bedingt neben persönlichen Schmerzen und Einschränkungen auch einen wirtschaftlichen Schaden durch die fehlende Arbeitskraft und Behandlungskosten für Arbeitgeber und Gesellschaft. Besonders für kleinere und mittelgroße Unternehmen stellt dies ein hohes Risiko dar. Darum gilt es, potenzielle Gefahrensituationen bereits frühzeitig zu erkennen und durch präventive Maßnahmen zu vermeiden.

Die Zielsetzungen des Projektes waren zum einen, grundsätzliche Methoden der Bildanalyse zur Steigerung der Arbeitssicherheit beim Einsatz motorbetriebener Flurförderzeugen aufzuzeigen. Dafür sollten unterschiedliche Methoden entwickelt, evaluiert und hinsichtlich ihrer Wirkung, Zuverlässigkeit und Effizienz gegenübergestellt werden. Zum anderen erfolgte die Entwicklung eines herstellerunabhängigen Demonstrator-Assistenzsystems, mit dem der Nachweis geführt werden kann, dass sich mit 2D- und 3D-Bildverarbeitung die Arbeitssicherheit steigern lässt. Dieses Assistenzsystem ist für beliebige manuell bediente Gabelstapler unterschiedlicher Hersteller nachrüstbar und kann somit branchenübergreifend seinen zukünftigen Einsatz finden. Hierbei wird eine sehr hohe Unabhängigkeit von proprietären, fahrzeugeigenen Sensoren erreicht, deren Integration heute in der Nachrüstung eine Hürde darstellt. Ferner erfolgte die Auslegung der Bildanalysemethoden unter der Berücksichtigung der Vielfältigkeit der intralogistischen Arbeitsumgebung, ohne sich auf ein einziges Umgebungsszenario festzulegen.

Zunächst wurde eine Betrachtung unterschiedlicher Methoden zur Erhöhung der Arbeitssicherheit an motorbetriebenen Gabelstaplern durchgeführt. Im Anschluss wurden auf Basis einer Analyse der vorliegenden Unfalldaten mit Flurförderzeugen kritische Bereiche im Arbeitsbereich des Flurförderzeuges sowie Situationen ermittelt, die häufig Unfälle zu verzeichnen hatten. Auf Basis dieser Analyse wurde im Folgenden das zu entwickelnde Assistenzsystem identifiziert und im Rahmen eines Systementwurfes konzipiert. Neben den technologischen Komponenten für die Sensorik und Auswertealgorithmen wurden zudem unterschiedliche Maßnahmen für die Gestaltung der Schnittstelle zum Fahrer des Flurförderzeuges konzipiert. Für die Simulation der Sensordaten und das Nachstellen von relevanten Testfällen entstand im Projekt ein realitätsnahes Modell einer dynamischen Lagerumgebung. Hier können sich Flurförderzeuge und Personen beliebig bewegen und synthetische Sensordaten von diesen Situationen erzeugt werden. Die Daten wurden automatisch gekennzeichnet und für die Entwicklung der Methoden verwendet. Die Entwicklung der Analysemethoden fokussierte auf zwei unterschiedliche Ansätze, wobei jeweils ein Kollisionserkennungssystem mit einem Modul zur Personendetektion kombiniert wurde. Zum einen wurden klassische Methoden zur Bildverarbeitung hinsichtlich ihrer Eignung für den Anwendungsfall untersucht. Zudem wurden die aktuellen Entwicklungen im Bereich der künstlichen Intelligenz aufgegriffen und Ansätze entwickelt, die auf unterschiedliche Arten von tiefen neuronalen Netzen beruhen. Alle Ansätze wurden innerhalb von Laborbedingungen mit Hilfe von Demonstrator-Systemen evaluiert. Das Projekt schloss mit einem Feldtest innerhalb einer industriellen Arbeitsumgebung ab.

Im ersten Schritt konnten unterschiedliche Arten von Interaktionsmöglichkeiten mit einem Flurförderzeug-Fahrer systematisch analysiert werden. Als Ergebnis wurde ein Konzept für ein taktiles Warnsystem entwickelt. Die Sensor-Simulationssoftware war in der Lage, beliebige logistische Szenarien virtuell nachzustellen und Trainingsdaten zu erzeugen. Die Analyse der klassischen Methoden der Bildverarbeitung zeigte die grundlegende Eignung für den Anwendungsfall, jedoch traten häufig falsche Detektionen auf, die die Akzeptanz eines solchen Systems beim Fahrer verringern könnten. Der Ansatz über Deep Learning-Methoden erzielte sehr gute Ergebnisse in den Labor- und Feldtests hinsichtlich der Erkennungsqualität bei einer sehr geringen Anzahl von Fehldetektionen. Durch das Training mit den simulierten Daten ist der Ansatz zudem unabhängig von einem konkreten Anwendungsszenario und daher auf unterschiedliche Einsatzorte übertragbar. Als Gesamtergebnis ist ein Assistenzsystem entstanden, dass durch Verwendung moderner Sensortechnologien einen kritischen Bereich, bspw. hinter dem Fahrer, zuverlässig nach Personen durchsuchen und den Fahrer dementsprechend warnen kann. Die Nutzung eines taktilen Interaktionskanals führt zudem dazu, dass der Fahrer nicht durch weitere audio-visuelle Interaktionskanäle überlastet wird.

2. Kurzfassung englisch

Any work accident causes in addition to personal pain and disability also economic losses due to work-related downtime and treatment costs for employers and society. This represents a high risk, especially for small and medium-sized companies. Therefore, it is important to identify potential dangerous situations at an early stage and to avoid them by means of preventive measures.

The aim of the project was, on the one hand, to identify principal methods of image analysis to increase occupational safety when using engine-powered industrial trucks. Different methods should be developed, evaluated and compared in terms of their effects, reliability and efficiency. On the other hand, a manufacturer-independent demonstrator assistance system was developed which demonstrated that occupational safety can be increased with 2D and 3D image processing. This assistance system can be retrofitted to any manually operated forklift truck from any manufacturers and can therefore be used in a variety of industries in the future. Here, a very high independence from proprietary, vehicle-own sensors is achieved, whose integration represents a challenge in retrofitting today. Furthermore, the image analysis methods are designed taking into account the diversity of the intralogistic working environments without being limited to a single environmental scenario.

First of all, an analysis of different methods to increase safety at work on motorized industrial trucks was carried out. Subsequently, based on an analysis of the existing accident data with industrial trucks, critical areas in the work area of the industrial truck as well as situations that frequently resulted in accidents were identified. Based on this analysis, the assistance system to be developed was identified and designed as a part of a system design. In addition to the technological components for the sensors and evaluation algorithms, various measures were also made for designing the interface to the driver of the industrial truck. For the simulation of the sensor data and the adjustment of relevant test cases, a realistic model of a dynamic storage environment was developed. Here forklifts and people can move arbitrarily and sensor data from these situations can be simulated. The data is automatically labeled and used to develop the methods. The development of the analysis methods focused on two different approaches, each of them combines a collision detection system with a module for person detection. On the one hand, classical methods for image processing were examined with regard to their suitability for the application. In addition, current developments in the field of artificial intelligence were taken up and approaches were developed based on different types of deep neural networks. All approaches were evaluated within laboratory conditions using demonstrator systems. The last step in the project was a field test within an industrial work environment.

In the first step, different types of interaction options with a truck driver could be systematically analyzed. As a result, a concept for a tactile warning system was developed. The sensor simulation software was able to virtually simulate any logistical scenarios and to generate training data. The analysis of the classical methods of image processing showed the basic suitability for the application, but often false detections that might reduce the acceptance of the driver of such a system. The approach using deep learning methods achieved very good results in the laboratory and field tests with regard to recognition quality with a very small number of misdetections. When trained with the simulated data, the approach is also independent of a specific application scenario and therefore transferable to different locations. The overall result is an assistance system that, by using modern sensor technologies, can reliably monitor a critical area, for instance behind the driver, to detect people and warn the driver accordingly. The use of a tactile interaction channel also means that the driver is not overloaded by an additional audio-visual interaction channel.

3. Problemstellung

Jeder arbeitsbedingte Ausfall von Mitarbeitern bedingt neben persönlichen Schmerzen und Einschränkungen auch einen wirtschaftlichen Schaden durch arbeitsbedingte Ausfälle und Behandlungskosten für Arbeitgeber und Gesellschaft. Besonders für KMU stellt dies ein hohes Risiko dar. Darum gilt es potenzielle Gefahrensituationen bereits frühzeitig zu erkennen und durch präventive Maßnahmen zu vermeiden. Dies trifft insbesondere für den Einsatzbereich von motorbetriebenen Flurförderzeugen wie z. B. Gabelstaplern im innerbetrieblichen Transport zu, der ein hohes Gefährdungspotenzial birgt [Gün-2014]. Auf Grund wachsender Anforderungen an das Transportsystem hinsichtlich Flexibilität und Wandlungsfähigkeit stellt der Gabelstapler aber dennoch unverändert ein wichtiges innerbetriebliches Fördermittel dar [Hei-2006, Sta-2017] und erfordert somit eine besondere Gefährdungsbetrachtung für Mensch und Maschine. Unterschiedliche Statistiken zeigen weltweit deutliche Gefährdungen des Menschen [Sta-2017, Kan-2009, Bos-2009] z. B. auf Grund von Anfahr-, Kipp-, Ladegut- und Absturzunfällen [Wit-2006], wobei in den zu Grunde liegenden Datenbanken lediglich Unfälle mit Personenschaden erfasst werden. Es ist zu erwarten, dass mindestens auf gleichem Niveau Schäden an den Betriebsmitteln entstehen, die unmittelbar einen wirtschaftlichen Schaden darstellen und mittelbar durch die eingeschränkte Funktionsfähigkeit die bereits erfasste Gefährdung für den Menschen erweitern können.

3.1 Bedarf an sensorgestützten Fahrerassistenzsystemen

Entsprechende Gefährdungen begründen sich überwiegend in der Unachtsamkeit des Fahrers und weiterer beteiligter Personen, die z. T. aus zu hoher oder einseitiger Arbeitsbelastung resultiert. Fahrerassistenzsysteme können hier ansetzen, um die Aufmerksamkeit im entscheidenden Moment des Auftretens einer Gefahr zu erhöhen und den Fahrer in seiner Entscheidung zu unterstützen. Hierdurch kann zudem abgesichert werden, dass Verhaltensregeln zur Unfallvermeidung (z. B. aus der Richtlinie 89/655/EWG des Europäischen Rates [Ewg-89] auch befolgt werden. Allerdings ergibt sich im stapler-geführten Transportprozess ebenso ein Gefährdungspotenzial durch eingeschränkte Sicht auf den Interaktionsraum. Konstruktive Maßnahmen am Transportmittel zur Reduktion dieser Ursache haben bereits einen hohen Umsetzungsgrad und sind Bestandteil von Produktbewertungen in Fachmedien [Stw-2014], so dass ein Bedürfnis für weiterführende Lösungen zur Erweiterung der Sicht besteht. Bildgebende Sensorsysteme stellen hier einen vielversprechenden Ansatz dar, weisen bisher aber keine große Verbreitung im angesprochenen Einsatzbereich auf.

Dass entsprechende Maßnahmen zur Senkung des Gefährdungspotenzials ihre Wirkung hinsichtlich reduzierter Arbeitsunfälle aufzeigen ist landläufig anerkannt. So beziffert die Occupational Safety and Health Administration (OSHA) das Potenzial einer Senkung von Flurförderzeug (FFZ)-Unfällen mit Personenschäden mit 70%, wobei von gegenwärtig ca. 110.000 FFZ-Unfällen p.a. in den USA ausgegangen wird [Bos-2009]. So erfolgte bereits auch eine Sensibilisierung der Verantwortlichen im FFZ-Einsatz z. B. durch die Kampagne „Risiko raus“ [Ris-2011]. Allerdings ist grundsätzlich eine einseitige Fokussierung auf die Gefährdung des Menschen zu erkennen [Wit-2006]. Selten wird in entsprechenden Maßnahmen die Gefährdung für die Lagertechnik, das Transportmittel, die Transportgüter und das Gebäude mit eingeschlossen, wie z. B. in der DGUV BGI/GUV-I 5160 für den Einsatz von Flurförderzeugen in Schmalgängen [Dgu-2011] in den Grundzügen zu erkennen ist.

Sensorsysteme zur Fahrerassistenz dienen allerdings nicht nur dazu, Unfallgefahren zu vermeiden, wie sie z. B. in der Unfallverhütungsvorschrift BGV D27 beschrieben sind. Als günstiger Nebeneffekt erhöht sich zudem häufig die Effizienz der Prozessausführung, da der Personalaufwand durch weniger Aufsichtspersonal oder durch eine beschleunigte Ausführung und somit besonders für KMU die Wirtschaftlichkeit des kostenintensiven Lagerbetriebs steigt. Entsprechende Zusatzfunktionen der Sensorsysteme dienen somit der Steuerung und Automatisierung von Arbeitsabläufen. In der Praxis ist ein hoher Bedarf an Lösungen zur sensorbasierten Routenführung und automatischen Be- und Entladung von FFZ zu beobachten. Damit besteht der Wunsch nach wirtschaftlichen Sensorsystemen, die hohe Synergien zwischen Sicherheits- und Automatisierungsfunktionen aufzeigen.

3.2 Offene Herausforderungen in der Umsetzung sensorgestützter Fahrerassistenzsysteme

Mit der Aufgabe der sensorgestützten präventiven Vermeidung von Arbeitsunfällen am motorbetriebenen Flurförderzeug ist eine Vielzahl an Herausforderungen verbunden. Diese beginnen damit, dass die Einsatzumgebung sehr abwechslungsreich und zeitlich instabil ist [Bos-2009], so dass zu detektierende Gefahren laufend andersartige Ausprägungen aufweisen und dennoch durch die gleiche Sensorik sicher erfasst werden müssen. Darüber hinaus ist die konstruktive Gestaltung eines jeden Flurförderzeugs zwischen den Herstellern und den Modellreihen individuell, so dass eine auf Grund des hohen Fahrzeugbestands geforderte nachträgliche Sensorintegration sich häufig aufwendig gestaltet. Aber auch in ihrer Funktion dürfen Sensorsysteme durch den Bediener nicht als Einschränkung wahrgenommen werden oder seine Aufmerksamkeit reduzieren, da am manuell geführten Flurförderzeug der Bediener am Ende immer noch für sein Fahrzeug verantwortlich ist [Bos-2009].

Zwar warnen aktuelle Assistenzsysteme generell vor möglichen Kollisionen, führen jedoch keine Klassifikation des Objektes durch, mit dem eine Kollision droht. Im innerbetrieblichen Materialfluss sind allerdings teilweise auch Kollisionen erwünscht, wie beispielsweise bei der Aufnahme einer beladenen Palette. Daher ist eine Unterscheidung der Objekte im Arbeits-/Gefahrenbereich eines FFZ zwingend erforderlich. Die Klassifikation des Objektes muss dabei unter hohen zeitlichen Anforderungen (echtzeitfähig) durchgeführt werden, da die Arbeitsumgebung durch hohe dynamische Einflüsse (Bewegung des FFZ, statisches Objekte, enge Arbeitsräume, Personen) geprägt ist.

Mit diesem Handlungsbedarf sahen sich die Forschungsinstitute und Industriefirmen dieses Projektvorhabens in zahlreichen weiteren anwendungsnahen Forschungsprojekten konfrontiert, was insbesondere hinsichtlich des Bedarfs an Sicherheitssystemen durch die Berufsgenossenschaft Handel und Warendistribution (BGHW) bestätigt wurde.

4. Forschungszweck/-ziel

Die Zielsetzung des Projekts war zum einen, grundsätzliche Methoden zur Steigerung der Arbeitssicherheit beim Einsatz motorbetriebener FFZ aufzuzeigen. Zum anderen sollte der Nachweis geführt werden, dass sich damit die Arbeitssicherheit durch die Anwendung der 2D- und 3D-Bildverarbeitung in einem Demonstrator-Assistenzsystem steigern lässt. Die Kombination beider Sensortechnologien sowie entsprechender Bildverarbeitungsmethoden ermöglicht eine Zusammenführung der jeweiligen Stärken beider optischer Technologien. Anhand der 2D-Bilddaten können Konturen und Texturen erkannt werden, während 3D-Bilddaten Informationen über räumliche Zusammenhänge bereitstellen. Dadurch kann bspw. der Gefahrenbereich automatisch anhand der räumlichen Informationen der 3D-Bilddaten ohne großen Aufwand segmentiert und anschließend mit den robusten und etablierten Verfahren der 2D-Bildverarbeitung analysiert werden. Abbildung 1 stellt die zum Projektstart geplante Verarbeitungsarchitektur der 2D- und 3D-Bilddaten dar.

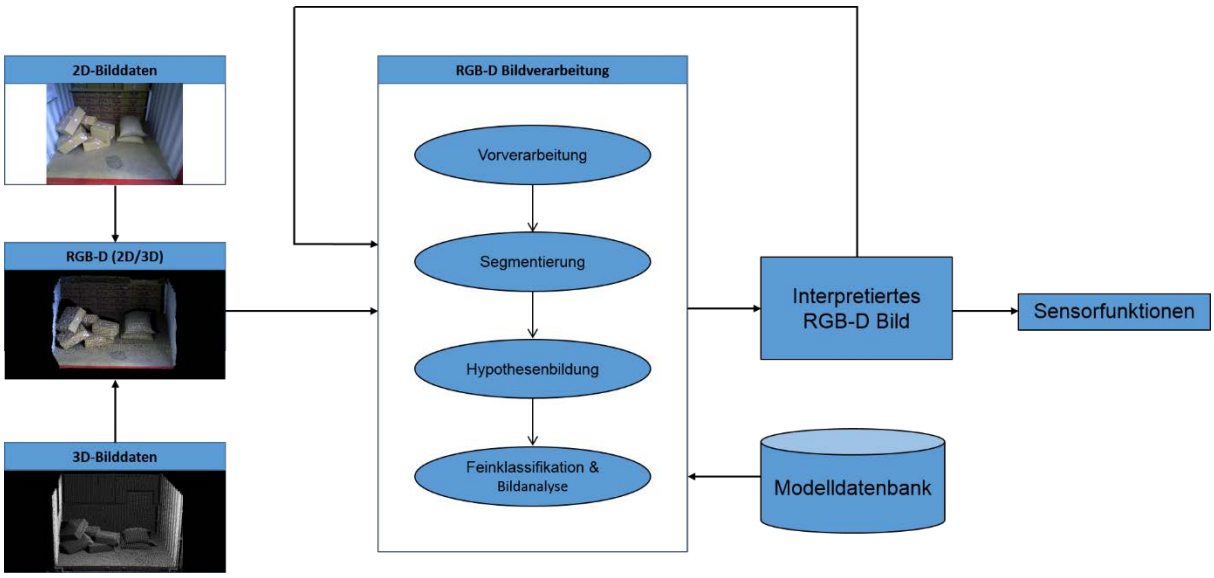

Abbildung 1: Geplantes Zielsystem zum Projektstart

Das erste Ziel des Projektes bestand darin, unterschiedliche Analysemethoden basierend auf den genannten Sensortechnologien zu entwickeln, um die Arbeitssicherheit zu erhöhen. Aufgrund der, parallel zum Projekt verlaufenden, großen Fortschritte insbesondere im Bereich der künstlichen Intelligenz, wurde jedoch von der geplanten Architektur abgewichen. Eine Segmentierung und separate Hypothesenbildung, wie auf Abbildung 1 verdeutlicht, ist z. B. bei der Verwendung von Deep Learning obsolet.

Die Analysemethoden sollten primär für die Detektion von Personen ausgerichtet werden, jedoch auch für die Vermeidung von Betriebsmittelschäden erweitert oder konzeptionell anwendbar sein. Ferner erfolgt die Entwicklung unter der Berücksichtigung des technischen Fortschrittes insbesondere in der optischen 2D- und 3D-bilderfassenden Sensorik. Im Ergebnis sollten unterschiedliche Methoden entwickelt und evaluiert werden, die hinsichtlich ihrer Wirkung, Zuverlässigkeit und Effizienz gegenübergestellt werden.

Mit der zweiten Zielsetzung erfolgt die Entwicklung eines herstellerunabhängigen Assistenzsystems, welches für beliebige FFZ unterschiedlicher Hersteller nachrüstbar ist und somit branchenübergreifend seinen zukünftigen Einsatz finden kann. Hierbei wird eine sehr hohe Unabhängigkeit von proprietären, fahrzeugeigenen Sensoren erreicht, deren Integration heute in der Nachrüstung eine Hürde darstellt. Ferner erfolgt die Auslegung der Analysemethoden unter der Berücksichtigung der Vielfältigkeit der intralogistischen Arbeitsumgebung, ohne eine vorherige Kennzeichnung dieser vorauszusetzen. In der Sensorwahl ist auf ein günstiges Kosten-/Nutzen-Verhältnis der Systemkomponenten zu achten, um einen wirtschaftlichen Einsatz des Assistenzsystems am motorbetriebenen, manuell geführten FFZ zu ermöglichen.

Zudem werden konzeptionelle Perspektiven und Ansätze aufgezeigt, die eine Skalierung des Assistenzsystems hin zu einem Sicherheitssystem ermöglichen. Hierzu werden Wirkungsketten untersucht, um Schwachstellen und Redundanzen im System zu erkennen und mit einem Sicherheitskonzept zu beantworten. Der Fokus der Forschungsarbeit lag auf der Entwicklung softwarebasierter Methoden unter der Verwendung marktüblicher Hardware. Um das Umsetzungspotenzial dieser Lösung zu steigern, schließt das Projektvorhaben mit einem Feldtest ab, welcher weitere konkrete Handlungsschritte für die Entwicklung zu marktfähigen Systemen aufzeigt.

5. Methodik

Durch die Anpassungen bei der Entwicklung der Methoden sowie durch verzögerte Tests unter realen Einsatzbedingungen konnte der ursprüngliche geplante Ablauf der Arbeitspakete nicht eingehalten werden. Die nachfolgende Tabelle stellt den geplanten Ablauf dem tatsächlichen, zeitlichen Projektverlauf gegenüber.

Tabelle 1: Geplanter gegenüber tatsächlichem Projektverlauf

		Projektjahr 1				Projektjahr 2				Projektjahr 3			
Arbeitspakete	Projektquartal	I	II	III	IV	I	II	III	IV	I	II	III	IV
1	Methoden Arbeitssicherheit												
2	Anforderungsanalyse												
3	Simulation												
4	Hypothesenbildung												
5	Bildanalyse												
6	Systemvalidierung												
7	Feldtest												

Arbeitsplan bei Antragstellung Tatsächliche Dauer

Das erste Arbeitspaket umfasste die Analyse unterschiedlicher Methoden zur Erhöhung der Arbeitssicherheit an motorbetriebenen FFZ. Hauptverantwortliche Projektpartner waren hier der Lehrstuhl fml sowie die BGHW. Dieses Arbeitspaket diente zur Identifikation der Funktionen des im Rahmen des Projektes zu entwickelnden Assistenzsystems. Es wurde zunächst die relevanten Gefährdungssituationen recherchiert und analysiert (AP 1.1 – AP 1.3). Unter Berücksichtigung dieser Daten wurden anschließend in einem gemeinsamen Workshop Methoden zur Ursachenvermeidung identifiziert und ein Bewertungskatalog für die Priorisierung dieser Methoden erstellt (AP 1.4 und 1.5). Abschließend wurde durch die Auswertung der Bewertung die umzusetzende Assistenzfunktion identifiziert. (AP 1.6 und 1.7). Dieses Arbeitspaket konnte analog zur ursprünglichen Zeitplanung abgeschlossen werden.

Im zweiten Arbeitspaket wurden für die umzusetzende Assistenzfunktion Anwendungsszenarien und Testfälle für die Systemevaluierung spezifiziert sowie ein Anforderungskatalog erstellt (AP 2.1 – AP 2.3). Ziel des AP2 war die Erstellung eines Systementwurfes in Hardware- und Softwaresicht (AP 2.4) sowie die Recherche der notwendigen Erweiterungen für die Entwicklung eines Sicherheitssystemes (AP. 2.5). Ein weiterer Arbeitsschwerpunkt lag in der Analyse unterschiedlicher technischer Schnittstellen für die Warnung des Fahrers des FFZ durch STILL. Im Arbeitspaket waren alle Projektpartner im ähnlichen Arbeitsumfang beteiligt, wobei das BIBA hier die Hauptverantwortung übernahm. Auch dieses Arbeitspaket konnte analog zur ursprünglichen Zeitplanung umgesetzt werden.

Grundlegende Veränderungen in der Zeitplanung ergaben sich in den folgenden Arbeitspaketen. Im dritten Arbeitspaket sollte die beste Sensorposition bzw. -anordnung simulationsgestützt bestimmt sowie eine Bilddatenmenge für die Entwicklung der Algorithmen erstellt werden. Nach einer genaueren Untersuchung der Sensorik (AP 3.1) wurden Bewertungskriterien für die Anbringung der Sensorik am FFZ erstellt (AP 3.2). Mithilfe einer entwickelten Sensorsimulation wurden die neun Testfälle aus dem ersten Arbeitspaket nachgestellt und verschiedene Sensorkonfigurationen und -ausrichtungen getestet. Insbesondere wurde dabei ermittelt, wie viel Prozent der Körperoberfläche im jeweiligen Testfall über den Zeitverlauf von der Sensoranordnung erfasst wird (AP 3.3). Anhand der Bewertungskriterien und der Simulationen sollte die finale Sensoranordnung ausgewählt und ein Sensorintegrationskonzept erstellt werden (AP 3.4). Aufgrund der Änderungen in den nachfolgenden Arbeitspaketen wurde die Bedeutung der Simulation im Laufe des Projektes insbesondere für die Entwicklung der Analyse-Verfahren

immer stärker und wurde im kompletten Verlauf des Projektes kontinuierlich weiter verfolgt. Hauptverantwortlicher Projektpartner war hier das BIBA, wobei SICK Unterstützung bei der Modellierung und Simulation der einzelnen Tiefensensoren leistete.

