Konzept zur Gefährdungserkennung im städtischen Verkehrsraum durch Personendetektion in MLS-Punktwolken

BJÖRN BORGMANN[1,2], MARCUS HEBEL[1], MICHAEL ARENS[1] & UWE STILLA[2]

Zusammenfassung: Die Detektion von personenbezogenen Gefährdungen im städtischen Verkehrsraum stellt eine Aufgabe dar, deren automatische Erfüllung durch ein technisches System z.B. im Hinblick auf die Entwicklung von Fahrerassistenzsystemen wünschenswert ist. In diesem Beitrag wird zunächst das Konzept einer Methode vorgestellt, welche diese Aufgabe unter Verwendung von 3D-LiDAR-Sensoren lösen kann. In der Methode wird die Aufgabe in mehrere Teilschritte zerlegt, zu denen eine Vorverarbeitung, die Detektion und das Tracking von Personen sowie eine abschließende Situationsbewertung gehören. Die zweite Hälfte des Beitrags konzentriert sich auf den Teilschritt der Detektion von Personen und auf die dafür notwendigen Schritte der Datenvorverarbeitung. Die Datenvorverarbeitung generiert dabei eine Reihe von Punktwolkensegmenten, bei denen es sich um Personen handeln könnte. Die Personendetektion erfolgt dann durch Klassifizierung dieser Segmente auf Basis eines „Implicit Shape Models". In ersten Untersuchungen wurde eine Genauigkeit von ca. 66 % bei einem Recall von ca. 76 % erreicht.

1 Einleitung

Besonders im städtischen Verkehrsraum bestehen Gefährdungen für Personen, die z.B. durch Fahrzeuge verursacht werden. Solche Gefährdungen sind für einen Fahrer nur schwer umfassend zu erkennen, da die Vielzahl an Strukturen in städtischen Umgebungen wie z.B. Fahrzeuge, Stadtmöbel, Verkehrsschilder, Ampeln und andere Personen es ihm erschweren, alle Personen in seinem Umfeld rechtzeitig wahrzunehmen und zu erkennen. Oft fehlt ihm ein vollständiges Bild der Lage, weswegen seine Fähigkeit, auf eine mögliche Gefahrensituation zu reagieren, eingeschränkt ist. An dieser Stelle sind technische Systeme wünschenswert, die z.B. in Form eines Assistenzsystems dabei unterstützen, Gefährdungssituationen im Zusammenhang mit Personen im eigenen Umfeld rechtzeitig zu erkennen. Neben dem Einsatz derartiger Systeme in Fahrzeugen ist auch die stationäre Verwendung z.B. bei der Nutzung schwerer Maschinen oder für besondere Ereignisse bzw. Überwachungsaufgaben hilfreich.
Systeme zur automatischen Gefährdungserkennung können mit unterschiedlichem Funktionsumfang ausgestattet sein. Es ist möglich, dass sie lediglich bei der Wahrnehmung des Umfelds unterstützen und z.B. Personen, die ein Mensch evtl. noch nicht selbst wahrgenommen hat, hervorheben bzw. auf diese aufmerksam machen. Wünschenswert ist, dass sie darüber hinaus auch Rückschlüsse aus dem erkannten Verhalten von Personen ziehen und dadurch mögliche Gefahrensituationen selbständig detektieren.

[1] Fraunhofer IOSB, Abteilung Objekterkennung, Gutleuthausstr.1, D-76275 Ettlingen, E-Mail: [bjoern.borgmann, marcus.hebel, michael.arens]@iosb.fraunhofer.de
[2] Technische Universität München, Photogrammetrie und Fernerkundung, Arcisstraße 21, D-80333 München, E-Mail: stilla@tum.de

Bei der Realisierung eines solchen technischen Systems stellt die jederzeitige Erfassung des Umfeldes durch das technische System eine notwendige Voraussetzung dar. Hierfür eignen sich unterschiedliche Arten von Sensoren. Neben Kameras im sichtbaren oder infraroten Spektralbereich gehören dazu auch verschiedenartige Sensoren zur direkten 3D-Erfassung einer Szene. Zu diesen zählen LiDAR-Systeme. Diese sind unter Kenntnis ihrer Position und Ausrichtung in der Lage 3D-Koordinaten von Oberflächen zu messen, welche durch Punktwolken repräsentiert werden können. Solche Systeme sind in der Lage, die Geometrie der beobachteten Oberflächen direkt und unabhängig von den Lichtverhältnissen zu erfassen.
Der vorliegende Beitrag stellt das Konzept einer Methode vor, durch die basierend auf 3D-Punktwolken wie sie von LiDAR-Systemen aufgenommen werden, oben genannte Gefährdungssituationen automatisch erkannt werden sollen. Hierfür sind verschiedene Teilaufgaben zu erfüllen. In diesem Beitrag wird neben dem Konzept der gesamten Methode auf eine dieser zu erfüllenden Teilaufgaben, nämlich die Detektion und Erkennung von Personen, detailliert eingegangen.

2 Aufgabenstellung

Aus der bereits beschriebenen Zielsetzung, personenbezogene Gefährdungen im städtischen Verkehrsraum zu detektieren, lassen sich verschiedene Aufgaben ableiten, die zu lösen sind. Bei dem Design sowie der Umsetzung der gesamten Methode muss beachtet werden, dass die dafür eingesetzten Verfahren schritthaltend mit der Datenaufnahme arbeiten können, da die Zielsetzung vorsieht ein System zu schaffen, welches die Gefährdungen im Live-Betrieb detektiert. Hieraus ergibt sich ein Kriterium für die maximale Laufzeit bzw. den Aufwand der verwendeten Verfahren.
Zunächst ist es erforderlich, Personen in den aufgenommenen Sensordaten zu detektieren. Hierbei gibt es die Schwierigkeit, dass aufgrund der möglichst vollständigen (bis zu 360°) und zeitlich schnell aufeinanderfolgend wiederholten Umgebungserfassung die jeweils aktuell verfügbare Datendichte (d.h. die Anzahl der aufgenommen 3D-Punkte bezogen auf den aufgenommenen Raum) vor allem in größeren Entfernungen gering ist. Personen werden daher oft nur grob erfasst. Das verwendete Detektionsverfahren sollte daher auch bei geringen Datendichten robust funktionieren. Abb. 1 verdeutlicht diese Problematik am Beispiel von Personen, die mit einem Velodyne HDL-64E LiDAR-Sensor in einer Entfernung von ca. 6 m bzw. 12 m aufgenommen wurden. Derartige LiDAR-Sensoren werden zurzeit z.B. im Bereich des autonomen Fahrens verwendet.
Die reine Position einer Person liefert allerdings nur sehr begrenzte Informationen darüber, was die Person gerade tut, wohin sie sich bewegt oder was sie vorhat. Diese Informationen sind jedoch für eine Gefährdungserkennung nützlich und sollten soweit möglich ebenfalls ermittelt werden. Ein weiteres zu lösendes Teilproblem besteht daher darin, bereits detektierte Personen im Zeitverlauf zu verfolgen um Informationen über ihr Bewegungsverhalten zu erhalten. Ebenfalls interessant sind die Körperposen detektierter Personen, da diese Rückschlüsse darüber erlauben was die Personen gerade tun. So haben zwei Personen, die z.B. am Straßenrand stehen und ein Gespräch führen, meist eine andere Pose als eine Person, die aktuell im Begriff ist auf die Fahrbahn zu treten. Es ist daher wünschenswert, Merkmale über die Körperpose wie z.B. die Position der Extremitäten oder auch die Blickrichtung zu ermitteln.