Die deutlichsten Änderungen in der Zeitplanung und Durchführung der Arbeitsinhalte ergaben sich in den folgenden Arbeitspaketen 4 und 5. Parallel zu den Arbeiten im Projekt wurden insbesondere im Bereich der künstlichen Intelligenz (Deep Learning) große Fortschritte erzielt. Daher wurde sich im Projekt entschieden, unterschiedliche Ansätze bzgl. der Problemstellung zu fokussieren. Dabei bearbeitete der Lehrstuhl fml den klassischen Ansatz des Shallow Learning (SL) über sogenannte Support Vector Machines (SVM), die selbstdefinierte Bildmerkmale zur Klassifizierung nutzen. Das BIBA untersuchte hingegen neue Methoden des sogenannten Deep Learning (DL). Je nach Art der untersuchten Methode passte daher die Definition und die angedachten Arbeitsaufteilung des Arbeitspaket 4 und Arbeitspaket 5 nicht mehr, da prinzipiell beide Verarbeitungsschritte bei beiden Ansätzen jeweils zusammen betrachtet werden müssen und sich vom grundlegenden Ablauf deutlich unterscheiden. Durch die kontinuierlichen Fortschritte und des Aufkommens im weiterer Netzarchitekturen wurden die Methoden bis zum Ende des Projektes weiterentwickelt.

Alle in den Arbeitspaketen 4 und 5 entwickelten Methoden zur Personen- und Kollisionsprüfung wurden zunächst in einer Laborumgebung, d. h. mit Aufnahmen aus den Versuchshallen des Lehrstuhls fml und BIBA getestet und separat innerhalb des sechsten Arbeitspaketes evaluiert. Parallel wurde durch die Projektpartner SICK und STILL eine konzeptionelle Erweiterung des Gesamtkonzeptes im Hinblick auf die Weiterentwicklung zu einem Sicherheitssystem vorgenommen. Die Arbeiten starteten analog zum Zeitplan, wurden jedoch auch noch parallel zur Durchführung der Feldtests weitergeführt.

Im abschließenden Arbeitspaket 7 wurde ein Feldtest bei einer Partnerfirma von STILL durchgeführt, der Blackforxx GmbH. Dieser verzögerte sich aus organisatorischen Gründen sowie durch die kontinuierliche Weiterentwicklung der Methoden, sodass eine kostenneutrale Verlängerung des Projektes bis zum 31.03.2018 beantragt und bewilligt wurde. In der Organisation und Durchführung der Tests waren alle Projektpartner beteiligt.

6. Ergebnisse des Gesamtvorhabens

Die Forschungstätigkeiten in diesem Projektvorhaben verfolgten das Ziel, Software-Methoden und Algorithmen zu erarbeiten, die der präventiven Steigerung der Arbeitssicherheit mittels optischer Sensorik dienen. Hierzu wurde auf offene Standards gesetzt, so dass eine Anpassbarkeit an weitere Flurförderzeuge und eine Übertragbarkeit auf artverwandte Anwendungen in anderen Einsatzbereichen möglich ist. Im Folgenden werden die einzelnen Ergebnisse der Arbeitspakete im Detail beschrieben.

6.1 AP1: Methoden zur Erhöhung der Arbeitssicherheit am motorbetriebenen FFZ

6.1.1 Recherche der Gefährdungssituationen (BGHW)

Die Unterarbeitspakete 1.1 – 1.3 beinhalten die Recherche und Analyse der Gefährdungssituationen für Fahrer und FFZ sowie die anschließende Identifikation der Testfälle.

Aktuelle Zahlen zum Unfallgeschehen mit Flurförderzeugen liefert der jährliche Bericht der Deutschen Gesetzlichen Unfallversicherung (DGUV) über das Arbeitsunfallgeschehen [Sta-2017]. Der zurzeit aktuelle Bericht der DGUV für 2016 listet 12.671 meldepflichtige Unfälle mit Staplern auf. Bei den meldepflichtigen Unfällen wird zu ca. 70 % nicht der Fahrer des Staplers verletzt, sondern ein weiterer Beschäftigter. In 44% der Fälle wird der Beschäftigte vom Stapler angefahren, überfahren oder eingequetscht.

Im Rahmen des Projektes wurde von der BGHW ein Datensatz von ca. 1400 Unfalluntersuchungsberichten eingehend analysiert. Aus der Gesamtmenge wurden 320 Unfalluntersuchungen, die in die Unfallkategorie „Anfahren von Personen“ fallen, separiert. Diese wurden aufgeteilt nach Gegengewichtstapler und Schubmaststapler näher analysiert. Der Unterschied des Schubmaststaplers zum Gegengewichtsstapler ist, dass der Fahrer im Falle des Schubmaststaplers quer zur Fahrtrichtung sitzt (siehe Abbildung 2).

Abbildung 2: Gegengewichtsstapler (links) und Schubmaststapler (rechts, Foto: Jungheinrich AG)

Der Analyse lag folgendes Raster zugrunde:

1. Fuhr der Stapler vorwärts oder rückwärts?
2. Wie schnell war der Stapler als sich der Unfall ereignete?
3. Wie weit war die verletzte Person vom Stapler entfernt, als sie frühestens vom Fahrer hätte erkannt werden können?
4. Welche Körperteile wurden verletzt?

Die Fragen 1 und 4 konnten präzise aus den Unfalluntersuchungsberichten heraus beantwortet werden. Bei den Fragen 2 und 3 wurde die Antwort unter Zuhilfenahme von Erfahrungswissen und der Bildung von diskreten Merkmalen gegeben. Zusammengefasst ergibt sich folgendes Bild:

Bei Schubmaststaplern ist keine valide Auswertung möglich, da die Anzahl der Unfalluntersuchungsberichte relativ gering ist (65) sowie die Vor- und Rückwärtsfahrt nicht eindeutig zuzuordnen sind. Evident ist jedoch, dass

Anfahrunfälle mit Schubmastern eher leichte Unfallfolgen haben. Bei den Gegengewichtstaplern wurden 255 Datensätze ausgewertet. Das Anfahren von Personen oder Hindernissen war hier in 65% der untersuchten Fälle die Hauptunfallkategorie. Es ergaben sich dabei 2 Schwerpunkte.

Erster Schwerpunkt: 55% der Unfälle ereigneten sich bei der Rückwärtsfahrt mit kleiner Geschwindigkeit oder beim Anfahren aus dem Stand, tendenziell sind eher leichte bis mittlere Verletzungen im Fuß- und Beinbereich die Folge. In nur 2% der Fälle wurde eine hohe Geschwindigkeit angegeben, in 51% eine mittlere und in 47% eine niedrige. Die theoretisch erstmalige Sichtbarkeit der Geschädigten liegt zu 77% im Bereich von 1-3m, in 13% waren sie im direkten Umfeld des FFZ.

Zweiter Schwerpunkt: 34% der untersuchten Unfälle passierten während der Vorwärtsfahrt mit mittlerer bis hoher Geschwindigkeit, tendenziell ereigneten sich eher schwere Verletzungen auch im Brust und Kopfbereich. Die Geschädigten waren in 41% der Fälle erstmalig erst in einer Entfernung von nur 3m sichtbar, in 59% weiter als 3m. Die Fälle mit später Sichtbarkeit der Geschädigten z.B. durch Heraustreten aus einem verdeckten Bereich stellen im Zusammenhang mit einer hohen Geschwindigkeit des FFZ deutlich höhere Anforderungen an das Assistenzsystem als im ersten Schwerpunkt.

Ausgehend davon wurden 12 Szenarien entwickelt, die die wesentlichen Gefährdungssituationen beschreiben. Nach Abzug der drei spiegelbildlich identischen Szenarien ergeben sich die in Tabelle 2 aufgelisteten neun Szenarien. Die geplante Handlungsempfehlung konnte aufgrund der parallel zum Projektverlauf weiterentwickelten relevanten Technologien und die damit verbundene Anpassungen der Arbeitsinhalte nicht durchgeführt werden. Allerdings konnten die im Projekt erlangten Kenntnisse bereits jetzt innerhalb der Arbeitstätigkeiten der BGHW verwendet werden.

Tabelle 2: Übersicht über die 9 Unfallszenarien

Nr.	**Fahrtrichtung / Geschwindigkeit**		**V_{rel} Beschäftigter zu Stapler**	**Was tut der Fahrer?**	**Priorität**
1	**Vorwärts**	hoch	im Winkel von 90°	Fährt geradeaus (Transportfahrt)	1
2		**niedrig**	steht in frontal rechts/links (an der Ladung)	Rangierfahrt	2
3			**steht/geht langsam**	Heck schwenkt nach links	1
4			**Stapler: Fährt mit großem Einschlag an**	Heck schwenkt nach rechts	2
5	**Rückwärts**		steht hinter/neben Stapler	Fahrer setzt rückwärts hat vorher **nicht** nach hinten geschaut	1
6			von links oder rechts im 45° oder 90° Winkel	Fahrer setzt rückwärts hat vorher nach hinten geschaut (Beschäftigter im toten Winkel)	1
7			steht seitlich hinter Stapler	Fahrer setzt rückwärts blickt dabei über rechte Schulter	1
8			steht seitlich vor/neben Stapler	Fahrer setzt mit Lenkeinschlag zurück, blickt zurück (Fahrer übersehen)	2
9			läuft seitlich an Stapler vorbei (parallel)	Fahrer setzt mit Lenkeinschlag zurück, blickt zur Gabel/Last	2

6.1.2 Methoden der Ursachenvermeidung und Wirkungsreduktion (FML)

Um zukünftig den in Kapitel 6.1.1 aufgezeigten Unfallgefahren begegnen zu können bedarf es geeigneter Methoden, mit denen sich Unfälle von Personen unter Beteiligung von FFZ vermeiden oder wenigstens in Ihrer Wirkung auf Personen reduzieren lassen. Der Einsatz solcher Methoden soll die Unfallzahlen signifikant reduzieren und ggf. die Schwere der Unfälle abmildern.

Zur Auffindung geeigneter, innovativer Methoden zum Schutz von Personen im Umfeld von Flurförderzeugen wurden im Rahmen des gemeinsamen Workshops am BIBA mittels Kreativitätstechniken eine Analyse unter folgender Fragestellung durchgeführt: „Wie lassen sich die Ursachen der vorgestellten Unfallarten vermeiden oder

die Unfälle in Ihrer Wirkung reduzieren“. Im Rahmen der Bewertungsphase erhielt jeder Teilnehmer 4 Punkte zur Gewichtung der Relevanz einzelner Vorschläge. Die Auswertung der gesammelten Vorschläge zeigt Abbildung 3.

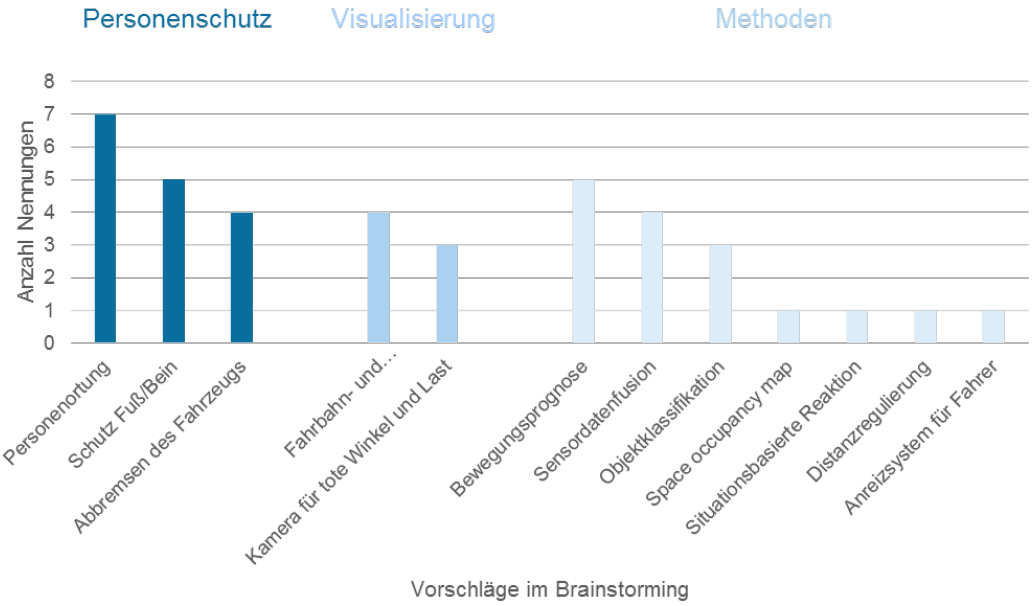

Abbildung 3: Zuordnung der genannten Methoden zu Funktionen

Die Zahl der Vorschläge zum Personenschutz lässt erkennen, dass die Beteiligten dem Schutz von Personen einen hohen Stellenwert einräumen. Dies deckte sich mit dem Ziel des Projektes und bestätigte den gewählten Ansatz. Die Personenortung im Umfeld des Flurförderzeugs wurde am häufigsten genannt (7 Punkte), spezielle Ansätze zum Schutz der unteren Extremitäten erhielten 5 Punkte. Ein Eingriff in die Steuerung zur Unfallvermeidung wurde ebenfalls vorgeschlagen (4 Punkte).

6.1.3 Ableitung und Auswahl einer Assistenzfunktion (FML)

Für den aktiven Schutz von Personen müssen zwangsläufig Kollisionen mit Menschen zuverlässig präventiv erkannt werden, indem Personen, die sich im Arbeits-/Gefahrenbereich des FFZ befinden automatisch detektiert werden. Wenn eine Person erkannt wurde, muss der Fahrer möglichst früh gewarnt werden. Dies verhält sich analog zu autonomen Flurförderzeugen, welche bei etwaigen Kollisionen immer eine Kollision mit einem Menschen annehmen und dementsprechend früh warnen.

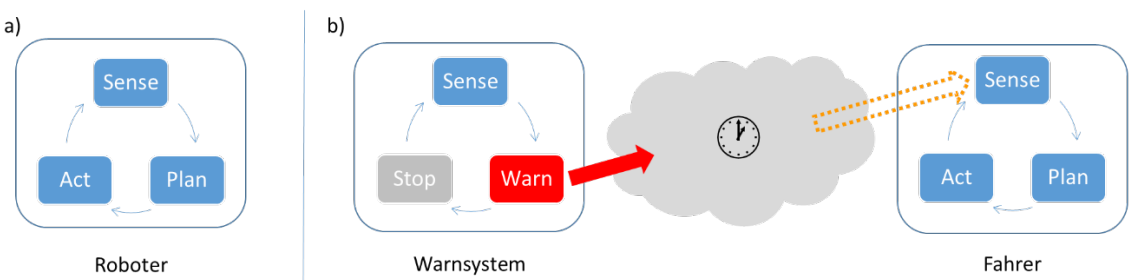

Abbildung 4: Vereinfachte Regelkreisläufe für Roboter und Warnsysteme

Innerhalb autonomer Systeme werden die Sensordaten auf demselben Gerät erhoben, das auch für die Planung zuständig ist. Somit kann das Fahrzeug direkt in Abhängigkeit der Umgebung gesteuert werden. Auch wenn diese Funktionen in separaten Modulen untergebracht sind, findet eine Kommunikation zwischen diesen Modulen ohne größere Verzögerung statt (siehe Abbildung 4, Teil a). In einem Warnsystem ist eine starke Trennung zwischen Sensorik (Warnsystem) sowie dem wesentlichen Plan und der Aktorik (Mensch) inhärent. Das wesentliche Problem ist hierbei die stark verzögerte und unzuverlässige Kommunikation zwischen Warnsystem und Fahrer, die eine sofortige Reaktion durch den Fahrer unmöglich macht. (siehe Abbildung 4, Teil b).

Da der Fokus primär der Schutz von Personen ist, wurde im Projekt daher ein allgemeines Kollisionswarnsystem mit einer Personenerkennung gekoppelt. Auf Basis der Distanzen, Relativbewegungen und des Zustands des Fahrzeugs wird dabei berechnet, mit welchen Bereichen im Fahrzeugumfeld eine Kollision möglich wäre. Dabei wird unterschieden, ob sich in diesen Bereichen Personen befinden oder nicht. Wenn sich in einem der Bereiche eine Person befindet, wird früher gewarnt, als in dem Fall, dass ein Objekt kollisionsgefährdet ist.

6.2 AP2: Anforderungsanalyse und Systementwurf

6.2.1 Anforderungskatalog (FML)

Grundsätzlich soll die Assistenzfunktion den Fahrer eines FFZ in vermeidbaren Unfallsituationen rechtzeitig auf eine Person aufmerksam machen, mit der eine Kollisionsgefahr besteht. Umgekehrt soll in den Fällen, in denen definitiv keine Gefahr einer Kollision besteht auch möglichst keine Warnung erfolgen. Daraus ergeben sich unter Berücksichtigung von Umgebungsbedingungen folgende Anforderungen:

Tabelle 3: Anforderungskatalog für die Assistenzfunktion im Rahmen des Forschungsprojektes

Anforderung	Wertebereich	Begründung
Funktional	Siehe SOLL-Szenarien	
Umgebungsbedingungen	Warenlager (Innenbereich)	Nebel, Regen Schnee erschweren Erkennung erheblich
Verarbeitungszeit	Echtzeit	Rechtzeitige Reaktion des Fahrers notwendig.
Robustheit	Hoch	
Fehlerrate der Personenerkennung	<20%	Einschränkungen durch Stand der Technik
Fehlerrate zur Erkennung der Kollisionsgefahr	<1%	Gemäß Stand der Technik mit vertretbarem Aufwand umzusetzen.

Eine weitere zentrale Anforderung bei der Entwicklung der Methodik bestand darin, dass die Methode und die Qualität der Erkennung unabhängig der intralogistischen Arbeitsumgebung funktionieren.

Die Echtzeitfähigkeit wurde im Projekt auf die Möglichkeiten der Sensorik bezogen. Das heißt, wenn eine Sensorik verwendet wird, welche im 30 Hz Takt Daten liefert, muss das System dazu in der Lage sein alle Berechnungen innerhalb von 33 Millisekunden abzuschließen.

Die Anforderung nach dem Bremsweg ergibt sich aus der Norm DIN ISO 6292, welche unter anderem die maximalen Bremswege für Flurförderzeuge vorschreibt. Zum Vergleich der Daten aus der Norm mit realen Werten wurden für die Geschwindigkeiten von 2 bis 8 km/h Bremstests mit einem unbeladenen Gegengewichtsstapler durchgeführt. Dabei wurde dem Fahrer akustisch der Befehl zum Bremsen gegeben, sobald er eine markierte Linie überfuhr. Die auf diese Weise ermittelten Distanzen können der Abbildung 5 entnommen werden. Dabei ist zu beachten, dass bei den empirisch ermittelten Werten die Reaktionszeit des Fahrers enthalten ist. Die Norm beschreibt den Bremsweg ab der Betätigung des Bremspedals, daher wurde eine Reaktionszeit von einer Sekunde bzw. der entsprechende Weg im Schaubild hinzugefügt.

Nach der Norm ergibt sich somit bei einer Geschwindigkeit von 8 km/h der maximale Bremsweg von 3,2 Metern. Diese Entfernung stellt zugleich die minimal notwendige Reichweite der Sensorik für diese Geschwindigkeit dar. Je nach deren Positionierung und Orientierung werden allerdings höhere Werte notwendig sein.

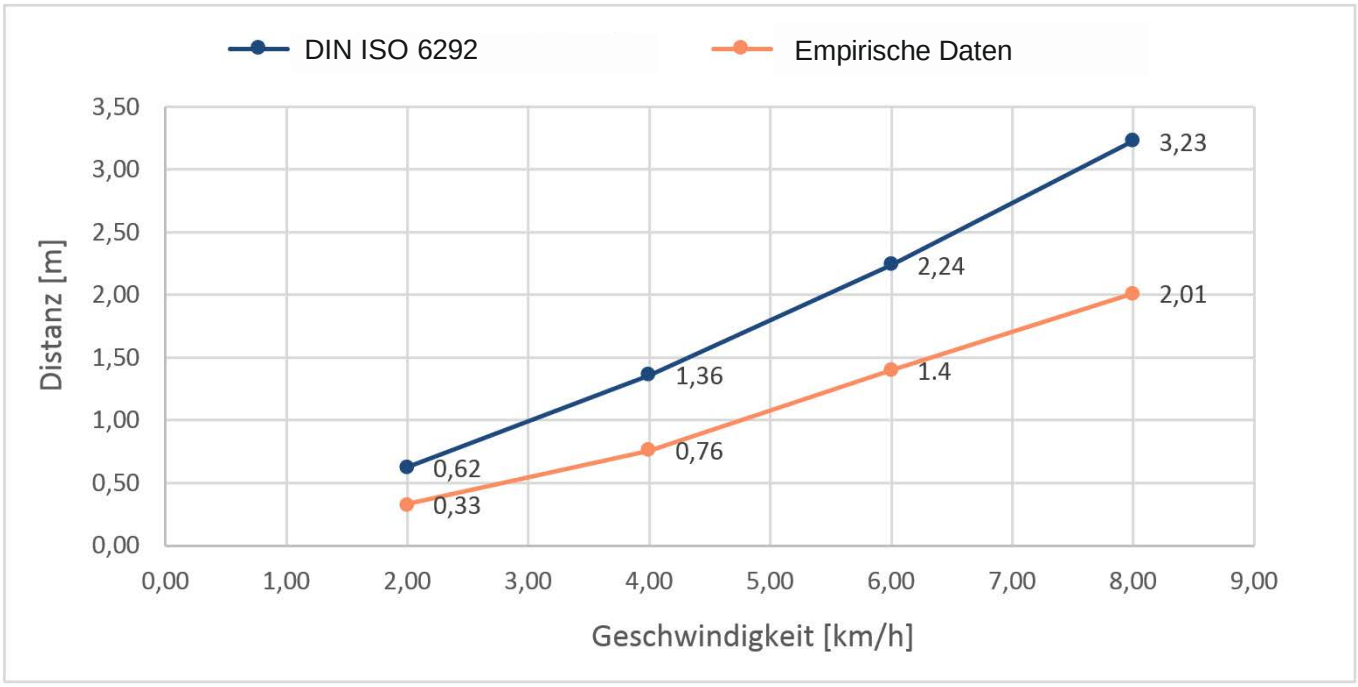

Abbildung 5: Vergleich der Bremswege nach DIN ISO 6292 und empirischen Tests.

6.2.2 Systementwurf (BIBA, FML, STILL)

Der Systementwurf dient als Konzept für den in Kapitel 6.1.3 ausgewählten zweistufigen Warnansatz. Der Entwurf des Gesamtsystems ist der folgenden Abbildung zu entnehmen.

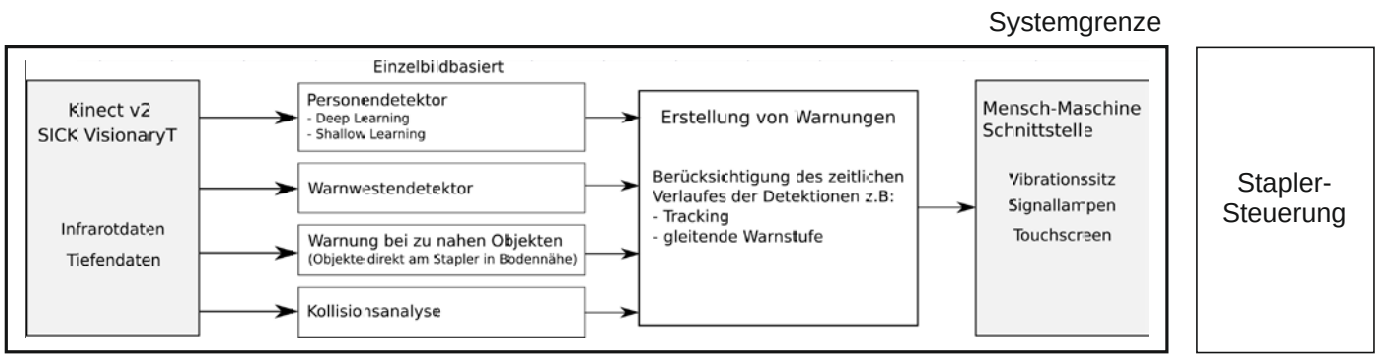

Abbildung 6: Systementwurf

Die von den Sensoren Microsoft Kinect v2 und SICK Visionary-T erzeugten Rohdaten werden softwaretechnisch abgegriffen und für die weitere Verarbeitung aufbereitet. Anschließend werden sie zur parallelen Ausführung an mehrere verschiedene Verarbeitungsmodule weitergeleitet. Diese sollen die Personendetektion durchführen sowie Kollisionen im 3D-Raum erkennen und geben die Ergebnisse für jedes Einzelbild aus. Denkbar ist auch die Integration weiterer einfacher Module, wie z.B. der räumlichen Überwachung der direkten Staplerumgebung oder das direkte Erkennen von Warnwesten. Diese Informationen werden gemeinsam betrachtet und zu Warnhinweisen verarbeitet. Zudem ist es auch denkbar, Personen zu verfolgen (engl.: tracking) oder die Warnung gleitend abschwellend zu realisieren. Ein Alarm würde so nicht sofort verschwinden, wenn z.B. die Person einmal nicht detektiert wird, sondern erst nach einer bestimmten Zeit. Die Kommunikation mit dem Fahrer wird im Demonstrator mit einem Bildschirm realisiert, in der Praxis sind auch einfache Signalmittel wie Lampen oder Vibrationsalarme denkbar.

Die verwendeten Systemkomponenten auf Hardware und Software-Seite sind der nachfolgenden Tabelle zu entnehmen:

Tabelle 4: Systemkomponenten Hardware und Software des Demonstrators

Hardware		Software	
Sensorik	SICK Visionary-T Microsoft Kinect v2	**Betriebssystem**	Ubuntu Linux 16.04 LTS
Recheneinheit	AMD Ryzen5 1600 8 GB RAM NVIDIA 1050Ti GPU 256 GB SSD Systemspeicher 2 x 6 TB HDD Datenspeicher	**Softwarebibliotheken**	OpenCV 3 PCL Caffe 1.0

Die Interaktion mit dem Fahrer findet über die Mensch-Maschine-Schnittstelle statt. Diese soll den Fahrer vor anwesenden Menschen im Arbeits-/Gefahrenbereich warnen. Das für den Demonstrator entwickelte Modul entspricht dabei nicht den möglichen Anforderungen für ein industrielles Produkt, sondern soll Einblicke in die Algorithmen und Datenlage für den Entwicklungs- und Evaluationsprozess geben.