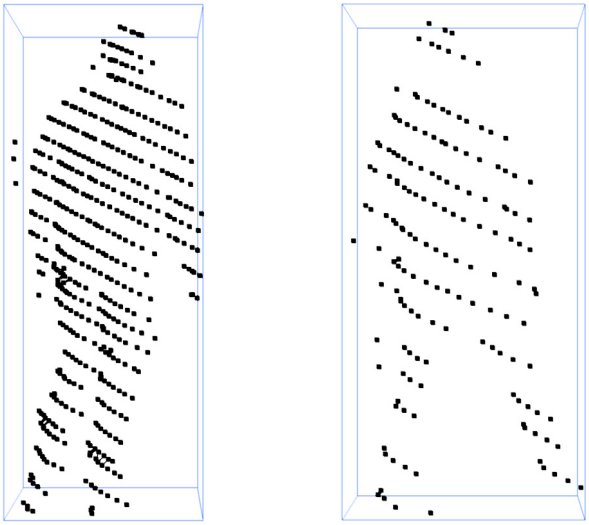

Abb. 1: Beispiele für Personen aufgenommen mit einem Velodyne HDL-64E LiDAR in ca. 6 m (links) und ca. 12 m (rechts) Abstand.

Der letzte Schritt einer automatischen Gefährdungserkennung besteht darin, die gesammelten Informationen auf semantischer Ebene zu einer Situationsbewertung zusammenzuführen und so den eigentlichen Output des Systems zu generieren.

3 Verwandte Arbeiten

In diesem Abschnitt werden verschiedene relevante Arbeiten vorgestellt, die sich bereits mit den im vorherigen Abschnitt angeschnittenen Themen beschäftigten und somit den gegenwärtigen Stand der Technik beschreiben. Insbesondere im Bereich des autonomen Fahrens sowie der Fahrerassistenzsysteme entstanden in den letzten Jahren viele Arbeiten, die sich mit ähnlichen Fragestellungen beschäftigen. So stellen beispielsweise KELLER et al. (2011) ein Fahrerassistenzsystem vor, welches der Fußgängersicherheit dient, indem es bevorstehende Kollisionen mit diesen detektiert und dann selbständig Brems- bzw. Ausweichmanöver durchführt. In ihrer Arbeit verwenden sie ein Stereokamerasystem und fusionieren eine Fußgängerdetektion und eine Detektion bewegter Objekte miteinander. Sie nutzen die so gesammelten Daten um die Kennzahlen TTB (engl. *Time to Break*) und TTS (engl. *Time to Steer*) zu ermitteln, welche die Basis für ihre Situationsanalyse darstellen. Eine umfassende aber bereits etwas ältere Übersicht über verschiedene Aspekte und Ansätze im Hinblick auf Fahrerassistenzsysteme zur Erhöhung der Fußgängersicherheit wurde von GANDHI & TRIVEDI (2007) vorgestellt.

Zur Detektion von Personen in 3D-Daten gibt es unterschiedliche Verfahren. Dazu gehören verschiedene Ansätze, die auf *Support Vector Machines* (SVMs) als Klassifikator zurückgreifen. Bei diesen wird in einer Trainingsphase im Merkmalsraum eine Hyperebene ermittelt, welche die unterschiedlichen Klassen voneinander trennt. NAVARRO-SERMENT et al. (2010) detektieren Personen in LiDAR-Punktwolken und verwenden dafür zwei hintereinander angewendete SVMs. Die erste erhält als Eingabe eine Reihe geometrischer Merkmale, die durch Projektion der Punkte eines