Im Rahmen des Projektes ist ein multimodales (MMS) Anzeige- und Warnsystem für einen Fahrerarbeitsplatz in einem FFZ konzeptionell entstanden, um Fahrern eine gezielte Aufmerksamkeitslenkung auf eine Gefahrensituation zu ermöglichen und eine erleichterte Handlungsentscheidung und Reaktionsauswahl zu unterstützen. Zum anderen wurde für dieses Konzept ein technischer Prototyp mit einer Steuerung und den entsprechenden Anzeigen für nachfolgende Versuche erstellt. Für die Konzeptionierung sind zunächst die Grundlagen wie Definitionen und Fahrzeugtypen beleuchtet worden. Für das Konzept wurden verschiedene Normen zur ergonomischen Gestaltung von Anzeigen und Studien aus dem PKW-Bereich herangezogen und auf Flurförderzeuge angepasst. Bei der Auswahl der Anzeigen wurde vor allem beachtet, dass möglichst mehrere Sinneskanäle beansprucht werden, um im späteren Versuchsverlauf ein geeignetes System für die Fahrzeugserie umzusetzen.

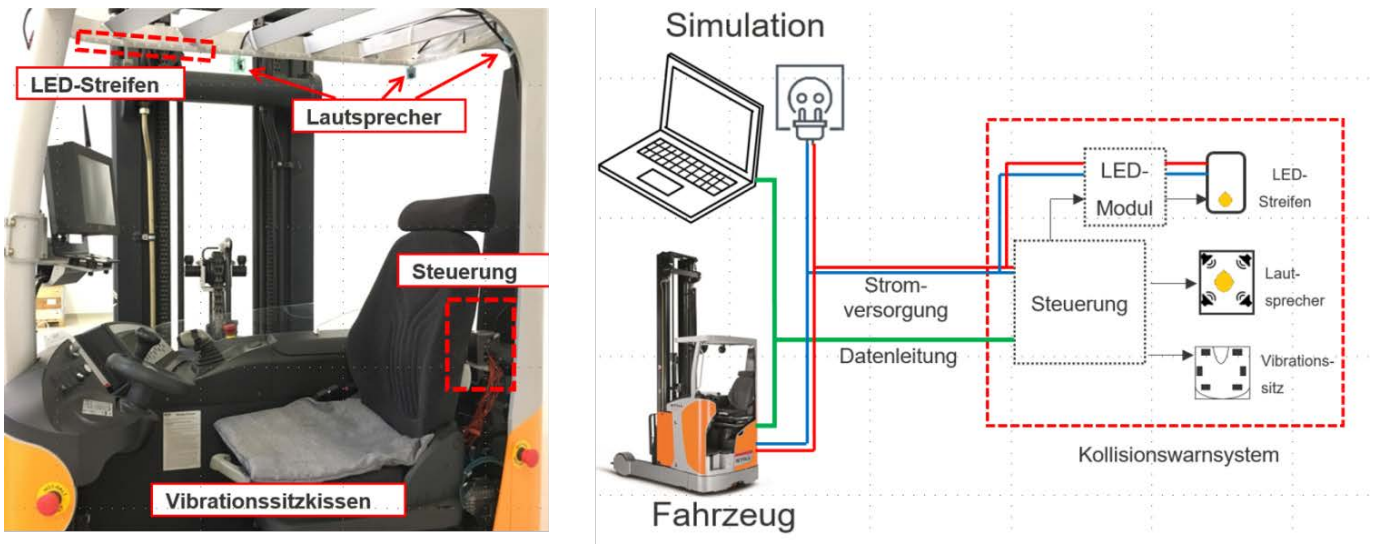

Abbildung 7: MMS-Anzeige- und Warnsystem und Steuerungskonzept

Bei der Recherche hat sich gezeigt, dass es zwar durch mehrere angesprochene Sinneskanäle zu einer besseren Wahrnehmung und schnelleren Reaktion kommen kann, aber in Anbetracht, dass der Fahrer ohnehin hochbelastenden und zeitkritischen Stresssituationen ausgesetzt ist, bedarf es einer innovativen benutzerfreundlichen Anzeige- und Warnlogik. Daher wurde eine MMS-Systematik auf Basis von drei richtungsorientierten Warnanzeigen in vier Kritikalitätsstufen entwickelt, die sämtliche Warnmeldungen logisch miteinander verknüpft. Die MMS-Systematik baut auf visuellen, akustischen und haptischen Sinneswahrnehmungen auf, um dadurch im späteren Versuch verschiedene Konstellationen zu ermöglichen und zu entscheiden, welche dieser Modalitäten sinnvoll für Flurförderzeuge sein können. Die Positionskodierung nach van Gijssel et. al. [Gij-2007] mit acht Segmenten wurde für den PKW-Bereich entwickelt, ob und wie diese Einteilung auch für

Flurförderzeuge hilfreich sein kann bleibt offen, eine Reduzierung auf sechs oder vier Segmente wäre möglich. Die vier Kritikalitätsstufen sind für Flurförderzeuge möglicherweise zu viel, da die Umgebung für Flurförderzeuge gänzlich anders ist als im PKW. Sie bieten aber für den Versuch die Möglichkeit drei Modalitäten aufbauend gestuft zu testen. Hier wäre es möglich die Kritikalitätsstufen auf drei oder zwei zu reduzieren, wenn weniger als drei Modalitäten zusammenwirken. Das visuelle Ergebnis zeigt die Möglichkeit einer LED-Warnanzeige mit der Aufteilung von acht Segmenten im nahen Blickfeld des Fahrers auf.

Die technische Umsetzung konnte, bis auf die Helligkeitseinstellung, bereits im Versuchsstand getestet werden. An dieser Stelle bedarf es jedoch noch weiteren Messungen, in wieweit die Helligkeitsunterschiede umgesetzt werden sollen. Es ist auch möglich die LEDs einzeln und nicht als Segmentaufteilung anzusteuern. Ein großer Nachteil der visuellen Anzeige ist jedoch, dass die Signale nicht omnidirektional wahrnehmbar sind. Daher bietet diese Modalität schlechte Voraussetzungen um Richtungswarnungen überhaupt zu ermöglichen. Omnidirektionale Signale sind dagegen mit akustischen Tönen wahrnehmbar. In der Konzeptgestaltung wurde dazu, für eine Richtungswahrnehmung ab der zweiten Kritikalitätsstufe, das Prioritätsprinzip nach Lundkvist et. al. [Lun-2016] angewendet. Leider konnten mit dieser Anzeige noch keine Tests durchgeführt werden, da dies im Projektrahmen noch nicht weitreichend fortgeschritten ist. Die Lautstärke wurde anhand von wissenschaftlichen Studien festgelegt, hier bedarf es daher noch an Messungen, um den Informationston nach [Lun-2016] entsprechend umzusetzen. Weiter ist der Quadrofonieansatz mit vier Lautsprechern in Frage zu stellen. Lundkvist et. al. [Lun-2016] empfiehlt mindestens sechs (alle 60°), um Richtungshören zu ermöglichen. Die Vibrationsmotoren, um taktile Warnungen zu ermöglichen, konnten in einem Aufbau teilweise getestet werden.

In der Erprobung sollte möglichst darauf geachtet werden wie die Position der Vibrationsmotoren optimiert werden kann, um die vorderen Sensoren auch mit den Unterschenkeln wahrnehmen zu können. Der wenig genutzte Sinneskanal eignet sich im Gegensatz zu den anderen möglicherweise am besten um Richtungswahrnehmungen zu gestalten. Dies muss allerdings in den aufbauenden Versuchen weiter erprobt werden. Aufgrund der zeitlichen Einschränkung des Projektes ist es nicht möglich das gesamte Potential der unterschiedlichen Anzeigen im vollen Umfang zu testen. Neben diesen offenen Fragen und Anregungen zur Optimierung des Anzeige- und Warnkonzeptes sind noch weitere Schritte auf dem Weg hin zur vollständigen Implementierung der MMS-Systematik nötig. Demnach sollte eine Versuchsreihe mit mehreren Probanden durchgeführt werden, um die Anzeigen in einem Flurförderfahrzeug zu testen. Hierzu bedarf es einen Versuchsablauf, der möglichst verschiedene Fahrprozesse und Gefahrensituationen abbildet, um dadurch festzustellen, ob und wie ein solches Warnsystem die Fahrer von Flurförderzeugen unterstützen kann.

6.2.3 Konzeptioneller Entwurf zum Sicherheitssystem (SICK, STILL)

Die im Projektrahmen genutzten Verfahren zur Klassifikation und Extraktion von Personen durch Verbindung von 2D-Kamerabildern und 3D-Tiefenbildern sind zunächst nicht ohne weiteres in ein Sicherheitssystem zu überführen. Dies liegt einerseits an der verwendeten Sensorik, die nicht für sicherheitstechnische Einsätze spezifiziert ist und andererseits an den verwendeten Verfahren zur Beurteilung der Personen im sensorischen Erfassungsbereich. Ein wichtiger Aspekt der im Projekt noch nicht beantwortet werden konnte ist die Kategorisierung der erstellten Algorithmen in eine mittlere Betriebsdauer bis zu einem gefahrbringenden Ausfall oder Versagen (mean time to failure MTTF). Im Folgenden wird erörtert, welche systemischen Änderungen am vorgestellten Assistenzsystem, insbesondere der 3D-Kamera und den verwendeten Methoden, vorgenommen werden müssen, um ein sicheres System, gemäß der entsprechenden Richtlinien, zu schaffen.

Initial muss hierfür die Sicherheitsfrage beantwortet werden bzw. die Sicherheitsfunktion genau definiert werden. Im konkreten Fall wäre das Ziel also ein sicheres Fahrerassistenzsystem an Flurfahrzeugen durch Fusion und Analyse von 2D- und 3D-Bilddaten. Final muss dann z.B. die Erkennung einer Person zu einem Schaltsignal führen. Alle involvierten, systemischen Komponenten, z.B. vom Sensor bis zur Bremse, müssen dabei das geforderte Sicherheitsniveau aufweisen. Letzteres wird über die Konkretisierung der Sicherheitsfrage im Rahmen der

Risikobeurteilung ermittelt. Das erforderliche Sicherheitsniveau wird dabei durch die Anwendung einer Norm ermittelt. Sofern eine passende C-Norm (maschinen-spezifische Norm) existiert, gibt diese das entsprechende Sicherheitsniveau für alle beteiligten Geräte, wie z.B. Sensor, auswertende Logik und Aktoren, vor. Für Sicherheitsfragen bei Flurförderzeugen wäre dies die DIN EN 1525. Diese Richtlinie ist von 1997 und wird daher aktuell überarbeitet, um mit aktuellen Technologien und Anforderungen Schritt zu halten. An ihrer Stelle soll zukünftig dann die EN ISO 3691 stehen. Bis zum ihrem Erscheinen und der Veröffentlichung wird aber empfohlen weiterhin die EN 1525 anzuwenden. Hier muss individuell entschieden werden, ob die pauschalen Annahmen und die daraus abgeleiteten Vorgaben der DIN EN 1525 für die vorliegenden Technologien (3D-Kamera, Machine Learning Algorithmen, usw.) passend sind oder, ob die Risikobeurteilung alternativ durch das Heranziehen der ISO 13849 erfolgt. Die Risikobeurteilung ist hier Folge von logischen Schritten, die die systematische Analyse und Bewertung des Risikos erlaubt. Im Kern müssen dabei die Fragen beantwortet werden: Wie schwer sind mögliche Verletzungen? Wie häufig und wie lange tritt die Gefährdung auf? Welche Möglichkeiten der Vermeidung gibt es? Hierbei werden leichte (reversible) und schwere (irreversible) Verletzungen, selten/kurze und häufig/lange Gefährdungen sowie mögliche und kaum mögliche Begrenzung des Schadens unterschieden. Die Beantwortung der Fragen führt zum erforderlichen Performance Level (PLr, performance level required) der in fünf diskreten Stufen definiert ist (siehe Abbildung 8).

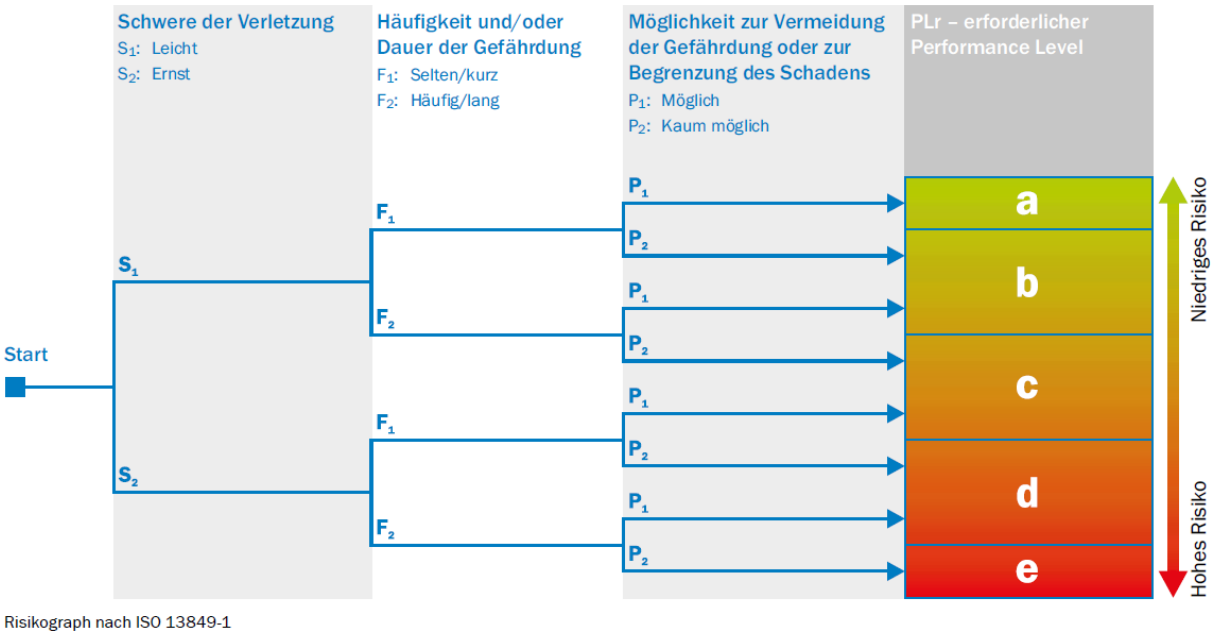

Abbildung 8: Stufeneinteilung Performance Level nach ISO 13849-1

Die Norm stellt dabei sicher, dass der Aufwand der Realisierung in angemessenem Verhältnis zum festgestellten Risiko steht. Der Schutz vor einem kleinen, langsam fahrenden FTS, von dem maximal das Risiko einer Prellung der unteren Extremitäten ausgeht, verlangt eine andere Betrachtungsweise als der Schutz vor einem großen, bemannten Flurförderzeug, welches eine Person tödlich verletzen kann. Ausgangspunkt der Risikobeurteilung ist immer die Maschine ohne Schutzeinrichtung.

Der ermittelte Performance Level gilt dann für alle sicherheitsbezogenen Teile. Im vorliegenden Forschungsprojekt sind dies insbesondere die 3D-Kamera und die auswertende Logikeinheit. Der Performance Level gibt dabei die Anforderungen an die sicherheitsbezogenen Komponenten vor. Dies umfasst Vorgaben an den Entwicklungsprozess (z.B. V-Modell), die Architektur (ein- oder zwei-kanalig, mit oder ohne Testung) und die systematischen Eigenschaften. Diese müssen so sein, dass der Diagnosedeckungsgrad und die mittlere Zeit bis zu einem gefahrbringenden Ausfall entsprechend des benötigten Sicherheitsniveaus sind. Im Falle der 3D-Kamera können die allgemeinen Anforderungen und Prüfungen für berührungslos wirkende Schutzeinrichtungen, wie Sie in DIN EN

61496-3 an diffuse Reflexion nutzende, aktive optoelektronische Sende- und Empfangselemente gestellt werden, herangezogen werden. Hier werden neben Anforderungen an die Mechanik und die elektromagnetische Verträglichkeit insbesondere Vorgaben an die Eigenschaften des optischen Pfades gemacht. Dies umfasst unter anderem die Fragen danach wie das System auf Verschmutzungen, Fremdlicht und Störsender reagiert, aber auch wie mit Manipulationen, teilweisem Ausfall des Bildsensors und, im Falle von 3D Time-of-Flight, mit Mehrwegsreflexionen umgegangen wird.

Schon heute kann die im Projekt eingesetzte 3D-Kamera *Visionary-T* von SICK mit entsprechender software-programmierbarer Sicherheitssteuerung (SICK FlexiSoft) kombiniert werden und damit Performance Level C erreichen. Müssen höhere Performance Level erreicht werden, gelten die oben dargelegten Anforderungen.

6.3 AP3: Simulative Bestimmung der Sensorauswahl und Generierung von Testbildern

6.3.1 Sensorevaluierung (BIBA)

Um diese geplanten Aufgaben umzusetzen, wurden zunächst genauere Untersuchungen der im Projekt betrachteten Sensoren (*SICK Visionary-T* und *Microsoft Kinect v2*) durchgeführt und die wichtigsten Sensorcharakteristiken bestimmt. Beide Sensoren ermitteln die Tiefendaten nach dem Time-of-Flight Prinzip. Hierbei wird von IR-Leuchtdioden ein im Nahinfrarotspektrum (NIR) moduliertes Licht ausgesandt. Das NIR-Spektrum grenzt an dem sichtbaren roten Licht an, ist für das menschliche Auge jedoch nicht mehr sichtbar. Der Tiefensensor erfasst in diesem Wellenlängenbereich im Gegensatz zu Farbbildsensoren nicht nur pixelweise die Intensität des zurückreflektierten Lichtes, sondern auch die Phasenlage. Ähnlich wie bei einem Radar ist der Sensor durch einen pixelweisen Vergleich zur Ausgangsphase in der Lage eine Abstandsmessung durchzuführen, wodurch das Tiefenbild erzeugt wird. Das Intensitätsbild für sich genommen stellt dabei lediglich ein NIR-Bild dar, wobei die Sensoren die Szene aktiv beleuchten. Damit es nicht zu Verwechslungen mit den Wärmebildsensoren kommt, welche im mittleren (MIR) bis langwelligen Infrarotbereich (FIR) arbeiten, wird das NIR-Bild im Folgenden als Intensitätsbild bezeichnet.

Der Sensor *SICK Visionary-T* benötigt zum Betrieb eine mindestens 100Mbit/s schnelle Ethernet-Verbindung sowie eine Versorgungspannung von 24 Volt. Wenn man von den Bildvorverarbeitungsoperationen absieht, die für das Sensormodul vorgesehen waren, stellt der *SICK Visionary-T* bereits selbst ein Sensormodul dar.

Der Sensor *Kinect 2* als Consumer-Produkt benötigt hingegen einen „Kinect Adapter for Windows“, auf Computerseite einen eigenen USB3.0 Hostcontroller und eine Spannungsversorgung von 12V Gleichspannung bzw. 220V Wechselspannung. Softwareseitig kann zur Anbindung unter Windows und Linux der freie Treiber „libfreenect2“ benutzt werden. Die vom Treiber zur Berechnung des Tiefenbildes benötigte Rechenleistung ist dabei nicht trivial. Praktisch benötigt wird dafür eine Grafikprozessor, der CUDA, OpenCL oder OpenGL unterstützt. Für die angedachten Sensorknoten wurden die ARM-Einplatinencomputer NVIDIA Jetson TK1 Ubuntu/Linux mit CUDA und der Odroid XU3 mit OpenCL getestet. Nach Anpassungen des libfreenect2-Treibers (Vermeidung der Übermittlung der RGB-Daten) konnten mit beiden Platinen die echtzeitfähigen 30 FPS ausgelesen werden. Der Odroid XU3 war jedoch mit dem Auslesen bereits nahezu ausgelastet, während der Jetson TK1 weitere Ressourcen für die Bildverarbeitung besitzt. Auf Grund der weiteren Rechenanforderungen für die Kollisions- und Personenerkennung wurde der Demonstrator durch einen handelsüblichen Computer ersetzt und auf separate Sensorknoten verzichtet.

Es folgte die Erkenntnis, dass prinzipiell beide Sensoren für die Tests der Szenarien nutzbar sind. Die *Kinect v2* ist in gegenwärtiger Konfiguration nicht industrietauglich (USB3.0, Gehäuse) und stellt höhere Rechenanforderungen, da jeweils 10 Teilbilder zu einem Tiefenbild zusammengesetzt werden müssen. Der *Visionary-T* besitzt eine geringere Auflösung, liefert aber bereits verwendbare Daten, sodass für die Datenakquise weniger Rechenleistung notwendig wird. Es wurde untersucht, inwiefern sich der *Visionary-T* und die *Kinect v2* sich gegenseitig beeinflussen. Festgestellt wurde, dass beide Systeme im Demonstrator nebeneinander betrieben werden können sofern sie nicht parallel installiert werden und die Oberflächen aus demselben Winkel beleuchten.

Im Rahmen des AP 3.1 wurde zudem eine Vermessung beider Sensoren durchgeführt. Dafür wurden zwei Teststände aufgebaut und in Betrieb genommen. Der Teststand (vgl. Abbildung 9) ist eine computergesteuerte Linearachse und ermöglicht die automatische Erfassung von Tiefendaten mit Tiefen von bis zu sieben Metern. Der zweite Teststand ist ein automatisch angesteuerter Rotationstisch, mit dem gängige Materialen auf einem Träger gradgenau vermessen wurden. Durch diese Daten konnten Sensormodelle mit Abhängigkeiten der Entfernungen, Materialien und Oberflächenwinkel auf das Rauschen der Tiefen- und Intensitätsdaten erzeugt und damit synthetische Daten generiert werden. Zudem wurden sowohl die Ansteuerung der Anlagen als auch die Auswertung der Daten eigene Programme geschrieben.

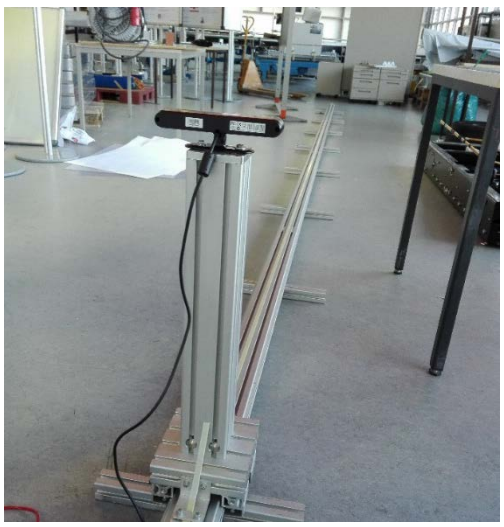

Abbildung 9: Linearteststand

Zu erwähnen ist, dass beide Sensoren starke Artefakte bei im Nahinfrarot-Spektrum stark reflektierenden Materialien wie z. B. Warnwesten erzeugen. Weiterhin treten Verzerrungen im Tiefenbild auf, sobald sich ein Objekt näher als der Mindestabstand zum Sensor befindet. Beide Sachverhalte sind in der Entwicklung des Systems berücksichtigt worden.

6.3.2 Sensordatensimulation (BIBA, SICK, STILL)

Mit den in AP 3.1 gesammelten Daten wurde die Erzeugung synthetischer Sensordaten möglich. Die in diesem Arbeitspaket erstellte Simulation wurde während der folgenden Arbeitspakete iterativ erweitert, sodass neben dem reinen Nachstellen der Testfälle für die Ermittlung der Sensorposition (AP 3.2 und AP 3.3) und dem Testen der Algorithmen (AP 3.5) auch geeignete Daten zum Training dieser Algorithmen erstellt werden konnten (AP 4 und AP 5). Obwohl sich die Arbeiten über 3 Arbeitspakete verteilen, werden aus Gründen der Übersicht alle simulationsbezogenen Arbeiten nachfolgend in diesem Kapitel beschrieben.

Es wurden zwei Simulationen durch Verwendung von Methoden der Computergrafik erstellt. Beide ermöglichen sowohl das Nachstellen dynamischer Testszenarien als auch die Generierung von Trainingsdaten. Die erste Simulation basiert auf der Cycles-Renderer von Blender und ermöglicht das direkte physikalisch korrekte, jedoch langsamere, Rendern von 3D-Szenen. Die zweite Simulation basiert auf einer Grafikengine und rendert über eine polygonbasierte Computergrafik, wie sie in Computerspielen verwendet wird, in Echtzeit Idealbilder. Das erzeugte Idealbild im Farb- und Tiefenbereich wird im Postprocessing durch Hinzufügen der Sensorcharakteristika über ein vom BIBA erstelltes Tool in synthetische Sensordaten umgewandelt. Die Sensorcharakteristika umfassen unter anderem den Sichtbereich, das Sensorrauschen, tiefen- und oberflächenabhängige sowie systematische Fehler.

In beiden Fällen wurde eine komplette Logistikhalle inkl. Infrastruktur wie z.B. Treppen und Türen sowie technischer Gerätschaften wie den Flurförderfahrzeugen und weiterer Produktionsanlagen erstellt.

Die möglichen Variationen im Bereich der Sensorik umfassen weiterhin unter anderem die Art, Anzahl und Anbringungspunkte der Sensoren, das Vorhandensein zusätzlicher Sensorik wie z. B. IMU, Kompass, Lagegeber oder odometrischen Messeinrichtungen. Die virtuellen Personen sind ebenfalls parametrisiert und können unter anderem in Ihren Dimensionen, im Gangverhalten und in der Kleidung variiert werden. Die beteiligten Stapler sind über die CAD-Daten variierbar, ebenso können unterschiedliche Bewegungsmuster verwendet werden. Hinzugefügt wurde zudem die Möglichkeit, gezielt die gesamte Infrastruktur sowie das FFZ inkl. Last auszublenden.

Nachfolgend werden beispielhaft beide Simulationsarten sowie zum Vergleich eine Realszene aus dem BIBA vorgestellt. Sowohl Sensoren als auch Simulationen erzeugen dabei pro „Bild“ bzw. Framesatz mehrere Teilframes, die jeweils unterschiedliche Informationen beinhalten und zur Weiterverarbeitung genutzt werden. Gleichartige Daten, wie z.B. Tiefendaten können dabei je nach Quelle in den Abbildungen unterschiedlich skaliert sein.

Für die Spiele-Engine-basierte Version wurden neben den Ablauf- und Steuerungsskripten auch spezielle Shader- und Exportskripte erstellt, mit denen es möglich ist aus dem laufenden Programm Rohdaten wie das Normalen-Bild im Grafikkartenspeicher zu berechnen und aus diesen in den Hauptspeicher zu exportieren. Die Shader sind Programme, die auf der Grafikkarte laufen und die Farbwerte der Bildpunkte berechnen. Während die Berechnung der Teilbilder selbst mit deutlich mehr als 30 FPS möglich war, wurde der Export der Rohdaten durch die Festplattengeschwindigkeit auf 0.3s für alle 4 Teilframes reduziert.[1] Im Tiefenbild ist die Entfernung der Bildpunkte in Bezug zur Sensorebene dargestellt. Ein absolut schwarzer Punkt steht hier für eine Entfernung von 0 m, ein weißer Punkt für 10 m. Die Helligkeitswerte des Normalen-Bildes beschreiben die Neigung der jeweiligen Oberfläche zur Sensorhauptachse in Grad. Ein hellerer Punkt steht hier für einen geringeren Neigungswinkel. Im Label-Bild sind die Objektklasse und Objektidentifikation farblich gekennzeichnet. Daraus können die Personen- und Palettenlabels berechnet werden, die in der Abbildung 10 im Tiefen – und Farbbild dargestellt werden. Es ist durch die Anwendung eines Sensormodelles möglich, diese idealen Daten in synthetische Sensordaten umzuwandeln.

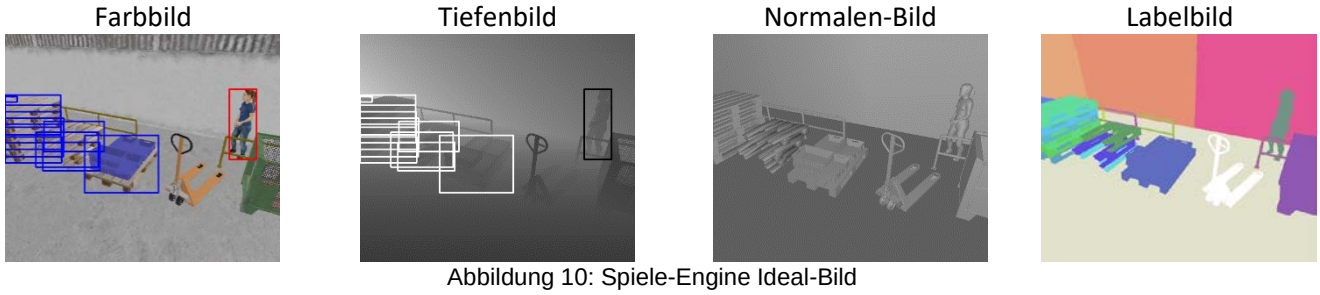

Abbildung 10: Spiele-Engine Ideal-Bild

In Abbildung 11 ist ein gerenderter Framesatz aus dem Blender-Rendering abgebildet, wobei auch hier die Shader für das Tiefen- und Intensitätsbild ebenfalls selbst entwickelt wurden. Im Falle der Abbildung wurden sie auf die Simulation von *Kinect v2* Daten parametrisiert. Die Berechnung der 3 Teilframes hat trotz eines schnelleren Computers unter Inanspruchnahme der Grafikkarte 59 Sekunden gedauert, also ca. 200-fach länger als die durch die Spiele-Engine erzeugten Bilder.[2]

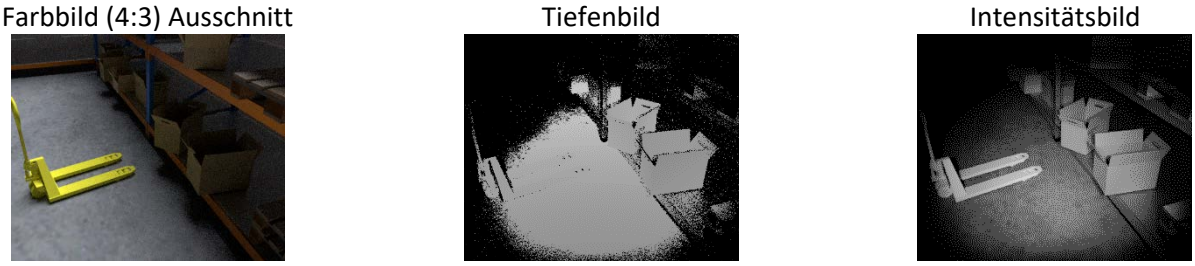

Abbildung 11: Blender Cycles Renderer

[1]Intel-i7 3770K, Nvidia Grafikkarte 980TI
[2]Intel-i7 5830K, Nvidia Grafikkarte 980TI

Nachfolgend ist eine reale, vergleichbare Szene im BIBA dargestellt, die mit Hilfe eines Sensors auf einem FFZ aufgezeichnet wurde. Der Farbbildsensor besitzt ein größeres Seitenformat von 16:9. Aus Darstellungsgründen wird hier nur der zum Tiefenbild korrespondierende Abschnitt gezeigt.

Farbbild (4:3) Ausschnitt | Tiefenbild | Intensitätsbild

Abbildung 12: Kinect v2 Beispielszene

In Abbildung 13 ist zum Vergleich dieselbe Szene, aufgenommen vom *Visionary-T*, dargestellt. Der Sensor bietet kein Farbbild, jedoch ein Confidence-Bild, der ein Vertraulichkeitswert für die Messung darstellt. Ein höherer Intensitätswert ergibt eine höhere Vertraulichkeit.

Confidence-Bild | Tiefenbild | Intensitätsbild

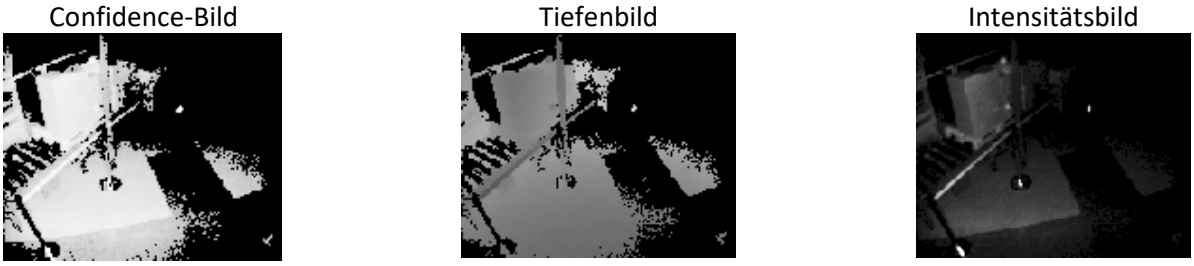

Abbildung 13: Visionary-T Beispielszene

Die zweite Anwendung der Simulation ist die Generierung von gelabelten Trainingsdaten, d.h. synthetischen Sensordaten, in denen die Personen automatisch durch einen umschließenden Rahmen (engl.: „Bounding Box“) markiert werden. Um auch die winkelabhängigen Einflüsse auf die Sensortechnik zu simulieren, wurde die Simulation um die Ausgabe eines Normalen- sowie Materialbildes erweitert. Dadurch wird die pixelweise Berechnung des Einfallswinkels des Sichtstrahles auf die Oberfläche sowie die Bestimmung des jeweiligen Materiales ermöglicht. Auf anisotrope (richtungsabhängige) Reflexionseigenschaften, wie sie z.B. bei gebürsteten Oberflächen auftreten, wurde keine Rücksicht genommen, da in der Praxis nahezu alle Flächen lackiert oder beschichtet sind. Um geeignete Daten für die Sensorsimulation zu generieren, wurde hierfür am BIBA eine automatische Rotationseinrichtung inkl. zugehöriger Auswertesoftware erstellt, mit der die Oberflächen bezogen auf die Intensität und Standardabweichung der Entfernung gradgenau vermessen werden können. Der gesamte Prozess für die Generierung synthetischer Daten ist schematisch in Abbildung 14 dargestellt.

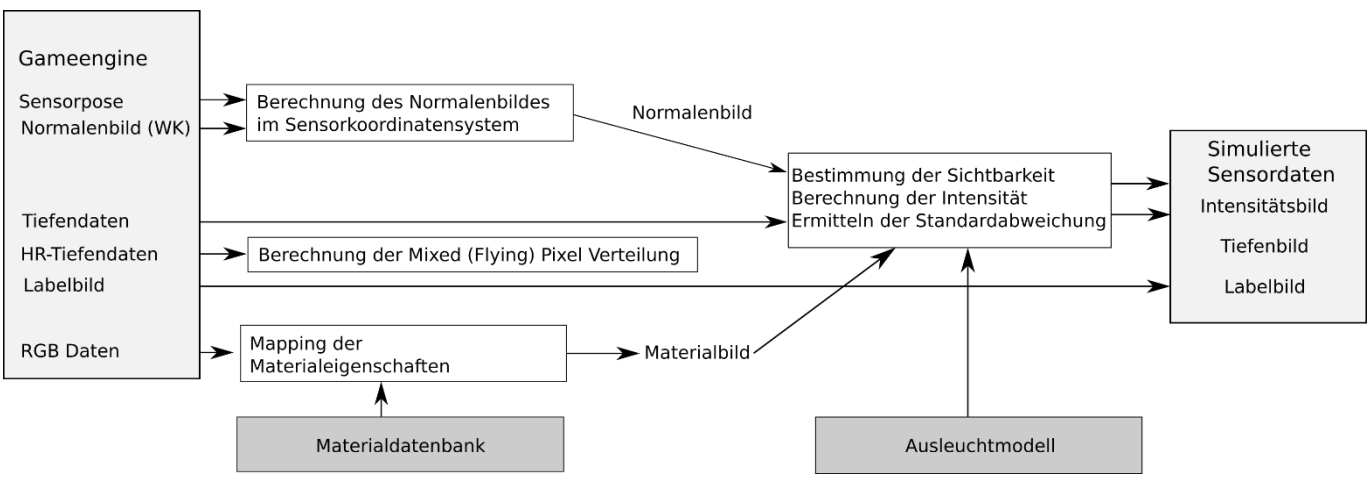

Abbildung 14: Pipeline für die Erzeugung synthetischer Sensordaten

Die Generierung synthetischer Testdaten ist für viele Bereiche der Entwicklung vorteilhaft. Es können so dynamische Szenarien abgebildet werden, die in der Realität aus Gründen der Arbeitssicherheit nicht durchführbar sind. Durch die „Ground-Truth“-Daten und Wiederholbarkeit können Algorithmen bereits während ihrer Entwicklung an definierten Szenarien getestet und verbessert werden. Zudem ist eine Automatisierung der Evaluation durch die Variation der Szenen- und Systemparameter innerhalb definierter Testszenarien möglich.

6.3.3 Simulative Bestimmung der optimalen Sensoranordnung (BIBA)

Für die Bestimmung der optimalen Sensoranordnung wurde die Spiele-Engine-basierte Simulation verwendet. Die Daten wurden wie folgt erzeugt: Zunächst wurden die Testfälle simulativ abgebildet. Für jeden Testfall wurden je drei Sensoren front- und heckseitig verwendet. Diese wurden mit den Parametern des *Visionary-T* sowie der *Kinect v2* gerendert. Für die Ermittlung der Verdeckungen wurde jedes Frame in drei Varianten berechnet, einmal ohne Verdeckungen, einmal mit Verdeckungen der Person durch die Infrastruktur und einmal mit Verdeckungen von der Infrastruktur sowie vom Gabelstapler selbst. (hauptsächlich der Mast in Frontrichtung).

In diesen Daten wurde ausgewertet, in welchen Anteilen die Körperteile in den Testfällen von dem Stapler oder der Infrastruktur verdeckt werden. Die Szenarien wurden zur Ausrichtung der Sensorneigung mit vier verschiedenen Winkeleinstellungen berechnet. Ein Beispielframe aus dem ersten Testfall ist in Abbildung 15 a) dargestellt. Das Diagramm in Abbildung 15 b) stellt die anteilige Sichtbarkeit der Person, sowie die Verdeckung durch das FFZ und der Infrastruktur dar. Die maximale Verdeckung im Frontbereich bei dem Einsatz von zwei Sensoren konnte so durch Kombination der beiden Sensoren ermittelt werden. Eine Anbringung am Mast wurde als zusätzliche Option im Rahmen des Projektes weiter untersucht.

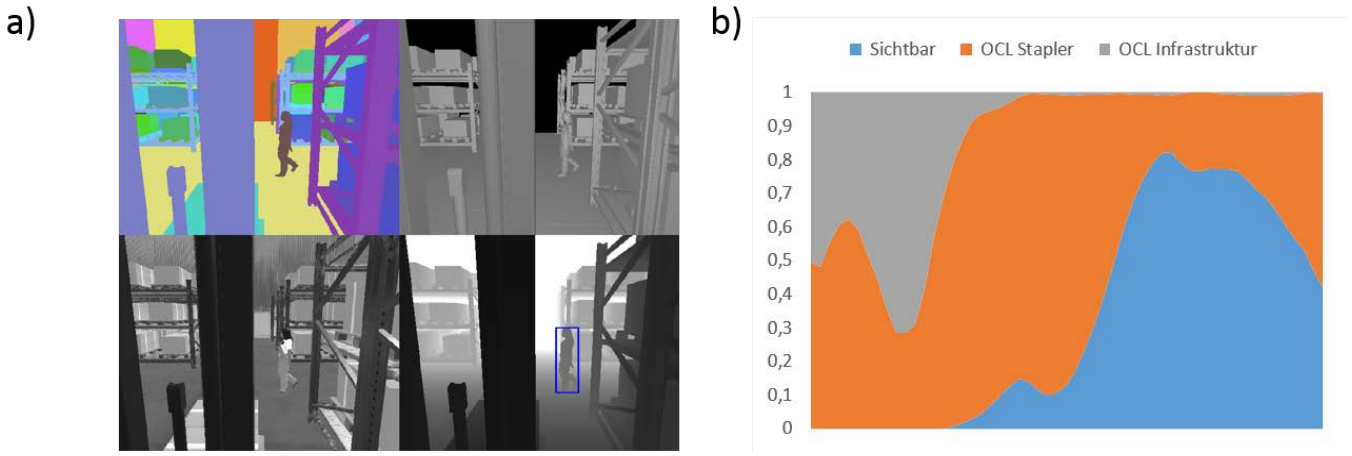

Abbildung 15: a) Testszene b) Auswertung relative Sichtbarkeit

Zwei Aspekte sprachen jedoch für die Nichtverwendung der hier erzeugten Ergebnisse. In der von der BGHW durchgeführten Unfalldatenauswertung wurde ermittelt, dass in 123 Unfällen in rückwärtiger Richtung die Geschädigten auf der linken Seite hinter dem FFZ standen, dagegen gab es jedoch nur 10 Unfälle auf der rechten Seite. Erklären könnte man das mit der deutlich schlechteren Sichtbarkeit der Geschädigten für den Fahrer, dessen Sitz im Falle des FFZ in der Regel auf der linken Seite verbaut ist. Diesen Fakt in die Auswertung zu integrieren würde auch simulativ dazu führen, dass genau dieser Bereich im Sinne der Unfallvermeidung besser abgedeckt ist. In Abstimmung mit dem Projektbegleitkreis wurde mit Hinblick auf Anpassung des Menschen an das Assistenzsystem sich jedoch stark für ein symmetrisches System ohne Vorzugsrichtung ausgesprochen. Die theoretisch möglichen Sensorpositionen und Ausrichtungen wurden dadurch stark eingeschränkt. Vorstellbar waren somit nur ein mittig positionierter Sensor oder mehrere gleichmäßig angeordnete Sensoren.

Der zweite Grund wurde mit der Entwicklung der Algorithmen deutlich. Für die Ermittlung der optimalen Kameraposition müssen Annahmen über die Fähigkeiten des Algorithmus aus dieser Perspektive getroffen werden. Da diese sich basierend auf den gemachten Erfahrungen je nach verwendeten Detektionsalgorithmus und Training

unterscheiden, kann die aus Erkennungssicht optimale Anbringung nicht im Vorfeld entschieden werden. Folgende unbekannte Aspekte spielen für die Erkennung eine größere Rolle:

- Welche Körperpartien werden zuverlässiger erkannt?
- Wie viele Pixel sind zur robusten Klassifizierung notwendig?
- Wie groß darf die Verdeckung einzelner Körperteile maximal sein?
- Wie groß muss insgesamt der Anteil des Körpers sein, der sich im Sensorbild befindet?

Es wurde daher unter Berücksichtigung dieser Gründe sowie den Erkenntnissen aus der Sensorevaluierung eine sinnvolle rückwärtige Sensoranordnung ausgewählt. Die entwickelten Methoden sind jedoch so ausgelegt, dass diese prinzipielle unabhängig von der Sensorposition mit Ausnahme der Verdeckungen durch den Mast funktioniert.

Die Annahmen dazu sind folgende:

- Aufgrund der notwendigen Mindestentfernung müssen die Sensoren auf dem Dach des Fahrzeuges installiert werden. Nahe Objekte und der Boden selbst würden die Erfassung stören.
- Die Sensorik muss die Personen, die sich direkt vor dem Stapler befinden, nicht erkennen – sondern nur als Objekt registrieren. In diesem Fall würde unabhängig von der Personenerkennung durch die Kollisionsprüfung gewarnt werden.
- Bei dem Einsatz von zwei Sensoren sollte der Überlappungsbereich so groß sein, dass eine Person nicht geteilt wird. Aus der Sensorevaluierung ist bekannt, dass beide Sensoren in den Randbereichen weniger gut ausleuchten.

6.3.4 Generierung von Trainings- und Testdaten (FML, BIBA)

Für die Überprüfung von Methoden zur kamerabasierten Kollisionserkennung werden zwangsläufig reale Sensordaten benötigt. Es wurden für die Evaluierung der Methoden auf zwei Arten Daten erzeugt: gespeicherte Sensordaten von der Mircosoft Kinect v2 und dem SICK Visionary-T und computergenerierte Daten aus der Simulation. Letztere ermöglichten es, Unfallszenarien mit Menschen zu testen, welche im Realfall nicht erprobt werden können. Die Speicherung realer Daten hat den Vorteil, dass sensorbedingte Bildfehler wie z. B. Rauschen in den Testdaten enthalten sind. Damit durch die Speicherung keine Informationen verloren gehen, wurden die Daten im Rohformat gespeichert.

Zur Evaluation der Kollisions- und Personenerkennung wurden folgende Aufnahmen gemacht:

- Freie Fahrten in der Versuchshalle des Lehrstuhls fml mit Personen
- Kollisionen mit Kartonagen und Schaufensterpuppe in der Versuchshalle des Lehrstuhls fml
- Freie Fahrten in der Versuchshalle des BIBA mit Personen
- Tagesgeschäft im Lager der Still GmbH
- Testfälle und freie Fahrten im Lager der Blackforxx GmbH
- Tagesgeschäft im Lager der Blackforxx GmbH

6.4 AP4+5: Methoden zur Hypothesenbildung und Bildanalyse

Um Kollisionen am Gabelstapler mittels kamerabasierter Technik erkennen zu können, sind mehrere Methoden aus dem Bereich des maschinellen Sehens nötig. Im Folgenden wird dargelegt, welche Methoden der Bildverarbeitung

zum Empfang von Kameradaten, zur Vorverarbeitung und zur Personen- sowie Kollisionserkennung eingesetzt werden können.

6.4.1 Ansatz über Support Vector Machines (FML)

6.4.1.1 Empfang und Vorbereitung der Kameradaten

Die Hardware des konzeptionierten Systems besteht aus einem Computer und einem Microsoft Kinect v2 Sensor. Letztere ist die einzige Hardwarekomponente, die an das System angebunden werden muss. Aufgrund der Festlegung auf die Programmiersprache C++ im Systementwurf können alle Systemkomponenten als Module oder Bibliotheken eingebunden werden.

6.4.1.2 Kalibrierung

Die Kalibrierung der verschiedenen Bildtypen entfällt bei dem SICK Visionary-T Sensor, da die Intensitäts- und Tiefenbilder vom gleichen Sensor geliefert werden und dementsprechend bereits zueinander kalibriert sind. Im Falle der Kinect v2 von Microsoft kann das Farbbild über eine bereits in der Bibliothek vorhandenen Funktion auf das Tiefenbild kalibriert werden. Dieses Bild wird im Folgenden als „registriertes Farbbild“ bezeichnet (siehe Abbildung 16). Diese Kalibrierung unterliegt allerdings unterschiedlichen Fehlern. Zum einen hat der Farbsensor eine andere Position als der Tiefensensor. Zum anderen besitzen die Sensoren unterschiedliche Linsen, aus denen in Kombination mit dem Sensor unterschiedliche Öffnungswinkel resultieren. Letztlich enthält das registrierte Farbbild vor allem an den Konturen Ungenauigkeiten. Dazu sei angemerkt, dass durch die Fusion von Tiefen- und Farbbild an den Stellen, an denen kein Tiefenwert vorliegt, der entsprechende Pixel im registrierten Farbbild fehlt. Aufgrund dieser Fehler wurde eine eigene Methode „preprocessing::fitBGR2ir“ zur Kalibrierung des Farbbilds auf das Intensitäts-/Tiefenbild erstellt.

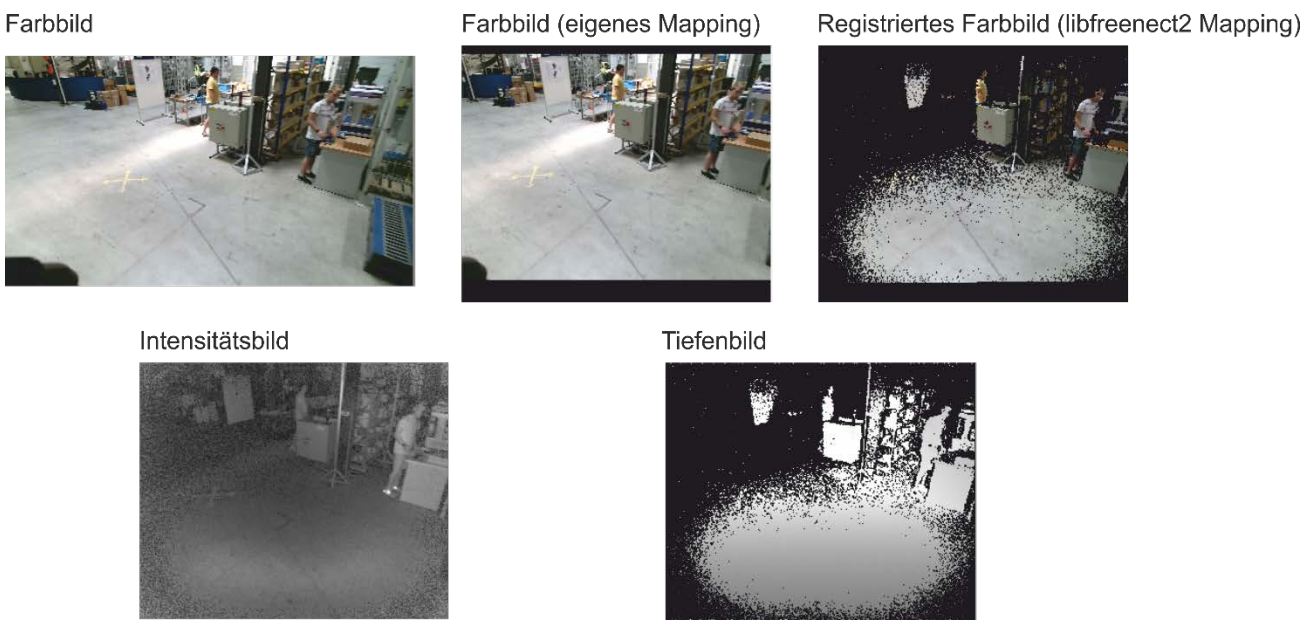

Abbildung 16: Beispiele der verschiedenen Aufnahmekanäle der Microsoft Kinect v2

6.4.1.3 Vorverarbeitung

Algorithmen im Bereich des maschinellen Sehens sind üblicherweise auf Farb-, Graustufen- oder Tiefenbilder ausgelegt. Im Projekt wurde unter anderem auch das Intensitätsbild als Eingangsbild verwendet. Zur Verwendung musste dieses erst vorverarbeitet werden, damit es einem Graustufenbild ähnlich ist. Ansonsten wäre eine Anpassung der Algorithmen die Folge gewesen. Zur optimalen Funktionsweise der nachfolgenden Personen- und Kollisionserkennung sollten allerdings auch die anderen Bildarten vorverarbeitet werden.

Je nach Bildtyp wurden folgende Vorverarbeitungsschritte verwendet:

Skalierung:	Verkleinern/Vergrößern des Bilds. (F/RF/I/T[3])
Farbkanaländerung:	Verminderung der Farbkanäle, z. B. von Farbbild zu Graustufenbild. (F/RF)
Farbwertänderung:	Anpassung der Farbwerte, z. B. Erhöhung der Helligkeit. (F/RF/I/T)
Normalisierung:	Spezielle Form der Farbwertänderung zur Egalisierung der Abstände aller Farbwerte. (F/RF/I/T)
(Selektiver) Gauß-Filter:	Weichzeichnen des Bildes um z. B. Rauschen zu verringern (unter Beibehaltung der Kanten). (F/RF/I/T)
Interpolation leerer Bereiche:	Berechnung kleiner Pixelbereiche durch Mittelwertbildung umliegender Pixel. (T)
Devignettierung:	Filterung der Randabschattung, die bei Kamerabildern zu den Ecken hin zu nimmt. (I)

6.4.1.4 Segmentierung – Hypothesenbildung

Ein weiterer Teil der Vorverarbeitung ist die Zerlegung eines Bildes in Teilbereiche. Spezielle Methoden sind zur Segmentierung oftmals nicht nötig. Es können z. B. bestimmte Farben im HSV[4]-Farbraum über die Filterung maximaler oder minimaler Pixelwerte erreicht werden. Für das Kollisionswarnsystem sind die Entfernung des Bodens und/oder Hintergrunds sowie das Bilden von Clustern vorgesehen. Hierzu müssen die Tiefendaten allerdings vorher in das Koordinatensystem des Staplers transformiert werden.

Das Entfernen des Bodens ist für das Clustern notwendig, da ansonsten alle Objekte über den Boden miteinander verbunden wären und dementsprechend immer ein Cluster bilden würden. Die Entfernung des Hintergrunds soll eine Minderung von Falscherkennungen der Personendetektion bewirken.