Detektionskandidaten auf zwei verschiedene zweidimensionale Ebenen ermittelt werden. Die beiden Ebenen werden dabei aus der Kombination von jeweils zwei Eigenvektoren der Punktwolke des Detektionskandidaten gebildet. Die zweite SVM erhält als Eingabe das Ergebnis der ersten sowie eine Reihe von Bewegungsmerkmalen, die in einem zuvor erfolgten Segmentations- und Tracking-Schritt ermittelt wurden.
Random Decision Forests sind eine Klassifikationsmethode bei der in einem Trainingsschritt mehrere Entscheidungsbäume mit einem bestimmten Zufallselement unabhängig voneinander trainiert werden. Sie sind so in der Lage dem Problem der Überanpassung (engl. Overfitting) zu begegnen. Sie sind geeignet in einzelnen Tiefenbildern sowohl Personen zu detektieren als auch dazu bestimmte Körperteile zu erkennen (SHOTTON et al. 2011 und SHOTTON et al. 2013). Sie klassifizieren hierfür jedes Pixel des Tiefenbildes dahingehend, zu welchem Körperteil es gehört und sind dabei ausreichend schnell für eine schritthaltende Verarbeitung, erfordern jedoch ein umfangreiches Training und ausreichend hoch aufgelöste Tiefenbilder.
Eine weitere Gruppe von Klassifikationsverfahren stellen *Bag of Words* Methoden dar. Bei diesen werden zunächst Merkmale ermittelt und diese einem „Wort“ (visuelles. bzw. geometrisches Wort) in einem „Wörterbuch“ zugeordnet. Die zugeordneten Wörter wiederum entscheiden in einem Abstimmprozess darüber, welcher Klasse ein Objekt zuzuordnen ist. Hierfür erfolgt zunächst ein Training, bei dem die Wörter sowie die Stimmen der Wörter für Objektklassen trainiert werden. *Implicit Shape Models* (ISM) erweitern klassische *Bag of Words* Methoden dahingehend, dass sie auch berücksichtigen wo an einem Objekt ein bestimmtes Merkmal zu finden ist, wodurch die Korrektheit der Klassifikation verbessert wird. Sie tun dies indem für die Wörter bzw. die den Wörtern zugeordneten Stimmen z.B. hinterlegt ist, wo sich von der Position des Wortes aus gesehen der Objektmittelpunkt befindet. Im Abstimmprozess wird dann ermittelt, an welcher Position die Stimmen für einen Objektmittelpunkt einer bestimmten Klasse konvergieren. Ursprünglich wurden sie für die Verwendung im Bereich der Objekterkennung in Bildern entwickelt (LEIBE et al. 2008, JÜNGLING & ARENS 2011) werden jedoch mittlerweile auch auf 3D-Punktwolken angewendet (KNOPP et al. 2010 und VELIZHEV et al. 2012). Ein an *Implicit Shape Models* angelehntes Verfahren wurde von SPINELLO et al. (2010) vorgestellt. Sie betrachten die Daten zunächst in mehreren flachen horizontal übereinander angeordneten Ebenen und bestimmen auf diesen Segmente und verschiedene Merkmale. Hierbei verwenden sie Methoden, die aus der Verarbeitung von Daten aus 2D-LiDAR-Sensoren (LiDAR-Sensoren mit nur einer Messebene) bekannt sind. Anschließend stimmen die extrahierten Merkmale für Objektpositionen im 3D-Raum ab.

4 Konzeptionelles Design der Methode

In diesem Abschnitt wird ein Gesamtüberblick über unsere Methode zum Erkennen von Gefährdungen im städtischen Verkehrsraum gegeben und die einzelnen Komponenten werden erläutert. Die Methode ist in Abb. 2 schematisch dargestellt. Sie erhält als Eingabe dreidimensionale Punktwolken, die bereits untereinander registriert sind, sodass alle ein gemeinsames Koordinatensystem verwenden. Eine solche Registrierung ist z.B. bei einem stationären System gegeben oder kann durch die Verwendung eines SLAM-Verfahrens (engl. *Simultaneous Localization and Mapping*) bzw. messtechnisch durch den Einsatz eines inertialen Navigationssystems (INS) gekoppelt mit

GNSS-Empfängern erzeugt werden (direkte Georeferenzierung). In diesem Beitrag wird davon ausgegangen, dass eine Registrierung dieser Art bereits vorhanden ist.

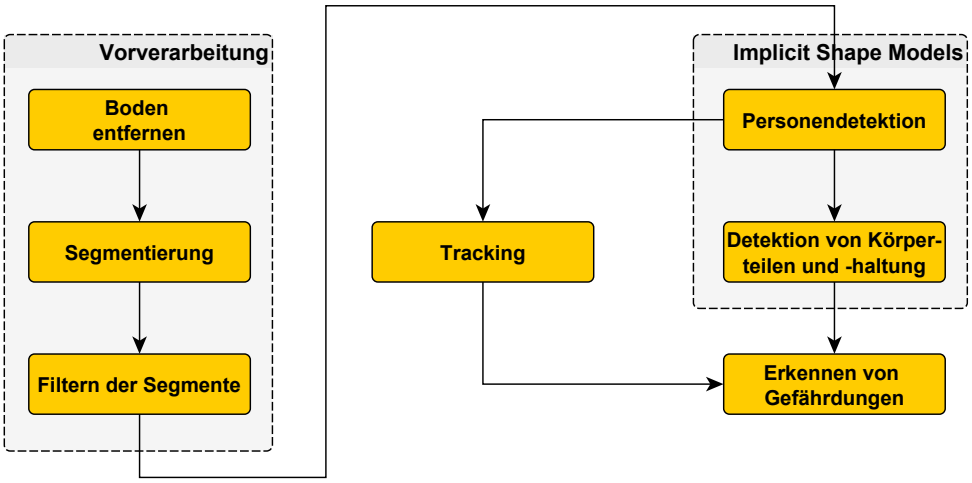

Abb. 2: Überblick über unsere Methodik.

Unsere bisherigen Arbeiten konzentrierten sich auf den Bereich der Vorverarbeitung und der Detektion von Personen sowie in einer späteren Ausbaustufe auf die Detektion von Körperteilen und der Körperhaltung der detektierten Personen. Auf diese beiden Bereiche wird in den beiden folgenden Abschnitten noch näher eingegangen, sie werden daher in diesem Unterabschnitt nur kurz zusammengefasst.

Für die Punktwolken erfolgt zunächst eine Vorverarbeitung, deren Ziel es ist, den Datenumfang mit Methoden kurzer Verarbeitungszeit möglichst stark zu reduzieren und in eine Form zu bringen, die eine parallele Verarbeitung in den nachfolgenden Schritten begünstigt. Für die Detektion von Personen, Körperteilen und der Körperhaltung verwenden wir einen Ansatz, der auf dem Konzept der *Implicit Shape Models* (ISMs) beruht. Diesen möchten wird für die Erkennung von Körperteilen in 3D-Punktwolken erweitern.

Das Tracking dient primär dazu, Informationen über das Verhalten der erfassten Personen im Zeitverlauf zu sammeln. Es wird in späteren Ausbaustufen jedoch auch möglich sein, es zusätzlich zu verwenden, um bereits detektierte Personen im Zeitverlauf nicht in jeder neu erfassten Punktwolke erneut detektieren zu müssen. Denkbar ist z.B., Segmente innerhalb eines Tracks solange nur mit niedriger Priorität bzw. nur in jeder x-ten Punktwolke erneut mit dem Personendetektor zu verarbeiten, wie der Konfidenzwert des Trackings über einem bestimmten Schwellwert liegt. Hierdurch kann eine Laufzeitersparnis erzielt werden, die eine schritthaltende Verarbeitung begünstigt. Für das Tracking ist derzeit die Verwendung eines Kalman-Filter geplant.

Die eigentliche Gefährdungserkennung erhält als Eingabedaten eine Vielzahl sehr unterschiedlicher Informationen. Neben den ermittelten Informationen über die erfassten Personen können noch weitere Daten z.B. über die Eigenbewegung des Sensorträgers in diese einfließen. Hierfür ist daher ein Verfahren notwendig, welches in der Lage ist, basierend auf sehr heterogenen Eingabedaten

zu arbeiten. Denkbar ist hier z.B. die Verwendung eines Verfahrens, welches auf Entscheidungsbäumen basiert. Diese Untersuchungen stehen noch am Anfang und werden Gegenstand unserer zukünftigen Arbeiten sein.

5 Vorverarbeitung

Die einzelnen Schritte der Vorverarbeitung sind in Abb. 2 dargestellt und in Abb. 3 sind ihre jeweiligen Ausgaben zu sehen. Die Vorverarbeitungskette besteht aus einem Entfernen der Bodenfläche, einer Segmentierung mithilfe eines Regionenwachstumsverfahrens (engl. *Region Growing*) und dem Filtern der daraus resultierenden Segmente anhand einiger einfacher Merkmale und Kriterien. Die einzelnen Schritte werden im Folgenden näher erläutert.