Zur sinnvollen Anwendung von Clustering-Verfahren ist eingangs die Entfernung des Bodens im Tiefenbild notwendig. Über den Boden sind alle Objekte miteinander verbunden, wodurch die verschiedenen Clustering-Verfahren Probleme hätten, verschiedene Objekte zu finden. Die Entfernung des Bodens kann nur realisiert werden, wenn die Höhe und Winkel der Kamera bekannt sind. Damit diese Werte nicht manuell gemessen werden müssen, wurde folgendes Verfahren verwendet:

[3] F: Farbbild, RF: Registriertes Farbbild, I: Intensitätsbild, T: Tiefenbild)
[4] Kanäle im HSV-Farbraum: Farbwert, Farbsättigung, Helligkeit

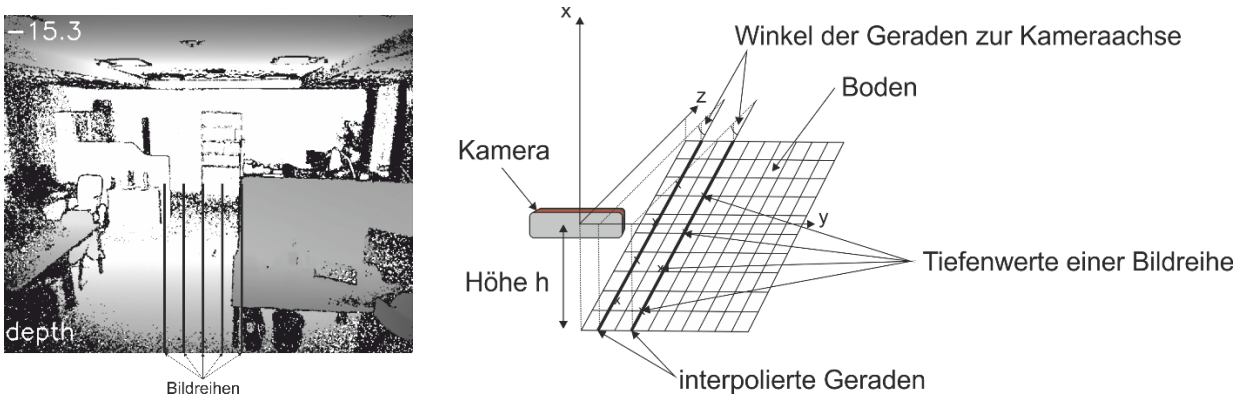

Abbildung 17: Ausgewählte Bildreihen im Tiefenbild (links) und Schema zur Berechnung von Winkel und Höhe der Kamera (rechts).

Es wurden fünf Geraden im zentralen unteren Bereich des Tiefenbildes interpoliert (siehe Abbildung 17 links). Dieser Bereich wurde gewählt, da er im Fahrweg des Staplers liegt und demzufolge meist frei von Gegenständen ist. Anhand der interpolierten Geraden wurden jeweils über geometrische Beziehungen die Höhen und die Winkel zur Kameraachse berechnet (siehe Abbildung 17 rechts). Invalide Werte (Winkel über 90°, Höhen unter 1 bzw. 3,5 Meter), die auftraten, wenn sich z. B. ein Hindernis in diesem Bereich befand, wurden gefiltert. Ebenso wurden Winkel und Höhen, die vom Median der fünf Geraden zu stark abwichen, gefiltert. Wenn mindestens drei valide Werte für Höhe und Winkel in einem Tiefenbild erkannt wurden, wurde der Median der validen Werte wie folgt verwendet: Durch die Kenntnis des Winkels der Kamera zum Boden konnte das Tiefenbild um den berechneten Winkel gedreht werden. Anschließend wurde die Höhe subtrahiert, sodass die Werte des Tiefenbilds im Koordinatensystem des Staplers vorlagen. In diesem Koordinatensystem wurde schließlich eine Maske für alle Pixel generiert, deren Höhe unter 20 cm lag. Anschließend konnte diese Maske auf das eigentliche Tiefenbild multipliziert werden. Dadurch werden alle Pixel im Tiefenbild, welche eine maximale Höhe von 20 cm zum Boden haben, entfernt.

Das genutzte Verfahren wurde zu Beginn nur zu Testzwecken implementiert. Da es aber sehr robust funktionierte und sehr geringe Rechenzeiten aufwies, wurde es bis zum Ende des Projekts für die Anwendung der Shallow-Learning Algorithmen verwendet.

Cluster, d. h. zusammenhängende Pixelbereiche, sind für die Erstellung von Objekthypothesen vorgesehen. Diese können für nachfolgende Objekterkennungsalgorithmen Suchbereiche darstellen. Das Clustering kann entweder im 2½-D Bild [5] oder in der Punktewolke durchgeführt werden.

Bezüglich des 2½-D-Clusterings wurden der „k-means“-Algorithmus [Ste-1956] und die Wasserscheidentransformation [Vin-1991] implementiert, da diese häufig im Bereich der Bildverarbeitung genutzt werden. Der k-means-Algorithmus findet Cluster über die Abweichung zum allgemeinen Mittelwert zugehörige Datenpunkte, allerdings muss die Anzahl zu bildender Cluster vorgegeben werden. Dieses Problem besteht bei der Wasserscheidentransformation nicht. Im übertragenen Sinn wird hierbei aus den Bildwerten ein Relief (Höhenprofil) erstellt, welches solange mit Wasser gefüllt wird bis zwischen den Wasserflächen eine bestimmte Dicke der Trennfläche unterschritten wird. Da das Tiefenbild bereits einem Höhenprofil entspricht, muss das Tiefenbild für diese Methode nicht vorbereitet werden.

6.4.1.5 Personendetektion

Die Verfahren der maschinellen Personendetektion können nach *Benenson et al.* In drei Kategorien untergliedert werden [Ben-2014]:

- Entscheidungsbäume (Decision Forests)
- Support Vector Machines (SVM) – Deformable Part Models (DPM)

[5] Matrix welche die Auflösung des Tiefensensors hat und die Tiefenwerte für jeden Pixel enthält

- Tiefe Netze (Deep Learning, Deep Neural Networks, Convolutional Nets)

Die Vorteile der „Decision Forests“ und „Support Vector Machines“ liegen in ihrem geringen Rechenaufwand. Dafür ist die Erkennungsrate moderat und die Anzahl der Falscherkennungen zum Teil sehr hoch. Deutlich bessere Erkennungsraten liefern „Deep Neural Networks“, allerdings sind diese rechenaufwändiger [Red-2016]. Da ein optimales Verfahren aufgrund der unterschiedlichen Vor- und Nachteile ohne Voruntersuchungen anhand der Beschreibungen aus Veröffentlichungen nicht bestimmt werden konnte, wurden Verfahren aus dem Bereich SVM und Deep Neural Networks untersucht. Entscheidungsbäume eignen sich im Allgemeinen nur zur Klassifikation, d. h. zur Unterscheidung von mehreren Objektklassen, und nicht zur Detektion. Sie wurden daher nicht weiter untersucht.

6.4.1.6 Support Vector Machines

Prinzipiell wird beim Training einer SVM eine Hyperebene berechnet, welche Objekte mindestens zweier Klassen voneinander trennt und den größtmöglichen Abstand zu den jeweiligen Datenpunkten besitzt. Die Ebene wird aufgespannt durch die Supportvektoren. Nach dem Training werden letztere gespeichert und beinhalten die Informationen zur Klassifizierung von Bildern. In Verbindung mit der Sliding-Window-Methode können Objekte auch lokalisiert werden. SVM gehören zu den flachen Maschinenlernmethoden. Bei diesen müssen die zu betrachtenden Datenmerkmale (engl.: features) vorgegeben werden. Sogenannte HOG-Features haben sich für die Personenerkennung etabliert. Diese repräsentieren die Wahrscheinlichkeitsverteilung von Kantenarten im Bild. [Dal-2006]

6.4.2 Ansatz über Deep Learning Methoden (BIBA)

Im Gegensatz zu Shallow-Learning Methoden, also z.B. der SVM mit HOG Merkmalen, erstellen Deep-Learning Methoden ihre Merkmale selbst. Daher wird grundsätzlich angestrebt, die Trainingsdaten möglichst roh in das Netz zu geben. Vorteile sind, dass wesentlich mehr Merkmale betrachtet werden können und dass das Netz eine große Anzahl von Variationsmöglichkeiten lernt, sowie das Wissen generalisieren kann. Der Hauptaufwand liegt in der Konzeption und Bereitstellung der Datenmengen und der Durchführung des Trainings der Netze.

Im BIBA wurden im Bereich der Personendetektion aus RGB-D-Daten durch- Deep-Learning-Verfahren verschiedene Ansätze untersucht. Anzumerken sind, dass die echtzeitfähigen Deep-Learning Methoden erst während des Projektzeitraumes veröffentlicht wurden und dass die synthetischen Trainingsdaten nicht sofort verfügbar waren.

Im ersten Ansatz wurden verschiedene bestehende 2D-Klassifikationsnetze in Zusammenhang mit einer selbst entwickelten Segmentierung zur Hypothesenbildung verwendet. In den folgenden Ansätzen wurden dagegen verschiedene aktuelle echtzeitfähige Architekturen verwendet, welche Personen ohne vorhergehende Hypothesenbildung detektieren – in den letzteren auch unter Anwendung synthetischer Trainingsdaten.

6.4.2.1 DL-Ansatz 1: Personendetektion durch Segmentierung, Deep-Learning zur Klassifizierung

Der erste Ansatz wurde durchgeführt, um die generelle Eignung von Deep Learning für Echtzeitanwendungen zu überprüfen. Die untersuchten Netze waren lediglich in der Lage, Bilder in bestimmte Klassen zu unterscheiden. Für eine Lokalisierung / Detektion wurde eine vorgeschaltete Hypothesenbildung verwendet.

Zunächst wurde eine Reduktion der Eingangsdaten durchgeführt. Verwendet werden hierzu die Farb- und Tiefendaten, die pro Sensor in eine Punktwolke umgerechnet werden. Um die Anzahl der Raumpunkte entfernungsmäßig zu harmonisieren (nähere zum Sensor gelegene Bereiche besitzen eine höhere Punktdichte), wurde ein sogenannter *Voxel-Grid*-Filter mit der Kantenlänge von drei Zentimetern angewendet. Bei diesem werden alle Punkte in eines 3cm Kubus zu einem Punkt zusammengefasst und deren Anzahl so auch auf ca. ein Fünftel reduziert, was die weitere Verarbeitung beschleunigt. Anschließend werden die Bodenebene und zu hoch liegende Punkte aus der Punktwolke entfernt, was je nach Anzahl der verbleibenden Objekte eine starke Reduktion der

Bildpunkte zur Folge hat. Die verbleibenden Punkte werden durch ein euklidisches Clustering segmentiert. Cluster mit zu wenigen Punkten, die zu klein für eine Person sind, werden dabei entfernt. Zu große Cluster werden, basierend auf ihrer maximalen Ausdehnung in der Höhe, in mehrere Cluster unterteilt. Somit werden Personen von etwaigen Gegenständen getrennt. Die verbleibenden Cluster werden als Hypothesen übernommen. Diese Schritte wurden parametrisch so schwach eingestellt, dass in jedem Fall die Menschen als Hypothesen abgebildet wurden. Es erfolgt die anschließende Rückprojektion jeder Hypothese in die 2D-Daten, die dadurch aus dem RGB- und Tiefenbild extrahiert werden können. Die extrahierten Bildbereiche werden zusammen auf eine quadratische Form von 256x256 Pixeln skaliert und anschließend durch das neuronales Netz klassifiziert.

6.4.2.2 DL Ansatz 2: Personendetektion direkt durch echtzeitfähiges Netz (SSD)

Der zweite Ansatz verwendet das erst im Jahr 2016 erschienene Echtzeit DL-Netz zur Klassifikation und gleichzeitiger Detektion, den „Single Shot Multibox Detector“ (SSD). [Wei-2016] Im Gegensatz zu den bisherigen Netzarchitekturen weist dieses Netz eine Reihe von selbstentwickelten Schichten auf und wurde auf Echtzeitfähigkeit optimiert. Unter anderem wurde hier auf rechenintensive voll vernetzte Schichten verzichtet, zudem findet die Bildskalierung nur im Merkmalsraum statt und es wird (statt einer „Sliding Window“ ähnlichen Methode) nur eine vorgegebene Anzahl von möglichen Bounding Boxes überprüft und diese dann zu einem Treffer kombiniert.

Im Rahmen des Projektes wurde das rein auf Farbdaten ausgelegte Netz auch auf dessen Eignung für die Verwendung von Tiefendaten untersucht. Das Klassifizierungsnetz wurde zudem durch eigene manuell gelabelte Farb- und Tiefendaten nachtrainiert. Das Netz ist vortrainiert, um unter anderem „Personen“ zu erkennen. Neben zusätzlichen Aufnahmen von Personen aus der Staplersicht wurden unter anderem auch die Klassen „Flurförderfahrzeuge“, „Boxen“ und „Paletten“ hinzugefügt. Als Datenbasis für die Flurförderfahrzeuge wurde eine Erstellung von Tiefenaufnahmen von Flurförderfahrzeugen bei der Blackforxx GmbH durchgeführt. Insgesamt sieben verschiedene Staplermodelle aus allen horizontalen Blickwinkeln mit insgesamt 1200 Einzelaufnahmen wurden dem Netz an Trainingsdaten hinzugefügt. Zudem kamen noch Tiefenaufnahmen von Personen, die im BIBA erzeugt wurden. Synthetischen Daten sind dagegen noch nicht in das Training integriert.

6.4.2.3 DL-Ansatz 3 Echtzeitfähiges Netz trainiert mit synthetischen Daten

In den vorherigen Ansätzen wurde direkt oder indirekt das Caffe[6] Framework verwendet. Da für das Projekt auch Aspekte wie die Softwareintegration der Netze und Geschwindigkeit eine Rolle spielen, wurde im nächsten Test auf das von Google erstellte Framework TensorFlow[7] [Goo-2015] basierende TENSORBOX[8] verwendet. TENSORBOX arbeitet vergleichbar zu dem im Ansatz 2 verwendeten SSD Netz, aber verwendet als Basisnetz, welches für die Merkmalsextraktion zuständig ist, anstatt des klassischen VGG16 Netzes [Sim-2014] das GoogleLeNet-Netz [Sze-2015]. Das GoogleLeNet verwendet anstatt klassischer Faltungsschichten sogenannte Inception Module. Diese zeigten (in nicht-Echtzeittests) trotz der weitaus weniger zu lernenden Parameter eine höhere Klassifizierungsrate und eine kürzere Berechnungszeit.

Zunächst wurden die synthetischen Daten, die für die Bestimmung der Sensorposition erstellt wurden, als Trainingsdaten verwendet. Die synthetischen Testdaten wurden zu 100 % richtig erkannt - ein Anzeichen für eine Überanpassung des Netzes ist, d.h. das Netz hat sich die virtuelle Person gemerkt und die Fähigkeit zum Generalisieren auf ungesehene Personen verloren. Im Ergebnis konnte das Netz in den vorhandenen Realaufnahmen aus der BIBA-Halle jedoch nur wenige Menschen erkennen, die sich unnormal bewegt haben. Zudem gab es eine größere Menge an Falscherkennungen von Personen. Untersuchungen haben gezeigt, dass das Netz sehr stark auf Farben reagiert hat und dass die Bewegungsmuster in den Daten ziemlich einseitig waren (Gehen und Stehen). Zudem waren außer den Personen nahezu alle sonstigen Objekte in den synthetischen Daten geradlinig, was eine potentielle Fehlerquelle für das Netz darstellt. Zudem ist klar geworden, dass die Verwendung eines vortrainierten

[6] http://caffe.berkeleyvision.org/
[7] https://www.tensorflow.org/
[8] https://github.com/Russell91/TensorBox

Netzes bei ausschließlichem Training mit synthetischen Daten die Ergebnisse stark verbessert und daher fortan verwendet wird. Ein Blick in die ersten Schichten des Netzes hat gezeigt, dass die ersten Schichten mit synthetischen Daten schlechter gelernt werden, da diese dafür sehr vielfältig sein müssten. Mit der Verwendung eines mit Realdaten vortrainierten Netzes befähigt man das Netz also mit Augen auf die virtuellen Daten zu schauen, die jedoch auf die Realität trainiert wurden. Die Gewichtungen des vortrainierten Netzes wurden dabei durch das Erlernen des ImageNet Datensatzes erzeugt, welches ca. 14 Mio. Farbbilder verschiedenster Kategorien enthält. Auf dieser Basis wurde dann mit den anwendungsbezogenen synthetischen Daten weitertrainiert. Im Ergebnis konnten die Netze dadurch besser von den synthetischen Daten abstrahieren und das gelernte dann in der Realität umsetzen.

Da die C++-API des Framework Tensorflow, welche für die Softwareintegration des Netzes notwendig wäre, die Anforderungen nicht erfüllt hat, wurden in den darauffolgenden Tests auf das in dem Caffe Framework implementierte DetectNet[9] von Nvidia verwendet. DetectNet verwendet als Basis ebenfalls das GoogleLeNet ebenso ohne die letzten vollvernetzten Schichten, und ist damit in Aufbau und in der Funktionsweise mit TENSORBOX vergleichbar.

Die daraufhin durchgeführten Entwicklungsarbeiten bezogen sich im Wesentlichen auf die Analyse und der Verbesserung der synthetischen Daten. Der iterative Workflow sah wie folgt aus:

1. Erzeugen von Datensätzen
2. Trainieren des Netzes
3. Tests mit Realdaten
4. Interpretation der Filter / Ergebnisse
5. Verbesserungen der Simulation und Trainingsparameter

Um den Einfluss der Datenmengen zu bestimmen, wurden in jedem Iterationsschritt jeweils 100, 550, 3450 und 12750 synthetischer Bilder für das Training erzeugt.

Tabelle 5: Anzahl Trainings- und Validierungsbilder

Trainingsbilder	Validierungsbilder
550	100
3450	550
12750	3450

Gleichbleibende Erkennungsraten bei einer Erhöhung der Datenmenge lassen darauf schließen, dass es aus den Daten nichts mehr zu lernen gibt. Im Gegensatz zum Lernen von SVM wird keine Unterteilung in Positiv- und Negativbildern benötigt. Dafür müssen die Personen gelabelt (gekennzeichnet) sein. Während des Trainings sieht das Netz jeweils nur einen Teil des Trainingsbildes; ist eine Person enthalten, wird lediglich der betreffende Part als Positiv gekennzeichnet. Zudem wurde das Netz mit jedem Bild 200-fach trainiert, d.h. es sah 200 mal eine in Farbe, Ausschnitt, Rotation und Spiegelung abgewandelte Form des Ursprungbildes. Dieser Vorgang wird Datenaugmentation genannt.

[9] https://devblogs.nvidia.com/detectnet-deep-neural-network-object-detection-digits/

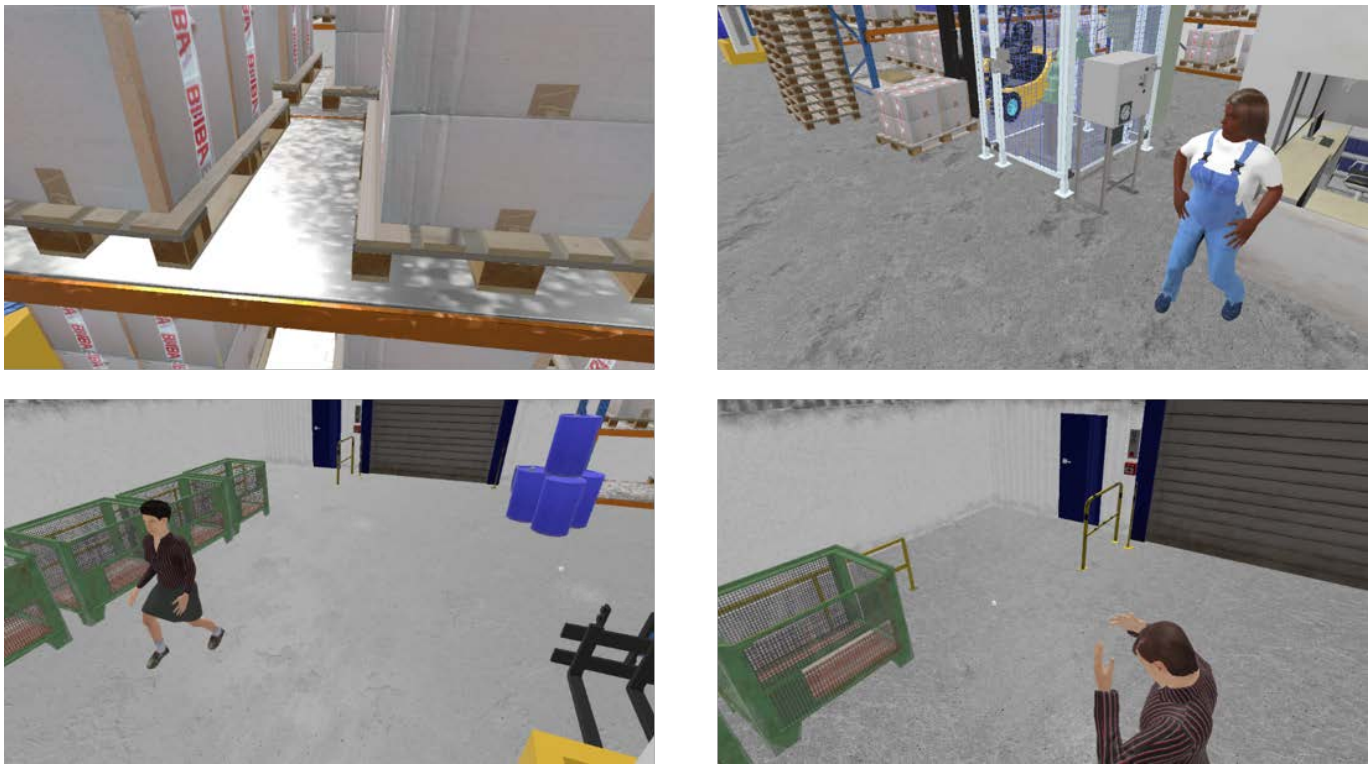

Abbildung 18: Beispieldaten aus der Simulation

Die ersten Iterationen zielten auf eine Erhöhung der Vielfalt an Trainingsdaten mit Personen ab. Es wurden mit der freien Software „MakeHuman“ zehn verschiedene Phänotypen von Personen erstellt. Mit einer gleichzeitigen Erhöhung des Farbrotationsparameters in der Datenaugmentierung konnte der farbliche Einfluss damit komplett eliminiert werden, das Netz achtet demzufolge nur noch auf Kanten, Mustern und darauf aufbauenden Strukturen.

Hiermit konnte das gleiche Netz auch für Intensitätsbilder angewandt werden, die ja letztendlich einem einkanaligen Farbbild entsprechen. Der Vorteil der drei Farbdimensionen liegt aber immer noch in der größeren Anzahl möglicher Kanten, es könnte z.B. eine Kante zwischen einer roten und blauen Fläche detektiert werden, die in einem einkanaligen Bild aufgrund gleicher Grauwerte nicht vorhanden wäre. Im Fall der Microsoft Kinect v2 bietet das Farbbild zudem einen weiteren Öffnungswinkel. Dem Gegenüber ist das Intensitätsbild nicht nur einfarbiger, sondern insbesondere in menschengeschaffenen Umgebungen auch einfacher aufgebaut und damit auch der Simulation ähnlicher. Viele vom Menschen erstellte Muster, insbesondere für Textilien, sind nur im sichtbaren Licht vorhanden. Fenster sind im NIR Spektrum überwiegend undurchlässig. Für die Erkennung von Menschen ist dieser Punkt nur vorteilhaft. Der Hauptvorteil der Verwendung von Intensitätsdaten ist in diesem Fall jedoch, dass diese aus dem ToF-Sensor und damit aus einem aktiven System stammt. Durch die Beleuchtung entsteht eine maximale Belichtungszeit und die Bildaufnahme ist unabhängig von externen Lichtquellen. Im Vergleich zum Farbbild der Microsoft Kinect v2 sind die Unterschiede insbesondere bei schlechten Sichtverhältnissen und Drehungen bemerkbar. (vgl. Abbildung 19) Im Sinne eines stabilen Systems werden daraufhin die Intensitätsbilder für die Erkennung verwendet.

Abbildung 19: Intensitätsbild vs. Farbbild der Microsoft Kinect v2 bei schwachem Licht

Die digitale Halle wurde umgebaut, um mehr verschiedene Objekte in einem Bereich zu konzentrieren. Zudem wurden 50 verschiedene Bewegungsprofile aus der Datenbank der „Carnegie Mellon University Graphics Lab“ ausgewählt und für die Simulation angepasst. Die Bewegungsprofile umfassen dabei unter anderem Laufen, Gehen, Sport und Tanzen. Um zu vermeiden, dass das Netz alle nicht geradlinigen Objekte als Personen erkennt, wurden zudem eine Auswahl kurviger Objekte wie Säcke, Tonnen, Roboter und Produktionsanlagen hinzugefügt. Zudem wurden die zu Training verwendeten Labels besser gefiltert, indem z.B. der sichtbare Anteil der Person und der Füllgrad des Labels bestimmt wurden.

Jeder dieser Schritte führte zu Verbesserungen in der Erkennungsrate, jedoch war beim Test mit Realdaten insbesondere in komplexen Szenen zu sehen, dass das Netz oft Fehlerkennungen von Personen ausgibt. Es mussten dem Netz noch mehr Fehlbeispiele gezeigt werden, die keine Personen enthalten. Ein zweiter gedanklicher Punkt war es, dass nahezu alle simulierten Trainingspersonen auf demselben Untergrund (und damit durch die Draufsicht oft auch Hintergrund) herumlaufen. Das Netz könnte also auf die Idee kommen, sich diesen Untergrund als Merkmal für eine Person zu merken. Damit dies nicht passiert und das Netz sich auf die Person konzentriert wurden prozedural erzeugte Bodentexturen eingesetzt. In jedem einzelnen Trainingsframe ist ein über sieben zufällige Parameter variierte betonähnliche Textur zu sehen. Die Umsetzung dieser Idee brachte sehr deutliche Verbesserungen bei den Fehlerkennungen.

Final kann der Ablauf der Erzeugung wie folgt beschrieben werden:

- Es wird eine Fahrbarkeitskarte für den Stapler erstellt, damit werden alle Bereiche markiert, in dem der Stapler positioniert werden kann
- Erzeugung der Trainingsbilder:
    - Der Stapler und mehrere weitere Objekte werden sowohl zufällig auf der Fahrbarkeitskarte positioniert als auch zufällig ausgerichtet.
    - Hinter dem Stapler ist eine nicht sichtbare Fläche definiert, die dem Detektionsbereich des Sensors entspricht (7m tief, bis an den Bildrand breit).
    - Auf dieser Fläche wird eine der 10 Personen zufällig positioniert und zufällig ausgerichtet.
    - Der Person wird zufällig eines der 50 Bewegungsprofile zugewiesen und innerhalb des Bewegungsprofils eine zufällige Pose ausgewählt.
    - Die Bodentextur wird zufällig neu generiert und angewandt.
    - Die Daten der Shader werden exportiert, das Labelbild wird wieder in einem erweiterten Sichtwinkel berechnet und ausgegeben.

Anschließend werden die von den Shader gesamten Daten durch ein in selbst entwickeltes Tool eingelesen und verarbeitet. Dessen Hauptaufgaben waren:

- Die Reduktion redundanter Teile, die vom Shader ausgegeben werden
- Die Berechnung der Labels aus dem Labelbild. Zu den Labels (vorstellbar als Rahmen) werden zusätzliche Informationen berechnet und gespeichert, wie z.B. die Anzahl der als Person klassifizierten Pixel, der Füllgrad des Labels usw.

Der Export der Labels erfolgte anschließend durch eine weitere Routine, Frames mit einem zu kleinen sichtbaren Anteil der Person oder zu stark verdeckte Personen, z.B. wenn nur ein Teil der Hand oder Fuß abgebildet wurden, wurden herausgefiltert.

6.4.3 Tracking (FML)

Zur Verbesserung der Erkennungsleistung können bereits erkannte Personen mit verschiedenen Verfahren aus dem Bereich Tracking verfolgt werden, bis sie den Bildbereich verlassen.

Texturbasierte Trackingverfahren extrahieren im Allgemeinen charakteristische Kontrastunterschiede im Bild und versuchen diese in neuen Bildern wiederzufinden. Der Nachteil dieser Verfahren ist, dass bei Personendetektionsverfahren meist nur Rechtecke, welche die erkannten Menschen umschließen, als Ergebnis geliefert werden und diese auch zum Teil den Hintergrund beinhalten. Dadurch unterliegen diese Verfahren einem „Drift“, der bei einer bewegten Kamera dazu führt, dass ggf. nicht mehr das Objekt, sondern der Hintergrund verfolgt wird. Der Vorteil der Verfahren ist die Verfolgung in jedem Bild, bis der Drift zu groß wird. Folgende Verfahren sind für das Projekt prinzipiell geeignet: BOOSTING [Gra-2006], MIL [Bab-2009], TLD [Kal-2012], MEDIANFLOW [Kal-2010], KCF [Dan-2014].

Ein konkurrierender Ansatz zu den texturbasierten Methoden ist die Anwendung des Personendetektionsalgorithmus in einem Vorhersagebereich, der über die letzten Erkennungen eines Objekts berechnet wird. Bei der Erkennung im Vorhersagebereich wird der Schwellwert, der bestimmt ob eine Erkennung als Mensch akzeptiert wird, niedriger gesetzt.

Aufgrund der Tatsache, dass Veröffentlichungen zu den genannten Tracking-Verfahren auf statischen positionierten Kameras beruhen, konnte keine Beurteilung über die Leistungsfähigkeit der Verfahren im Einsatz bei einer bewegten Kamera getroffen werden. Daher wurden alle genannten Verfahren implementiert und ausgewertet.

6.4.4 Kollisionserkennung (FML)

Die Kollisionserkennung kann alleinig mit den Daten der Tiefenbilder oder in Verbindung mit den Intensitätsdaten der Microsoft Kinect v2 umgesetzt werden. Es wurden nur Verfahren ausgewählt, welche die Geschwindigkeit des Staplers in Relation zu Objekten im Fahrweg berücksichtigen. Einerseits handelt es sich dabei um die Verwendung einer Top-Down-Karte und Verfahren des optischen Flusses.

6.4.4.1 Ansatz Top-Down-Karte

Ein Ansatz, der auf Tiefendaten basiert, ist die Erstellung einer Top-Down-Karte. Diese wird durch die Umwandlung aller Tiefenwerte in eine Punktwolke und der darauf folgenden Projektion aller Tiefenwerte auf die Bodenebene erstellt. Das Resultat ist ein Intensitätsbild, auf dem die Umrisse von Objekten abgebildet werden. Das Verfahren Enhanced Correlation Coefficient Maximization (ECC) [Bab-2009] fasst die einzelnen Punkte zu Objekten zusammen und berechnet die Transformationsmatrix, welche die Bewegung eines Objektes in zwei aufeinanderfolgenden Bildern beschreibt. Anhand der Transformationsmatrix werden die relativen Geschwindigkeiten zwischen dem Stapler und den Objekten sowie der Lenkwinkel zur starren Achse des Staplers berechnet. Zur Kollisionswarnung wird ein Gefahrenbereich um den Stapler definiert, der abhängig von der relativen Geschwindigkeit zu den Objekten, dem aktuellen Lenkwinkel, dem maximal möglichen Lenkeinschlag und der maximal möglichen Geschwindigkeitsänderung ist.

6.4.4.2 Ansatz Scene-Flow

Ein weiterer getesteter Ansatz zur Kollisionserkennung ist die Nutzung von „Scene-Flow“-Verfahren, die als Eingangsdaten ein Farb- oder Intensitätsbild und ein Tiefenbild erfordern. Scene-Flow-Methoden berechnen die Bewegung einzelner prägnanter Punkte in konsekutiven Bildern. Das resultierende Bild liefert den Geschwindigkeitsvektor jedes einzelnen Pixels in Relation zur Kamera. Der Winkel und die Höhe der Kamera müssen daher bekannt sein oder berechnet werden, damit die Bewegungsvektoren in das Koordinatensystem des Staplers umgerechnet werden können. Ob eine Kollision vorliegt, wird über das „Ray-Box-Intersection-Verfahren“ [Eri-2004] ermittelt. Bei diesem wird virtuell ein Quader um den Stapler definiert. Für jeden Pixel wird der Zeitpunkt bis zur Kollision mit dem Quader berechnet. Wird die minimale Kollisionszeit unterschritten, wird der Fahrer akustisch gewarnt. Es wurden die etablierten Scene-Flow-Algorithmen „Farneback“ [Far-2003], „Dual-TVL“ [Pér-2013] und der Algorithmus „Primal-Dual“ [Jai-2015] implementiert und ausgewertet.

6.5 Evaluation der Methoden zur Kollisions- und Personenerkennung

6.5.1 Kollisionserkennung (FML)

Zur Überprüfung der Kollisionserkennung wurden keine geeigneten Methoden in der Literatur gefunden. Daher wurden eigene Kennzahlen definiert, welche die Qualität der Kollisionserkennung beschreiben:

- Kollisionserkennung: In wie vielen Sequenzen wurde die Kollision erkannt, wobei falsche Erkennungen als negativ gewertet werden?
- Konstante Erkennung: In wie vielen Sequenzen wurde die Kollision konstant vom Anfang bis zum Ende der Kollision erkannt?
- Rechtzeitige Erkennung: In wie vielen Sequenzen wurde die Kollision vor der maximal zulässigen Kollisionszeit von drei Sekunden erkannt?
- Falsche Erkennungen: Wieviel Prozent der Kollisionswarnungen waren falsch?

Die Evaluation wurde anhand der in Abschnitt 6.1 definierten Testfälle durchgeführt. Dabei wurden allerdings keine Menschen, sondern Kartonagen mit der Höhe und Breite von Menschen und eine Schaufensterpuppe (siehe Abbildung 20) als Hindernisse benutzt. Als Eingangsbild diente das Intensitätsbild, da es schärfer als das Farbbild und unabhängig vom Umgebungslicht ist. Die maximal tolerierte Kollisionszeit von drei Sekunden wurde empirisch ermittelt. Bei einer Geschwindigkeit von 8 km/h wurde die Brems- und Reaktionszeit in Versuchen mittels einer inertialen Messeinheit und einem akustischen Warnsignal gemessen. Die Zeit bis zum Stillstand betrug maximal 2,5 Sekunden bei einem Stapler[10] ohne Ladung.

Abbildung 20: Beispiel für Kollisionsaufnahme mit Schaufensterpuppe (Microsoft Kinect v2)

[10] Modell: Jungheinrich EFG-220

Der Ansatz mit der Top-Down-Karte hat mit beschriebener Evaluationsmethode nicht funktioniert. Es stellte sich im Laufe der Versuche heraus, dass zur korrekten Erfassung der Geschwindigkeit des Staplers mindestens drei Objekte im Sichtfeld der Kamera sein müssen. Da die Testfälle mit einem Hindernis spezifiziert sind, konnte diese Methode nicht evaluiert werden. Eine weitere Evaluation war ohnehin nicht notwendig, da die Methode auf Grund dieser Problematik als nicht praxistauglich bewertet wurde.

Die Scene-Flow-Algorithmen funktionieren unabhängig von Objekten im Kamerasichtfeld, da diese nicht die Bewegung von vollständigen Objekten, sondern von einzelnen Pixeln berechnen. Die Auswertung (siehe Tabelle 6-6) der verschiedenen Scene-Flow-Algorithmen hat ergeben, dass der Algorithmus „Primal-Dual“ in allen Punkten die besten Ergebnisse liefert. Einzig bei der benötigten Rechenleistung ist der Algorithmus „Farneback“ deutlich schneller.

Tabelle 6-6: Evaluation verschiedener Scene-Flow-Algorithmen

	Kollision erkannt	**Rechtzeitig erkannt**	**Konstant erkannt**	**Falsche Kollision erkannt**	**Bilder pro Sekunde (FPS)**
Farneback	76,2%	31,3%	75,0%	52,4%	100 (CPU)[11]
Dual-TVL	57,1%	33,3%	41,7%	28,6%	14 (CPU)[11]
Primal-Dual	100,0%	85,7%	90,5%	14,3%	61 (GPU)[12]

Während der Evaluation wurde ein Problem mit den Scene-Flow-Algorithmen erkannt. Wenn sich der Gabelstapler um seine Achse drehte, konnten die Scene-Flow-Algorithmen aufgrund der Bewegungsunschärfe die Bewegung der Pixel nicht mehr zuverlässig berechnen. Jedoch trat dieses Problem nur auf, wenn als Eingangsquelle das Farbbild benutzt wurde. Aufgrund dessen wurde das Intensitätsbild als Ersatz für das Farbbild als Eingangsquelle verwendet, da dieses aufgrund des Sensors nur eine sehr geringe Bewegungsunschärfe aufweist. Dadurch wurde die Funktionsfähigkeit der Scene-Flow-Algorithmen auch in diesen Fahrsituationen gewährleistet, wobei die Leistung bei Geradeausfahrten gleich blieb.

Die beste Erkennung von Kollisionen wurde eindeutig mit dem Algorithmus „Primal-Dual“ erreicht, daher wurde im weiteren Projektverlauf dieser Algorithmus zur Kollisionserkennung verwendet.

6.5.2 Personenerkennung (FML)

Die Personenerkennung ist essentiell für das Funktionieren der zweistufigen Kollisionswarnung. Im Allgemeinen ist es wünschenswert, dass 100 % der Personen erkannt werden. Allerdings ist dies laut den Veröffentlichungen zu den Algorithmen bei keinem aktuell möglich, weshalb ein möglichst hoher Wert angestrebt wird. Zur Auswertung wurden Testdaten verwendet, welche am Lehrstuhl fml mit der Microsoft Kinect v2 aufgenommen wurden (siehe Abschnitt 6.3.3 und Abbildung 16).

[11] CPU: i7-6820HK
[12] GPU: NVIDIA Geforce GTX 1070

6.5.2.1 Evaluationsmethode

Zur Evaluation sind markierte Testdaten notwendig, damit die Evaluation automatisiert durchgeführt werden kann. Anhand der Markierungen können Erkennungen als falsch (engl.: „false positive“) oder (engl.: „true positive“) zugeordnet werden. Darüber hinaus können nicht gefundene (engl.: „false negative“) Personen erkannt werden.

Die Entscheidung, ob eine Erkennung richtig oder falsch ist, wird anhand eines Vergleichs der erkannten und markierten Rechteckfläche nach Dollar et al. durchgeführt [Dol-2012]:

$$x = \frac{A(BB_t \cap BB_e)}{A(BB_t \cup BB_e)} > 0{,}5$$

Entscheidend ist das Verhältnis x der Fläche der Schnittmenge des markierten ($\mathrm{BB_t}$) und des erkannten Rechtecks ($\mathrm{BB_e}$) zur Fläche der Vereinigungsmenge beider Rechtecke. Dieses Verhältnis muss laut Dollar et al größer als 0,5 sein, damit ein erkanntes Rechteck $\mathrm{BB_e}$ als „true positive“ gewertet wird. [Dol-2012]

Da es nicht sinnvoll ist jeden einzelnen Pixel einer Person zu markieren, werden üblicherweise Rechtecke verwendet. Die Markierung der Testdaten aus Abschnitt X erfolgte mittels Vorschlägen von Personenerkennungs-Deep-Learning-Netze mit nachträglicher manueller Verifikation bzw. Ergänzung.

Die Leistung (auch: Erkennungsrate) eines Algorithmus kann beliebig gesteigert werden, indem der Grenzwert zur Akzeptanz für die Wahrscheinlichkeit, dass es sich bei einem Bildausschnitt um einen Menschen handelt, niedrig gesetzt wird. Wird der Grenzwert zu niedrig angesetzt, kann allerdings nicht mehr von einer Klassifikation bzw. Lokalisierung gesprochen werden. Dadurch entsteht das Dilemma, bei welchem Grenzwert Personenerkennungsalgorithmen bezüglich der Detektionsrate zu vergleichen sind. *Dollar et. al* haben eine Methodik zur Auswertung in der Personenerkennung vorgeschlagen, die sich als Referenz etabliert hat. Bei dieser werden Stützpunkte durch verschiedene Akzeptanzgrenzwerte erzeugt und die resultierenden Graphen miteinander verglichen. Die Stützpunkte werden durch die jeweilige Anzahl an falschen Erkennungen pro Bild („false positives per image“) und der prozentualen Anteil an verfehlten Personen („miss rate“) bestimmt. Die Entscheidung, ob es sich um eine richtige oder falsche Erkennung handelt, wird über den Vergleich der Rechtecke der erkannten und der markierten Person getroffen: [Dol-2012]

Aufgrund der Vielzahl der Graphen werden im Folgenden zum Vergleich verschiedener Durchläufe die durchschnittlichen miss rates ($\overline{\mathrm{mr}}$) angegeben. Dafür wurde eine Formel erarbeitet, die sich auf empirische Vorversuche stützt:

$$\overline{mr} = \frac{1}{0.99} \int_{0{,}01}^{0{,}1} mr\left(x_{fppi}\right) dx_{fppi}$$

Die miss rate wird über den Wertebereich von 0,01 bis 0,1 false positives per image (fppi) integriert und normiert. Unter 0,01 fppi wurden in den Vorversuchen nahezu keine Menschen mehr erkannt. Bei Werten von über 0,1 fppi spricht man im Allgemeinen nicht mehr von einer Klassifikation, da in dem Bereich zu viele Falscherkennungen vorliegen.

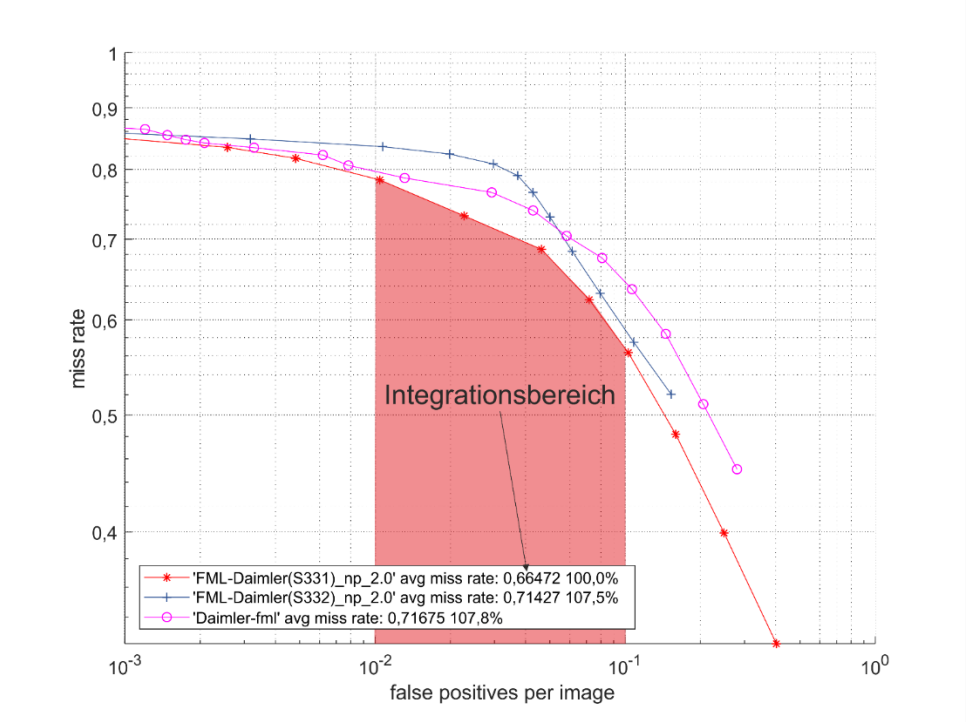

Abbildung 21: Beispiel für eine Auswertung nach Dollar et al.

Der Vergleich mehrerer Detektionsdurchläufe kann entweder grafisch oder rein quantitativ über die gemittelten miss rates erfolgen. Im Falle der grafischen Analyse ist ein Durchlauf umso besser, je näher dessen Graph am Koordinatenursprung ist. Im Beispiel aus Abbildung 21 wäre dies die rote Kurve. Das gleiche Resultat ergibt sich bei der quantitativen Analyse durch die kleinste mittlere miss rate. Beim Vergleich des blauen mit dem pink Graph zeigt sich der Vorteil der quantitativen Methode, da hier schwierig zu beurteilen ist, welcher Graph näher am Koordinatenursprung ist.

6.5.2.2 Einflussfaktoren der maschinellen Personenerkennung

Die Funktionsfähigkeit von Machine-Learning-Algorithmen hängt maßgeblich von den Daten ab, die zu deren Training und bei der jeweiligen Klassifikationsaufgabe verwendet werden.

Machine-Learning-Algorithmen müssen vor ihrer Anwendung trainiert werden. Für Klassifikationszwecke werden beim Training zwei Arten von Datensätzen verwendet: negative und positive. Im Bereich der Bildverarbeitung enthält der negative Datensatz Bilder ohne, der positive Datensatz Bilder mit dem zu klassifizierenden Objekt. Hierbei können die Bilddatenquellen, die Anzahl an Bildern und das Verhältnis zwischen positiven und negativen Bildern variiert werden.

Das Detektionsergebnis kann nach dem Training nur noch durch die Aufbereitung der Eingangsdaten oder je nach Algorithmus durch dessen Parameter verändert werden.

6.5.2.3 Ergebnisse zur Vorklassifikation durch Clustering

Anhand der in Abschnitt 6.3.3 erzeugten Testbilder wurden die verschiedenen Clustermethoden evaluiert. Als Bildeingang wurde das Tiefenbild verwendet, da es stärkere Gradienten als das Farbbild liefert und somit besser für die Clusteringmethoden geeignet ist. Es hat sich gezeigt, dass die Wasserscheidentransformation und k-means zu rechenintensiv (< 5 FPS) für eine echtzeitfähige Lösung sind[13]. Darüber hinaus tritt im Falle der Wasserscheidentransformation eine Übersegmentierung auf. Dadurch müssen entstandene Cluster wiederum durch Nachbarbeziehungen und Gradientengrenzwerte zusammengefügt werden. Wenn die Clusteranzahl beim k-means-Algorithmus nicht vorgegeben wird, sind mehrere Iterationen nötig, daher verringert sich die Bildrate

[13] Hardware: i7-6820HK (CPU), NVIDIA Geforce GTX 1070 (GPU)

nochmals deutlich (< 1 FPS). Aufgrund der genannten Gründe hat sich das Clustering im 2½-D Bild als Vorklassifikationsmethode nicht bewährt.

6.5.2.4 Ergebnisse der Support Vector Machine

Eingangs wurde untersucht, ob eine SVM bei Aufnahmen eines fahrenden Gabelstaplers in einer Lagerumgebung eingesetzt und wie diese gegebenenfalls optimiert werden kann. Dabei wurden vor allem die Auswirkungen des Trainings der SVM und der Einfluss der Variation von SVM-Parametern auf die Detektion betrachtet. Für die Auswertung wurden nur Aufnahmen der „Microsoft Kinect v2“ verwendet, da die Auflösung der Industriekamera „SICK Visionary-T“ für die SVM zu klein ist. Als Testdaten wurden markierte Aufnahmen aus der Versuchshalle des Lehrstuhls fml verwendet.

Es wurden folgende Untersuchungen durchgeführt:

- Training:
  - Variation der Bildquellen
  - Variation des Verhältnisses von trainierten Negativ- zu Positivbildern
  - Variation der Gesamtzahl trainierter Bilder
- Detektion
  - Variation der SVM-Parameter
  - Bearbeitung des Eingangsbildes
    - Skalierung
    - Normalisierung des Histogramms
    - Weichzeichnen
    - Segmentierung

Zum Training wurden Daten von den bestehenden Bilddatenbanken „Monocular Pedestrian Dataset“ (Daimler) und „INRIA“ als Positivbilder verwendet [Dal-2005, Enz-2009]. Darüber hinaus wurden die synthetisch hergestellten Daten (siehe Abschnitt 6.3.3) zum Training verwendet. Als Negativbilder wurden Aufnahmen aus der Versuchshalle des Lehrstuhls fml und Lageraufnahmen aus Werbevideos auf YouTube verwendet.

Tabelle 6-7: Auswertung der durchschnittlichen miss rates $\overline{mr}$ verschiedener Trainingsdatenquellen.

Negative \ **Positive**	**Daimler**	**INRIA**	**Generiert**
Farbbild			
fml	0,817162	0,711501	0,935724
fml + youtube	0,743076	0,657021	-
Registriertes Farbbild			
fml	0,934151	0,772104	0,973622
fml + youtube	0,770831	0,697621	-
Intensitätsbild			
fml	0,625815	0,659514	0,830962
fml + youtube	0,596541	0,662186	-

Die Auswertung in Tabelle 6-7 zeigt die durchschnittliche miss rate für den Bereich von 0,01 bis 0,1 fppi, kategorisiert nach den verschiedenen Eingangsbildarten. Im Falle der Nutzung des regulären und registrierten Farbbilds als

Eingangsbild ist der INRIA-Datensatz besser für das Training positiver Bilder geeignet, im Falle des Intensitätsbilds der Daimler-Datensatz. Darüber hinaus führte das Hinzufügen von den YouTube-Videos zu den Hallenvideos fast ausschließlich zur Verbesserung der Erkennung. Die generierten Bilder repräsentieren die Originale offensichtlich nicht gut genug, sie erbrachten das schlechteste Ergebnis.

Tabelle 6-8: Auswertung der durchschnittlichen miss rates $\overline{mr}$ in Abhängigkeit von der Gesamtzahl trainierter positiver Bilder.

Positive / Anz. Positive	Daimler	INRIA	Generiert
Farbbild			
3.500	0,787125	0,665960	0,93619
7.000	0,799158	0,671778	0,93345
Registriertes Farbbild			
3.500	0,934299	0,713729	0,94842
7.000	0,899081	0,715675	0,98974
Intensitätsbild			
3.500	0,580764	0,67855	0,83718
7.000	0,576535	0,638926	0,84972

Die Abhängigkeit der Detektionsleistung von der Gesamtanzahl an trainierten Bildern wurde für zwei Fälle evaluiert: 3.500 und 7.000 positive mit jeweils 7.000 bzw. 14.000 negativen Bildern. Da der INRIA-Datensatz nur 3.500 Bilder enthält, konnten maximal 7.000 Bilder getestet werden. Die 7.000 Bilder wurden durch die Addition 3.500 gespiegelter Bilder erzeugt. Im Einzelfall verbesserte sich die Erkennungsrate um bis zu 5,8% (INRIA – Intensität), wenn doppelt so viele Bilder trainiert wurden. Im Durchschnitt über alle Bilddatentypen lag die Verbesserung durch Verdopplung allerdings nur bei 0,12%; eine signifikante Verbesserung durch mehr Daten konnte daher nicht bewiesen werden.

Tabelle 6-9: Auswertung der durchschnittlichen miss rates $\overline{mr}$ in Abhängigkeit vom Verhältnis der Anzahl trainierter negativer und positiver Bilder.

Neg/Pos \ Positives	Daimler	INRIA
Farbbild		
1,4	0,784348	0,684578
1,6	0,809599	0,664515
1,8	0,832643	0,668071
2,0	0,798177	0,670463
2,2	0,829614	0,669067
2,4	0,834406	0,65259
Registriertes Farbbild		
1,4	0,904281	0,705897
1,6	0,922653	0,715457
1,8	0,950980	0,724587
2,0	0,878693	0,730740
2,2	0,962759	0,718293
2,4	0,939863	0,687821
Intensitätsbild		
1,4	0,551785	0,623717
1,6	0,573000	0,624082
1,8	0,655876	0,678587
2,0	0,637582	0,670200
2,2	0,665150	0,659958
2,4	0,656080	0,713950

In einem weiteren Test wurde die Anzahl der negativen Bilder bei konstanter Anzahl an positiven Bildern (3.500 und 7.000) verändert. Die Auswertung in Tabelle 6-9 zeigt, dass bei keiner Bildart eine Aussage getroffen werden kann, welches Verhältnis am besten ist. Im Durchschnitt ist die Erkennung beim besten Verhältnis (1,4) 5,8 % besser als im schlechtesten Fall (1,8).

Bei der Variation der SVM-Parameter wurden aufgrund der hohen Anzahl an Variationsmöglichkeiten nur die drei besten SVM der jeweiligen Bildarten evaluiert. Folgende Parameter wurden untersucht:

- hit threshold: Grenzwert mit welcher Erkennungswahrscheinlichkeit Menschen als solche akzeptiert werden.
- group threshold: Grenzwert zur Vereinigung von Detektionen an einer ähnlichen Position.
- h scale: Prozentuale Vergrößerung des Klassifikationsfensters pro Bilddurchlauf.
- n levels: Anzahl an Vergrößerungen des Klassifikationsfensters bzw. Anzahl der Durchläufe pro Bild.

Der Parameter „hit_threshold“ wird bei allen Durchläufen von 0,1 bis 1,4 zur Erzeugung der Stützpunkte variiert (siehe Evaluationsmethode). Ein weiterer Grenzwert ist der Parameter „group threshold“. Dieser beschreibt inwieweit sich überschneidende Rechtecke von Erkennungen zu einem einzigen zusammengefasst werden.

Tabelle 6-10: Auswertung der durchschnittlichen miss rates $\overline{mr}$ unterschiedlicher „group threshold“-Werte.

group threshold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Farbbild	0,71379	0,71021	0,68749	0,68408	0,68513	0,67982	0,66626	0,66118	0,66738
Registriertes Farbbild	0,66875	0,65615	0,65468	0,65310	0,65169	0,65567	0,65829	0,65732	0,65944
Intensitätsbild	0,58294	0,56772	0,55576	0,54882	0,54154	0,53346	0,52568	0,52179	0,52257

Die Auswertung der Unterschiedlichen „group threshold“-Werte in Tabelle 6-10 zeigt einen Trend möglichst keine Rechtecke zusammenzufassen (9). Ein Grund hierfür könnten nahestehende Personen sein. Werden diese zusammengefasst und als ein Mensch betrachtet, wird die Erkennung nach Dollar et. al als falsch gewertet.

Bei der Auswertung des Vergrößerungsfaktors „h scale“ wurde die maximale Anzahl möglicher Vergrößerungen („n levels“) berechnet, da sonst etwaige Erkennungen von sehr großen Personen fehlen könnten. Die Berechnung erfolgt auf Basis der Eingangsbildgröße:

$$n_{levels\_max} = \frac{\log(h_i * h_d)}{\log(h_{scale})}$$

$h_i = Bildhöhe\ [Px]$

$h_d = Höhe\ des\ SVM - Detektors\ [Px]$

Tabelle 6-11: Auswertung der durchschnittlichen miss rates $\overline{mr}$ unterschiedlicher „h scale“-Werte.

h scale	**1,04**	**1,05**	**1,06**	**1,07**	**1,08**	**1,09**	**1,10**	**1,11**	**1,12**
Farbbild	0,71337	0,68766	0,72358	0,70672	0,69579	0,70331	0,70069	0,68927	0,69738
Registriertes Farbbild	0,65079	0,65217	0,65781	0,65828	0,65308	0,65801	0,65860	0,66620	0,66939
Intensitätsbild	0,51635	0,51918	0,52117	0,52898	0,53608	0,54552	0,56093	0,56026	0,56172

Der optimale Vergrößerungsfaktor liegt nach Tabelle 6-11 bei 1,04 oder 1,05. Grundsätzlich ist ein Trend zu kleinen Werten zu erkennen. Werte unter 1,04 wurden nicht untersucht, da das Suchfenster in dem Fall nur noch minimal vergrößert wird. Die benötigte Rechenleistung würde deutlich steigen, wenn auch Menschen erkannt werden sollen die im Bild die Größe der Bildhöhe haben.

Für den besten „h scale“-Wert von 1,05 wurde schließlich die beste Anzahl an Vergrößerungen „n levels“ ermittelt. Um die beste Erkennung zu erreichen, kann dieser Wert wie bereits erwähnt berechnet werden. Mit jeder Vergrößerung erhöht sich allerdings auch der Rechenaufwand. Daher wurde untersucht, ob eine Sättigung bei einem bestimmten Wert eintritt. Dies konnte bei der Anzahl von 17 Vergrößerungen festgestellt werden. Die Erhöhung auf 18 Vergrößerungen ergab nur noch eine Verbesserung der Erkennung um 0,6 %.

Tabelle 6-12 zeigt die Abhängigkeit der Erkennungsleistung von der Skalierung des Eingangsbildes. Für das Farbbild sind andere Skalierungsfaktoren verwendet worden als für die anderen Bildarten. Die Ursache hierfür liegt in der Auflösung des Farbbildes von 1490x1232 Pixeln (registriertes Farb- oder Intensitätsbild: 512x424 Pixel). Kleinere Skalierungsfaktoren konnten aufgrund der verwendeten Sliding-Window-Methode nicht untersucht werden. Die Auswertung wurde ebenfalls mit der maximal möglichen Anzahl an Vergrößerungen nlevels_max durchgeführt.

Tabelle 6-12: Auswertung der durchschnittlichen miss rates $\overline{mr}$ unterschiedlicher Skalierungsfaktoren des Eingangsbildes.

Skalierung [%]	**0,15**	**0,25**	**0,35**	**0,40**	**0,45**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**1,0**
Farbbild											
	0,2432	0,5048	0,6562	-	0,6998	-	-	-	-	-	-
Registriertes Farbbild											
	-	-	-	0,7326	-	0,6470	0,7930	0,9231	0,9345	0,9397	0,9606
Intensitätsbild											
				0,6448	-	0,5486	0,7097	0,7080	0,7323	0,7558	0,7942

Bei allen Bildarten hat sich gezeigt, dass die Erkennung bei Verwendung einer Auflösung von etwa 256x212 am besten ist. Im Falle des Farbbilds liegt der Skalierungsfaktor bei 0,15, beim registrierten Farb- und Intensitätsbild aufgrund der anderen Auflösung bei 0,4. Der Grund für die bessere Erkennung bei kleinen Auflösungen liegt wahrscheinlich an den Trainingsdaten. Die Bilder aus den Personendatenbanken, welche für den positiven Trainingsdatensatz verwendet wurden, hatten entweder 48x96 oder 64x128 Pixel.

Weiter wurde untersucht inwieweit Normalisierung und Weichzeichnen des Eingangsbildes Verbesserungen bei der Erkennung bewirken. Darüber hinaus wurde die Entfernung des Bodens und Hintergrunds (Segmentierung) evaluiert.

Tabelle 6-13: Auswertung der durchschnittlichen miss rates $\overline{mr}$ unterschiedlicher Skalierungsfaktoren des Eingangsbildes.

Vorverarbeitungsmethoden	**Keine**	**Normalisieren**	**Weichzeichnen**	**Segmentierung**
Farbbild	0,68513	0,68513	0,68513	0,57980
Registriertes Farbbild	0,65340	0,65340	0,65340	0,62547
Intensitätsbild	0,51998	0,58896	0,51510	0,62017

Nach Tabelle 6-13 bewirkt das Normalisieren oder Weichzeichnen im Falle des Farbbilds keine Änderung bezüglich der Erkennung. Die Segmentierung verbessert die Erkennung mit bis zu 15 %. Bezüglich des Intensitätsbildes gibt es keine signifikanten Verbesserungen durch Bearbeitung des Eingangsbilds.

Zusammenfassend ist als Eingangsquelle das Intensitätsbild für die SVM in der Versuchsumgebung am besten geeignet. Damit die SVM für das Intensitätsbild optimal funktioniert, sollten folgende Punkte berücksichtigt werden:

- Das Training mit Positivbildern der Daimler-Datenbank und Negativbildern aus Realaufnahmen in Kombination mit (Werbe-) Videos andere Lager.
- Die Anzahl an Trainingsbildern muss empirisch ermittelt werden.
- Die Sliding-Window-Parameter sind so zu wählen, dass die Suchfenstergröße maximal 17 mal um 5 % erhöht wird.
- Als Vorverarbeitung sollte das Bild auf die Auflösung von 256x112 Pixeln reduziert und das Bild weichgezeichnet werden.

6.5.2.5 Ergebnisse zur Verbesserung der Personendetektion durch Tracking-Verfahren

Die Anwendung der Tracking-Verfahren auf die Aufnahmen aus der Versuchalle des Lehrstuhls fml hat gezeigt, dass alle Methoden die Erkennungsleistung wesentlich verschlechtern. Die Ursache hierfür lag in der Verfolgung von falschen Erkennungen. Zur Reduktion des Trackings falscher Erkennungen musste zwangsläufig der Akzeptanzwert,

dass es sich um einen Menschen handelt, erhöht werden. Dies führte wiederum zur Reduktion richtiger Erkennungen. Der Einsatz von Tracking-Verfahren ist demzufolge bei automatischen Detektionen nicht zielführend.

6.5.3 Evaluation der Deep Learning Methoden (BIBA)

Die im Folgenden dargestellte Ansätze 1 und 2 dienten zur Evaluierung und Weiterentwicklung einer Deep-Learning-basierten Personenerkennung. Daher werden die Ergebnisse für diese beiden Ansätze lediglich qualitativ beschrieben.

6.5.3.1 DL-Ansatz 1 : Personendetektion durch Segmentierung, Deep Learning zur Klassifizierung

Der erste Ansatz beinhaltete die Evaluation von Deep Learning für die projektspezifische Problemstellung unter Laborbedingungen. Für das Training der Netze wurden mit einer Microsoft Kinect inklusive Augmentation 18.352 Aufnahmen mit Menschen und 11.554 Aufnahmen ohne Menschen verwendet. Es wurden dabei vier verschiedene Netzarchitekturen (Cifarnet[14], AlexNet, Small Cifarnet, GoogLenet [Sze-2015]) getestet. Diese Netze wurden jeweils in vier verschiedenen Datenkombinationen (RGB, Tiefendaten, Graustufen und Tiefe, RGB und Tiefe) überprüft.

Der Vergleich der Datenvariation zeigt, dass die Tiefendaten für das Klassifizieren nicht erforderlich sind, jedoch im Bereich der Hypothesenbildung eine herausragende Rolle spielen. Falscherkennungen werden weitestgehend vermieden. Die Erkennungsgenauigkeit dieses Ansatzes ist als sehr gut zu bezeichnen, Personen in den Validierungsdaten wurden mit dem *GoogLeNe*t-Netz unter Verwendung von Farb- und Tiefendaten zu 99,3% richtig erkannt. Leider besitzen alle Klassifizierungsnetze als Endstufe voll vernetzte Ebenen, die sehr rechenaufwändig sind. Zudem hat der Ansatz den großen Nachteil, dass die Berechnungszeit sehr stark von der Anzahl der Hypothesen abhängt und somit nicht deterministisch ist. Dabei erreichte der Algorithmus maximal 3 FPS. Jedoch konnte gezeigt werden, dass Deep-Learning zur Personendetektion sehr gut geeignet ist.

6.5.3.2 DL Ansatz 2: Personendetektion direkt durch echtzeitfähiges Netz (SSD)

Das Ergebnis bei der Personenerkennung mit dem zweiten Ansatz war letztendlich, dass diese zum größten Teil zuverlässig erkannt wurden, aber nur aus der horizontalen Sensorperspektive. Das Netz war nicht darauf trainiert, Menschen mehr oder weniger von oben zu erkennen. Man hätte damit auch ein SSD-Netz mit synthetischen Daten trainieren können, jedoch versprach das vergleichbar funktionierende, später herausgebrachte *DetectNet* hierfür eine bessere Leistung, weshalb Ansatz 2 nicht weiterentwickelt wurde.

Zur Evaluation wurden freie Fahrten in der Versuchshalle des BIBA verwendet. Es konnte bei Verwendung der Farbdaten eine gute Detektionsleistung von Menschen festgestellt werden. Die hinzugefügten weiteren Objektklassen werden dagegen schlechter erkannt, dies hängt vor allem mit der vergleichsweise geringen Menge an Trainingsdaten zusammen. Diese manuell zu erstellen, bedeutet einen sehr großen Arbeitsaufwand. Im Vergleich zu einer auf HOG-Features basierten SVM sind deutlich weniger Falscherkennungen zu verzeichnen. Die Verwendung von Tiefendaten hat bei der Erkennung keine sinnvollen Ergebnisse hervorgebracht, vermutlich ist hierfür auch die geringe Anzahl an Trainingsdaten verantwortlich. Dieser Ansatz ist jedoch im Vergleich zum zuvor vorgestellten ersten Ansatz mit vorgeschalteter Hypothesenbildung klar zu bevorzugen. Der größte Vorteil ergibt sich in der Echtzeitfähigkeit, die zudem unabhängig von der Szenerie und Anzahl der Objekte gegeben ist. Die zum Zeitpunkt der Tests verwendete Highend Nvidia GTX980TI-Grafikkarte wurde nahezu konstant zu 51% ausgelastet. Abbildung 22 zeigt eine beispielhafte Analyse mit klassifizierten Objekten und Angabe der zugehörigen Klassifikationssicherheit.

[14]http://caffe.berkeleyvision.org/gathered/examples/cifar10.html

Abbildung 22: Beispielbilder vom Test des zweiten DL-Ansatzes (Microsoft Kinect)

6.5.3.3 DL-Ansatz 3 Echtzeitfähiges Netz trainiert mit synthetischen Daten

Dieser Ansatz wurde innerhalb der Laborumgebung des BIBA evaluiert. Hier wurden willkürliche Szenarien mit mehreren Personen nachgestellt, die zufällige Bewegungen durchführen und während der Bewegung Objekte transportierten. Die Aufnahmen wurden zum größten Teil in engen Teilen der BIBA-Versuchshalle durchgeführt, die häufig weitere komplexe Objekte enthielten. Abbildung 23 zeigt zwei Beispielszenen, aufgenommen von der Kinect v2 Kamera, die auf einem FFZ montiert ist. Alle Personen konnten in dem Durchlauf zuverlässig erkannt werden.

Abbildung 23: Beispielbilder aus den Laborexperimenten (Microsoft Kinect v2)

Die Experimente wurden für die kontinuierliche Weiterentwicklung der Methode verwendet und bestätigten das große Potential des Ansatzes. Die resultierende Konfiguration wird in dem nachfolgenden Kapitel in der Beschreibung der Feldtest verwendet, in der die quantitative Auswertung der Methode beschreiben wird.

6.6 AP7: Feldtests (BIBA, FML, SICK; STILL)

Auf Basis der Ergebnisse der beiden Ansätze des FML und BIBA unter Laborbedingungen wurde entschieden, die Feldtests ausschließlich mit dem Deep Learning-Ansatz durchzuführen, da dieser echtzeitfähig ist, ein größeres Potential hinsichtlich der Güte der Erkennung und insbesondere bei der geringeren Anzahl an Fehlalarmen aufweist.

Der Demonstrator besteht aus einem Schubmaststapler, 3 Sensoren (2 unterschiedliche Versionen des SICK Visionary-T Sensors sowie der Microsoft Kinect v2), einer Recheneinheit und Elemente für die Spannungsversorgung (Gleichstromumformer, Gleichspannungstransformator und Sicherungskasten). Die Schnittstelle zum Fahrer stellt ein Touchscreen-Panel dar, das direkt vor dem Fahrer platziert ist. Hier wird der Fahrer visuell über eine Person im Arbeitsbereich informiert. Ein Foto vom Demonstrator ist der Abbildung 24 zu entnehmen.

Abbildung 24: Demonstrator mit Sensoranbau

Alle Testszenarien wurden mit allen Sensoren gleichzeitig aufgezeichnet und analysiert. Die anschließende Analyse bezieht sich aufgrund der höheren Auflösung des Intensitätsbildes auf die Kinect v2.

Beschreibung Experimente

Es wurden an mehreren Tagen Testfahrten in der Testumgebung bei der Blackforxx GmbH durchgeführt, wobei acht unterschiedliche Testfälle mit dem Fokus auf der Evaluation der Personendetektion konzipiert, durchgeführt und ausgewertet worden sind. Tabelle 14 beschreibt die Testfälle und setzt diese in Kontext zu den ermittelten Anwendungsfällen aus Arbeitspaket 2. Neben den Testszenarien wurde ebenfalls der Einfluss unterschiedlicher Materialien und Farben bei Berufskleidung untersucht.

Tabelle 14: Szenarien der Testfälle im Feldtest

Nr.	Beschreibung	Dynamik	Komplexität	Anwendungsszenario
1.1	Person im blauen Arbeitsanzug (mit niedriger IR Reflektivität) läuft in ungeordneter Lagerumgebung jeweils frontal sowie im schräg von rechts und links im 45 Grad Winkel auf das FFZ zu und wieder weg. Das FFZ ist stationär. (insgesamt 3 Durchläufe mit unterschiedlichen Personen)	gering	mittel	1,2,5
1.2	Person, die eine Jacke trägt, läuft in ungeordneter Lagerumgebung jeweils frontal sowie im schräg von rechts und links im 45 Grad Winkel auf das FFZ zu und wieder weg. Das FFZ ist stationär. (insgesamt 2 Durchläufe mit unterschiedlichen Personen)	gering	gering	1,2,5
1.3	Person im Pullover/Shirt läuft in ungeordneter Lagerumgebung jeweils frontal sowie im schräg von rechts und links im 45 Grad Winkel auf das FFZ zu und wieder weg. Das FFZ ist stationär. (insgesamt 4 Durchläufe mit 3 unterschiedlichen Personen)	gering	gering	1,2,5
1.4	Person in reflektierender Warnweste läuft in ungeordneter Lagerumgebung jeweils frontal sowie im schräg von rechts und links im 45 Grad Winkel auf das FFZ zu und wieder weg. Das FFZ ist stationär.	gering	mittel	1,2,5

	(insgesamt 3 Durchläufe mit unterschiedlichen Personen)			
2	Längere Rundfahrt und simulierter Rangierfahrten durch Halle und engen Gängen mit vielen Flurförderzeugen mit sehr wenigen Personen zur Evaluation der Anzahl von False Positives	mittel	hoch	-
3	Drei Personen gehen vor dem Stapler herum und versuchen im gesamten Sensorbereich vertreten zu sein. Sie stellen dabei auch praktische und seltener vorkommende Posen zur Schau. (In einen Behälter schauen, Laufen, Schuhe zubinden, …) FFZ ist stationär. Durchgeführt in einer reinen Lagerhalle in einem deutlich dunkleren Bereich.	mittel	hoch	1,2,5
4	Fahrt mit teilweise verdeckten Personen in komplexer Lagerumgebung.	Hoch	Hoch	1,2,3,4,5,6,7,8,9
5	Verfolgungsfahrt einer Person mit Reflektorweste in einer dunkeln Lagerumgebung (1 Durchlauf)	Hoch	Hoch	2,5
6	Verfolgungsfahrt einer Person in einer dunkeln Lagerumgebung (2 Durchläufe mit verschiedenen Personen)	Hoch	Mittel	2,5
7	Verfolgungsfahrt von 2 Personen durch mehrere Hallen und Räume hinweg (2 Durchläufe in jeweils gegensätzlicher Richtung)	Hoch	Hoch	2,5
8	Vorbeifahrt an Personen 1 (2 Durchläufe)	Hoch	Hoch	1, 2, 3, 6, 8

Die Auswertemethodik wurde projektspezifisch entsprechend der weiteren Verarbeitung der Sensordaten festgelegt. Zur einfacheren Auswertung werden die Extremfälle nicht manuell aus der Statistik entfernt, sondern kurz in den Beschreibungen der Testfälle beschrieben. Zunächst wurden alle Frames im BIBA manuell und vollständig gelabelt. Dies umfasste alle im Frame vorhandenen Personen. Diese Labels werden als GroundTruth-(GT) Labels bezeichnet. Um direkt mit dem verwendeten Datenformat arbeiten zu können, wurde ein eigenes Programm für das Labeling erstellt.

Zunächst wurde ein Bereich im Bild festgelegt, der für das Personenerkennungssystem relevant ist und Bereiche ausgeschlossen, die anderweitig kontrolliert werden (bspw. durch das Kollisionserkennungssystem). Für den relevanten Bereich wurde anschließend der Detektor trainiert. Boundingboxen müssen vom Algorithmus erst erkannt werden, sobald die Boundingboxen der Personen, die diesen Bereich berühren überschneiden. Die verbliebenen Bildpunkte werden benötigt, um die Person zu erkennen. An den Seiten wurde jeweils ein Abstand von 64 Pixel definiert. Personen, deren zugewandte Oberfläche nur zu einem Drittel oder weniger erfasst wurde, können dabei vom Algorithmus nicht zuverlässig erkannt werden. Am oberen Bildrand werden 167 Pixel entfernt. Damit die Boundingbox einer Person diesen Bereich berührt, müssen sich Ihre Füße ca. 5m entfernt zum Sensor befinden. Tests aus AP 3 haben gezeigt, dass bei einigen Materialen (dunkle Jeans, blaue Jeans) ab 5m Entfernung keine Tiefenwerte gemessen werden können. Weiter entfernte Personen werden daher für die Evaluation der Personenerkennung ausgeschlossen. Am unteren Bildrand werden 85 Pixel ausgeschlossen. Objekte, die sich in diesem Bereich befinden, befinden sich unmittelbar am Stapler und müssen als Personen nicht erkannt werden.

Die Detektion wird vom Algorithmus für alle Frames durchgeführt. Zusätzlich werden zur Kontrolle Videos der Erkennung generiert, in dem die Boundingboxen der Erkennung sowie der Ausgang der Klassifizierungsschicht auf das Bild appliziert werden. Die anschließende Auswertung erfolgt wiederrum framebasiert. Sowohl die Erkennungen als auch die GT-Labels werden dahingehend gefiltert, dass sie sich mit dem “Rahmen” überschneiden müssen. Alle GT-Labels werden zunächst mit “nicht erkannt” markiert. Im Anschluss werden alle Detektionen sequentiell abgearbeitet. Jede Detektion wird daraufhin überpüft, ob Sie sich mit einem GT-Label überschneiden. Der

angewandte *Intersection over Union* (IoU) Wert wird hier mit geringen 0.1 angegeben: Im Falle der Überschneidung werden die betreffenden GT-Labels als “erkannt” markiert. Kann eine Detektion keinem GT-Label zugeordnet werden, wird sie als “False Positive” (Falscherkennung) gezählt.

Die einzelnen Testfälle wurden hinsichtlich der oben erläuterten Testdurchführung und Auswertemethodik durchgeführt. Dabei wurde für jeden Testfall analysiert wie viele Frames insgesamt aufgezeichnet wurden und auf wie viele Personen in den Frames insgesamt erkannt hätten werden müssen. Für jeden Testfall wird anschließend die Anzahl an „True Positives“ (TP) (wie viele Personen wurden richtigerweise erkannt), die Anzahl an „False Negatives“ (FN) (wie viele Personen wurden nicht erkannt) und die Anzahl an „False Positives (FP) (wie viele Personen wurden erkannt, obwohl keine Person zu erkennen war) aufgeführt. Zusammenfassend wird die Rate an Falscherkennungen pro 1000 Bilder sowie die prozentuale Angabe der Anzahl an richtiger Erkennungen im Kontext zur Gesamtanzahl an Personen aufgeführt. Pro Testfall werden zudem eine kurze Bewertung sowie ein Beispielintensitätsbild angegeben.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.1 1	1661	498	432	0	66	86,75	0

Kommentare

Die FP entstehen allesamt dadurch, dass Personen erkannt werden, die nach der o.g. Methodik nicht mehr erkannt werden müssten. Die FN sind nahezu ausschließlich in dem Bereich, wo die Person am Stapler steht und der Kopf sich gerade so im Soll-Detektionsbereich befindet. Da in dieser Situation sowieso gewarnt wird, daher ist die eigentliche Qualität der Erkennung deutlich über der automatischen Auswertung.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.1 2	1908	566	511	25	55	90,28	13,1

Kommentare

In der manuellen Kontrolle wurden keine tatsächlichen FP gezählt. Personen im Nahbereich am Stapler werden besser erkannt, FN entstehen sporadisch am entfernteren Soll-Detektionsbereich.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.1 3	1314	623	526	0	97	84,43	0

Kommentare

Personen werden im Nahbereich besser erkannt, da sie nicht so dicht an den Stapler herankommen, wie in 1.1.1 und 1.2.1. FN entstehen hauptsächlich beim Eintreten im Seitenbereich bei den 45 Grad Läufen. Hier ist die Detektionsleistung am geringsten.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.2 1	2928	681	595	9	86	87,37	3,07

Kommentare

Die FP sind durch die Hand der Person im Stapler erzeugt worden, mit der Anweisungen für die Tests gegeben wurden.
Daher ist die eigentliche Qualität der Erkennung deutlich über der automatischen Auswertung.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.2 2	1369	648	591	0	57	91,20	0

Kommentare

Die Jacke wird im seitlichen Bereich besser erkannt als der blaue Arbeitsanzug. Zudem ist das Ergebnis in Realität besser, da die Person nicht bis ganz an den Stapler herankommt

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.3 1	2069	529	449	1	80	84,88	0,48

Kommentare

Detektionsschwierigkeiten bestehen im nahen Seitenbereich. Dort ist die Sensordatenqualität auf Grund der schlechten Ausleuchtung nicht gut. Die Person läuft im schrägen Anlauf etwas flacher au das FFZ zu.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.3 2	2131	544	505	0	39	92,83	0

Kommentare

FN entstehen hauptsächlich beim Eintreten im Seitenbereich bei den 45 Grad Läufen. FP sind eigentlich TP außerhalb des abgegrenzten Bereiches.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.3 3	1306	602	566	0	36	94,02	0

Kommentare

siehe 1.3.2

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.3 4	1909	560	547	0	13	97,68	0

Kommentare
siehe 1.3.2

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.4 1	1523	517	406	0	111	78,53	0

Kommentare
siehe 1.3.2
Die Reflektorweste überblendet Details und macht die Detektion dadurch schlechter. Diese Situation könnte durch eine algorithmische Westenerkennung gelöst werden.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.4 2	2019	548	502	1	46	91,61	0,5

Kommentare
siehe 1.3.2
Die Reflektorweste überblendet Details und macht die Detektion dadurch schlechter. Diese Situation könnte durch eine algorithmische Westenerkennung gelöst werden.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
1.4 3	1231	522	436	4	86	83,52	3,25

Kommentare

siehe 1.3.2

Die Reflektorweste überblendet Details und macht die Detektion dadurch schlechter. Diese Situation könnte durch eine algorithmische Westenerkennung gelöst werden.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
2	20406	392	298	283	94	76,02	13,87

Kommentare

Mehrere Personen waren auf der Fahrt anzutreffen, sie liefen in einem größeren Abstand am Stapler vorbei, hieraus resultiert die höhere Anzahl an FN.

Die FP entstehen von vorkommenden strukturellen Anordnungen. Durch die vielfältigen Objekte sind die 14.11 ein valider Wert für diese Art von Lagern. Der Großteil der FP lag nicht in der Fahrtroute und damit außerhalb des Sichtfeldes des Sensors.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
3	1443	1989	1712	5	277	86,07	3,47

Kommentare

Die FP sind sämtlich „nicht erwünschte“ richtige Erkennungen außerhalb des Detektionsbereiches. Während des Bück-Vorganges konnte eine Person nicht erkannt werden, im gebückten Zustand dagegen wieder.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
4	3861	2332	1887	5	445	80,92	1,3

Kommentare

Die Fehlerkennung resultiert weniger aus den Verdeckungen als auf die vielen Fahrtmanöver dieser Szene. Dadurch befanden sich Personen sehr häufig in den äußeren Bereichen der Sensorfläche.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
5	1936	1438	1235	145	203	85,88	74,90

Kommentare

Das Gegenbeispiel zu Szenario 4. Trotz Reflektorweste konnten über 80% der Personen richtig erkannt werden. Die hohe Anzahl der FP sind real und müssen in Bezug zur Reflektorweste gesehen werden.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
6.1	1492	1352	1306	11	46	96,60	7,37

Kommentare

Sehr hohe Erkennungsrate, da sich die Person hauptsächlich im inneren Bereich des Detektionsrahmens bewegt hat.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
6.2	743	708	699	7	9	98,73	9,42

Kommentare

Sehr hohe Erkennungsrate, da sich die Person hauptsächlich im inneren Bereich des Detektionsrahmens bewegt hat.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
7.1	4344	2784	2461	44	323	88,40	10,13

Kommentare

Sehr hohe Erkennungsrate, da sich die Personen hauptsächlich im inneren Bereich des Detektionsrahmens bewegt haben.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
7.2	2958	2022	1714	28	308	84,77	9,47

Kommentare
siehe 7.1

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
8.1	391	105	64	4	41	60,95	10,23

Kommentare
Die Person befindet sich ausschließlich im Seitenbereich des Sensors. Dieser Test war für die seitlich ausgerichteten Visionary-T Sensoren gedacht, um den Einfluss der Fahrt zu ermitteln.

Nr.	Anzahl Frames	Anzahl Personen	TP	FP	FN	Richtige Erkennung (%)	Falscherkennung pro 1000 Bilder
8.2	361	114	57	0	57	50,00	0

Kommentare
Die Person befindet sich ausschließlich im Seitenbereich des Sensors und trägt eine reflektierende Weste. Diese könnte jedoch selbst gut erkannt werden. Dieser Test war für die seitlich ausgerichteten Visionary-T Sensoren gedacht, um den Einfluss der Fahrt zu ermitteln.

Die Ergebnisse verdeutlichen das große Potential von Deep Learning Methoden. Die Erkennungsqualität ist in jedem Testfall innerhalb des zu analysierenden Arbeitsbereiches in einem sehr hohen Bereich bei gleichzeitiger geringer Anzahl an Fehlalarmen, die die Akzeptanz des Systems bei FFZ-Fahrern schnell verringern könnten. Lediglich in Bereichen an den Rändern des Intensitätsbildes kommt es in wenigen Fällen zu Fehlerkennungen bei paralleler Vorbeifahrt an Personen (Szenarien 8.1 und 8.2). Jedoch befinden sich diese Bereiche nicht im analysierten Blickfeld des Sensors. Hier wäre ein größeres Sensorblickfeld oder weitere Sensoren hilfreich die Güte zu verbessern.

Die Ergebnisse bestätigen zudem, dass durch die Verwendung simulierter Trainingsdaten die Implementierung einer Insellösung verhindert wird, die bei der Verwendung von anwendungsspezifischen vom Einsatzort stammenden Trainingsdaten entsteht. Diese Trainingsdatenbasis wird in Abschluss an das finale Treffen des Projektbegleitkreises in Kombination mit den entwickelten Methoden veröffentlicht.

6.7 Konsequenzen aus dem Projektverlauf und der allgemeinen Forschung

Im Rahmen des Projektes bestätigte sich das große Potential von Deep Learning-basierten Methoden zur echtzeitfähigen Personenerkennung. Auch die Arbeiten anderer Forschungsprojekte und Universitäten bestätigen, dass diese Methoden den neuen Standard im Bereich der Objekterkennung darstellen. Die bekannten Limitierungen wie hohe Anzahl an Trainingsdaten sowie aufwendigen Hardwareanforderungen bestätigten sich zunächst ebenfalls im Projektverlauf. Daher wurden die Trainingsdaten anstatt der aufwendige Generierung aus einem einzigen und spezifischen Anwendungsszenario hier über die Sensorsimulation durchgeführt. Auch die Hardwareanforderungen werden in mittlerer Zukunft kein Problem darstellen, da Mikrocontroller immer leistungsfähiger und kostengünstiger werden. Daher ist davon auszugehen, dass mobile Systeme (analog zum PKW-Anwendungsszenario) immer stärker auf Methoden der künstlichen Intelligenz setzen werden. Neben der technischen Realisierbarkeit und Weiterentwicklung der Ansätze ist daher insbesondere eine Rechtssicherheit bzgl. der Verwendung dieser Verfahren herzustellen. Dies wird neben dem im Projekt adressierten Einsatzbereich auch weitere Einsatzfelder betreffen, in denen autonome Fahrzeuge oder mobile Robotersystem selbstständig Aufgaben unternehmen sollen.

6.8 Projektveröffentlichungen, Schutzrechtsanmeldungen und erteilte Schutzrechte (erfolgt oder geplant), Publikationen in Fachzeitschriften und Kongressbeiträge

Folgende Maßnahmen wurden im Projekt umgesetzt. Im Anschluss an das Projekt werden insbesondere die Arbeiten im Bereich Deep Learning veröffentlicht sowie die entwickelten Methoden und die simulierten Trainingsdaten online zur Verfügung gestellt.

Maßnahme	Erläuterung	Teil-Maßnahmen	Zeitraum
Maßnahme A: Vorträge und Veröffentlichungen	Die (Teil-)Ergebnisse des Forschungsprojektes werden bereits während und besonders nach dem Projektende veröffentlicht, um sie einem breiten Fachpublikum zur Verfügung zu stellen.	A1 Publikation in der Fachzeitschrift „Hebezeuge Fördermittel“	3. Quartal 2016
		A2 Vortrag und Veröffentlichung auf der HCII 2017 Konferenz	Juli 2017
		A3 Vortrag und Veröffentlichung auf dem WGTL-Kolloquium	September 2017
		A4 Vortrag auf dem Logistikseminar des Lehrstuhls fml	Oktober 2017
		A5 Veröffentlichung in der Zeitschrift „Technische Sicherheit“	Dezember 2017
		A6 Vortrag und Veröffentlichung auf der HCII 2018 Konferenz	Juli 2018
		A7 Veröffentlichung der Software-Bibliothek zur 2D- und 3D-Bildverarbeitung zur freien Nutzung	2. Quartal 2018
Maßnahme B: Internetdarstellung	Elektronische Verbreitung von Forschungsinhalten und -ergebnissen	B1 Bekanntmachung des Projektes über die Homepage der beteiligten Forschungsinstitute	1. Quartal 2016
		B2 Veröffentlichungen zum Projekt in den Newslettern der beteiligten Forschungspartner	1. Quartal 2015, 3. Quartal 2016, 4. Quartal 2016
		B3 Veröffentlichung auf der Publikationsseite der WGTL (Wissenschaftlichen Gesellschaft für technische Logistik)	1. Quartal 2016

Maßnahme	Erläuterung	Teil-Maßnahmen	Zeitraum
Maßnahme C: Präsentation bei Veröffentlichungen	Ergebnistransfer in die Wirtschaft und Wissenschaft	C1 Vorstellung des Projektes im Rahmen des 26. Deutschen Materialfluss-Kongress	1. Quartal 2017
		C2 Vorstellung des Projektes im Rahmen des BVL International Scientific Symposium on Logistics	2. Quartal 2018
Maßnahme D: Demonstrator und Feldtest	Praktische Umsetzung der erarbeiteten Methodik in einem Demonstrator zur Durchführung eines Feldtestes	D1 Aufbau eines Funktionsdemonstrators am Lehrstuhl fml und BIBA zur Validierung des Systems für den Feldtest und Durchführung von Präsentationen für Interessierte; der Funktionsdemonstrator verbleibt am Lehrstuhl nach Projektende	2017
		D2 Aufbau eines Demonstrators bei einem Industrieanwender mit Schulung spezifischer Fahrer für das System; Durchführung eines Feldtests in realen Einsatz	4. Quartal 2017
Maßnahme E: Dissertation	Die geplante Stelle des Projektbearbeiters an den Forschungseinrichtungen sieht die Möglichkeit zur Promotion vor	E1 Promotionsvorhaben zur Vertiefung und Veröffentlichung der Forschungsinhalte dieses Projektes am Lehrstuhl fml, TU München	2019
		E2 Promotionsvorhaben zur Vertiefung und Veröffentlichung der Forschungsinhalte dieses Projektes am BIBA, Bremen	2019

7. Auflistung der für das Vorhaben relevanten Veröffentlichungen, Schutzrechtsanmeldungen und erteilten Schutzrechte von nicht am Vorhaben beteiligten Forschungsstellen

Es sind keine relevanten Publikationen u. ä. von Dritten zur Forschungsthematik veröffentlicht und keine relevanten Schutzrechtsanmeldungen vorhanden bzw. Schutzrechte erteilt worden.

8. Bewertung der Ergebnisse hinsichtlich des Forschungszwecks/-ziels, Schlussfolgerungen

Seit Projektstart und bis Projektende von PräVISION hat es deutliche Entwicklungsfortschritte im Bereich der Objekterkennung im Allgemeinen und der Personendetektion im Besonderen, insbesondere durch echtzeitfähige tiefe neuronale Netze gegeben. Im Projekt wurden daher wie geplant klassische Methoden der Personenerkennung sowie grundlegend neue Methoden basierend auf Deep Learning untersucht. Um unabhängig von anwendungsspezifischen Trainingsdaten sowie aufwendigen manuellen Labeln der Daten zu sein, wurde im Projekt eine realitätsnahe Simulation einer dynamischen Einsatzumgebung von Flurförderzeugen (FFZ) entwickelt, um damit die erforderlichen Trainingsdaten für die tiefen neuronalen Netze zu generieren. Die Ergebnisse der Feldtests bestätigen das sehr große Potential des entwickelten Ansatzes. Die aktuell benötigte hohe Rechenleistung und Hardwareanforderungen werden zukünftig zu vernachlässigen sein, da die Entwicklungen in diesem Bereich mit hoher Geschwindigkeit voranschreiten.

Jedoch wurde im Projektverlauf ebenfalls deutlich, dass aus Kostengründen eine Ausstattung eines FFZ mit mehreren Kameras (2D/3D) für die Erfassung des kompletten Arbeitsbereiches des FFZ nicht umsetzbar ist. Daher erscheint eine Nutzung eines solchen Systems nur in kritischen Bereichen wie im Rücken des Fahrers sinnvoll. Alternativ könnte zusätzlich ein lenk- und geschwindigkeitsbasiertes, drehbares Kamerastativ die Notwendigkeit mehrerer Kameras vermeiden. Aus Zeitgründen war die Entwicklung eines solchen Systems im Projekt jedoch nicht möglich.

Die im Projekt erarbeiteten Methoden und Softwaresysteme ermöglichen eine Übertragung der Ergebnisse in zahlreiche angrenzende Bereiche, die nachfolgend vorgestellt werden. Dabei ist anzumerken, dass bei Weiterentwicklungen auch auf alle erreichten Teilergebnisse zurückgegriffen werden kann. Das bedeutet, dass Quellcode, der im Rahmen des Projektes entstanden ist, für die Weiterverwendung der Ergebnisse bei den jeweiligen Projektpartnern abgerufen werden kann.

Wie oben aufgeführt ist am BIBA im Rahmen des Projektes ein vollständiges virtuelles Lager sowie eine entsprechende Simulationssoftware entstanden, welche neben der Darstellung von lagertypischen Einrichtungen wie Flurförderzeugen, Stückgütern und Regalen auch Personen in typischen Posen abbilden kann. So ist es möglich, das entstandene virtuelle Lager für den Test neuer sicherheitstechnischer Funktionen an Staplern zu nutzen oder neue Sensoriken zu evaluieren. So können zum Beispiel kritische Lagersituationen zwischen Flurförderzeugen und Personen erstellt sowie das dazugehörige sensorische Abbild der Umgebung abgeleitet und untersucht werden. In diesem Rahmen können echte Sensordaten in der virtuellen Welt und damit ein realistisches sensorisches Umweltabbild einer virtuellen Umgebung erzeugt werden. Da in dieser Umgebung alle Daten zur Erzeugung der sensorischen Informationen aus einem zuvor erstellten Modell stammen, ist beim Erstellen der sensorischen Daten der wahre Ursprung der Daten bekannt. Der Vorteil, der daraus entsteht, ist in verschiedene Richtungen nutzbar. So ist es zukünftig möglich, kritische Lagersituationen in Verbindung mit Personen und die daraus entstehende algorithmische Reaktion eines möglichen Fahrzeuges zu überprüfen und zu evaluieren, ohne die eigentliche Gefahr real nachstellen zu müssen. Die Simulationssoftware wird kontinuierlich weiterentwickelt und deren Funktionsweise der Öffentlichkeit zur Verfügung gestellt.

Das Projekt hat zudem gezeigt, dass die Erkennung von Personen durch sensorische Informationen generisch nur effizient mit Hilfe von Machine-Learning-Techniken gelöst werden kann. Die verwendeten Netzwerke benötigen große Mengen an Eingangsdaten. Im Projekt kommen diese Eingangsdaten aus zuvor aufgezeichneten Kamerafahrten durch verschiedene Lager der STILL GmbH. Für die Erkennung von Personen oder anderen lagertypischen Einrichtungsgegenständen ist häufig ein aufwendiges Labeling notwendig. Dabei beschreibt das Labeling den Vorgang zur Markierung bestimmter Bildmerkmale, um diese für Trainingszwecke einsetzen zu können. In einer virtuellen Umgebung sind die Ursprünge der sensorischen Wahrnehmung bekannt, so dass während der

Erzeugung des sensorischen Abbildes automatisch ein Label erzeugt werden kann, welches für Trainings- oder Evaluationszwecke benutzt werden kann. Dies spart enorme Aufwände während der Entwicklung und führt zu weniger fehleranfälligen Daten, die auf Grund von fehlerhaftem Labeling entstehen können.

Im industriellen Umfeld hat sich herausgestellt, dass nicht alle der hier verwendeten Algorithmen in der Praxis ohne weiteres verwendbar sind. So erfordert Deep Learning, wie es im Projekt verwendet worden ist, eine dedizierte Hardware (Tensor Processing Units TPUs oder Graphical Processing Units GPUs), für die es derzeit noch keine Erfahrung in der Langzeitverfügbarkeit der Bauteile und dem Einsatz auf Flurförderzeugen im industriellen Maßstab gibt. Die ebenfalls im Projekt verwendeten Support Vector Machines zeigen gute Ergebnisse in der Erkennung von Personen und Objekten und bringen die Möglichkeit mit, auf Standardhardware gerechnet zu werden. Somit ist eine technische Verwendung in industriellen Maßstäben ebenfalls bei der Implementierung einer spezifischen Lösung denkbar.

Das Projekt zeigt, dass es möglich ist, ein Assistenzsystem zu erstellen, welches auf klassischer oder multi-modaler Ebene Fahrer vor möglichen Kollisionen mit Menschen warnt. Der industrielle Maßstab ist beim umgesetzten System noch nicht erreicht, was einerseits an der Zuverlässigkeit und andererseits an einer fehlenden geeigneten industriellen Umsetzung auf dem Flurförderzeug liegt. Ferner ist es denkbar, die erarbeiteten Algorithmen auf dedizierten stationärer Sensoriken ohne Einschränkung des Sichtbereichs und dynamischen Einflüssen zu nutzen, um damit Gefahrenpunkte wie beispielsweise Kreuzungen oder Tore abzusichern. Das grundsätzliche Gefahrenpotential beim alltäglichen Umgang mit Flurförderzeugen könnte damit gesenkt, aber bei weitem nicht ausgeschlossen werden. Auf Grund der noch hohen Kosten der Sensorik wäre ein wirtschaftlicher Einsatz überdies nur möglich, wenn der Gefahrenbereich in einem örtlich begrenzten Raum lokalisierbar ist. Für die Nutzung der im Projekt entstandenen Algorithmen fehlen derzeit Verfahren, Machine Learning nach klassischen Methoden der DIN 13849 oder angrenzenden Normlagen zu bewerten, um die Erfüllung von Sicherheitsanforderungen (Performance-Level-Requirements PLr) sicherzustellen. Die im Projekt entwickelten Methoden sind für eine Sicherheitsanwendung daher noch nicht geeignet.

Die im Projekt erzeugten Methoden zur Klassifikation und Merkmalsextraktion aus 2D- und 3D-Bildinformationen sind auch für andere lagertypischen Objekte nutzbar, die im aktuellen Projektrahmen noch nicht vollständig evaluiert werden konnten. So sind die Verfahren auch für die Erkennung und das Tracking (Verfolgung) von Ladungsträgern und Flurförderzeugen nutzbar. Die Verfahren müssen jedoch dafür angepasst werden und auf die speziellen Objekte trainiert werden. Hierfür bietet der Ansatz, dies über die entwickelte Simulationsplattform zu realisieren, sehr großes Potenzial.

9. Aktueller Umsetzungs- und Verwertungsplan

Die Projektpartner planen die Verwertung der Projektergebnisse im Anschluss an das Vorhaben wie folgt:

Die wissenschaftliche Verwertung erfolgt durch das BIBA und den Lehrstuhl FML als gemeinnützige Forschungseinrichtung bzw. universitärer Lehrstuhl, die aus diesem Grunde keine wirtschaftliche Vermarktung der Projektergebnisse anstreben. Das BIBA wird die gewonnenen Erkenntnisse, insbesondere der Sensorsimulation und Deep Learning-Methoden, kurzfristig nach Projektende in wissenschaftlichen Zeitschriften und auf Konferenzen veröffentlichen. Zudem werden die Ergebnisse zur Qualifikation von Studierenden beispielsweise im Rahmen von Vorlesungen und Abschlussarbeiten eingesetzt. Darüber hinaus können die Projektergebnisse durch Ausstellung des Demonstrators einem Fachpublikum präsentiert werden. In diese Aktivitäten werden ausdrücklich die Verbundunternehmen einbezogen, die durch vielfältige Mitgliedschaften in Vereinigungen, Verbänden, Gremien o.ä. ein breites Feld für den Transfer erreichen können. Zudem wird die Sensorsimulation kontinuierlich erweitert und in weiteren Forschungsprojekten eingesetzt. Die Funktionsweise der Simulationssoftware wird kontinuierlich weiterentwickelt und der Öffentlichkeit in Form von wissenschaftlichen Veröffentlichungen zur Verfügung stehen, damit die im Projekt erzielten Ergebnisse für die Weiterentwicklung kamerabasierter Assistenzsysteme von der Forschungsgemeinschaft genutzt werden können.

Der Lehrstuhl FML der TUM wird die Ergebnisse ebenfalls wissenschaftlich verwerten. Im Rahmen eines projektbezogenen Dissertationsvorhabens werden die Optimierungsmöglichkeiten von SVM in Kombination mit den Daten einer Time-of-Flight-Kamera ermittelt. Darüber hinaus wird untersucht, wie gut die Personenerkennung für die Vermeidung von Kollisionen am Gabelstapler sein muss, damit eine objektbasierte Kollisionserkennung möglich ist.

Die SICK AG wird die Erkenntnisse aus dem Projekt für die Weiterentwicklung der Sensorik sowie der Implementierung neuer Anwendungen für Flurförderzeuge (FFZ) verwenden. Insbesondere die simulierten Sensordaten können für die Entwicklung echtzeitfähiger Algorithmen zur Objekterkennung von vielfältigen Objektklassen in intralogistischen Anwendungen genutzt werden.

Für die STILL GmbH bietet insbesondere das konzipierte und entwickelte multimodale Anzeige- und Warnsystem zur Interaktion mit dem FFZ-Fahrer großes Potential, marktreife Assistenzsysteme zu entwickeln. Zudem kann durch die Simulation unterschiedlicher Sensoren an unterschiedlichen Positionen an einem FFZ der Entwicklungsprozess von neuen, intelligenten FFZ beschleunigt werden.

10. Literaturverzeichnis

[Bab-2009] Babenko, B.; Yang, M.-H.; Belongie, S.: Visual tracking with online multiple instance learning. Computer Vision and Pattern Recognition, 2009, S. 983–990.

[Bos-2009] Bostelman, R.: Towards improved forklift safety: White Paper. In: Association for Computing Machinery (Hrsg): Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems. New York: Association for Computing Machinery-2009, S. 297–302.

[Dal-2006] Dalal, N.; Triggs, B.; Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance. In: Leonardis, A.; Bischof, H.; Pinz, A. (Hrsg.): Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, S. 428–441.

[Dan-2014] Danelljan, M.; Khan, F. S.; Felsberg, M.; van de Weijer, J.: Adaptive Color Attributes for Real-Time Visual TrackingComputer Vision and Pattern Recognition - IEEE Conference on, 2014, S. 1090–1097.

[Dgu-2011] BGI/GUV-I 5160: Personenschutz beim Einsatz von Flurförderzeugen in Schmalgängen, Deutsche Gesetzliche Unfallversicherung, http://publikationen.dguv.de/dguv/pdf/10002/i-5160.pdf, Aufruf am 20.03.2018.

[Dol-2012] Dollar, P.; Wojek, C.; Schiele, B.; Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Jg. 34 (2012) Nr. 4, S. 743–761.

[Eri-2004] Ericson, C.: Real-Time Collision Detection. CRC Press, 2004.

[Ewg-89] Richtlinie 89/655/EWG des Rates über Mindestvorschriften für Sicherheit und Gesundheitsschutz bei Benutzung von Arbeitsmitteln durch Arbeitnehmer bei der Arbeit. Vorschriftensammlung der Gewerbeaufsicht Baden-Württemberg, Version 03/2007.

[Far-2003] Farnebäck, G.: Two-frame Motion Estimation Based on Polynomial ExpansionProceedings of the 13th Scandinavian Conference on Image Analysis, Berlin, Heidelberg, 2003, S. 363–370.

[Gij-2007] Van Gijssel et al.: Assistance for the Sovereign Vehicle Driver-A Driver-Vehicle Interface for Sustained Situation Expectancy and Anticipative Vehicle Control, VDI BERICHTE, (2015), S. 251.

[Gün-2014] Günthner, W.A.; Hohenstein, F.; Jung, M.: Das Staple-Auge – Integration bestehender Sensorik durch die Entwicklung geeigneter Verfahren zur optischen Identifikation und Ortung an/von Flurförderzeugen. Forschungsbericht zum IGFVorhaben FV 17244 N, Lehrstuhl für Fördertechnik Materialfluss Logistik, Garching, 2014.

[Goo-2015] Google Brain Team: TensorFlow, 2015.

[Gra-2006] Grabner, H.; Grabner, M.; Bischof, H.: Real-Time Tracking via On-line Boosting. In: Chantler, M.; Fisher, B.; Trucco, M. (Hrsg.): British Machine Vision Conference 2006, 2006, S. 6.1-6.10.

[Hei-2006] Heinecker, M.: Methodik zur Gestaltung und Bewertung wandelbarer Materialflusssysteme. Dissertation; Lehrstuhl für Fördertechnik Materialfluss Logistik, Technische Universität München, 2006.

[Jai-2015] Jaimez, M.; Souiai, M.; Gonzalez-Jimenez, J.; Cremers, D.: A Primal-Dual Framework for Real-Time Dense RGB-D Scene FlowProc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), 2015.

[Kal-2010] Kalal, Z.; Mikolajczyk, K.; Matas, J.: Forward-backward error: Automatic detection of tracking failures. Pattern recognition (ICPR), 2010 20th international conference on. IEEE, 2010, S. 2756–2759.

[Kal-2012] Kalal, Z.; Mikolajczyk, K.; Matas, J.: Tracking-learning-detection. In: IEEE transactions on pattern analysis and machine intelligence, Jg. 34 (2012) Nr. 7, S. 1409–1422.

[Kan-2009] Kany, H.-P.: Unfallgeschehen und mögliche Präventionsmaßnahmen. In: Hebezeuge und Fördermittel, 49 (2009) Sonderheft Flurförderzeuge, S. 48-49.

[Kri-2012] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (2012), S. 1097-1105.

[Lun-2016] Lundkvist, A.;Nykänen, A.: Response times for visual, auditory and vibrotactile directional cues in driver assistance systems, (2016) V4, SAE International Journal of Transportation Safety, S.8-14.

[Pér-2013] Pérez, J. S.; Meinhardt-Llopis, E.; Facciolo, G.: TV-L1 optical flow estimation. In: Image Processing On Line, Jg. 2013 (2013), S. 137–150.

[Red-2016] Redmon, J.; Farhadi, A.: YOLO9000 – Better, Faster, Stronger. In: ArXiv e-prints (2016)

[Ris-2011] Abschlussbericht „Risiko raus“, Deutsche Gesetzliche Unfallversicherung (DGUV), 2011.

[Sim-2014] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition, ICLR 2015.

[Sta-2017] Standke, W.: Arbeitsunfallgeschehen 2016. Herausgegeben von: Deutsche Gesetzliche Unfallversicherung (DGUV) Spitzenverband der gewerblichen Berufsgenossenschaften und der Unfallversicherungsträger der öffentlichen Hand, 2017.

[Ste-1956] Steinhaus, H.: Sur la division des corp materiels en parties. In: Bull. Acad. Polon. Sci, Jg. 1 (1956) Nr. 804, S. 801.

[Stw-2014] Staplercheck 01: Manitou ME 315. Staplerworld, http://www.staplerworld.com/fileadmin/Downloads/STAPLERCHECK/ME_315/STW_06_08_Check_03.pdf, Aufruf am 20.03.2014.

[Sze-2015] Szegedy, C. et al: Going deeper with convolutions, Cvpr.

[Vin-1991] Vincent, L.; Soille, P.: Watersheds in digital spaces – An efficient algorithm based on immersion simulations. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Jg. 13 (1991) Nr. 6, S. 583–598.

[Wei-2016] Liu, W., et al.: Ssd: Single shot multibox detector, European conference on computer vision. Springer, Cham, 2016.

[Wit-2006] Witt, G., Eschenbruch, J,: Entwicklung und Evaluation technischer Lösungen zur Vermeidung von Personenunfällen durch Gabelstapler, Forschungsabschlussbericht, 2006.