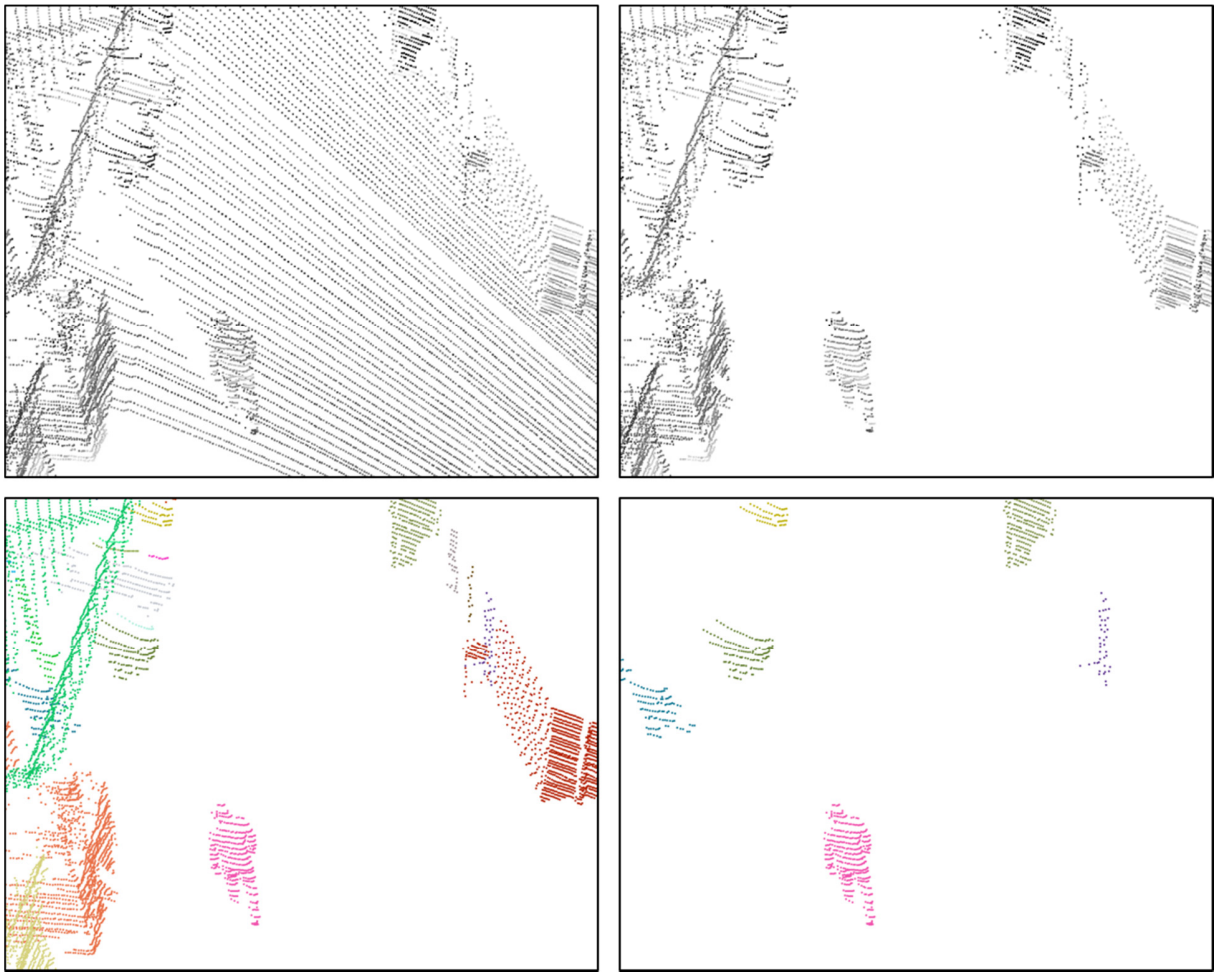

Abb. 3: Verarbeitungsschritte der Vorverarbeitung. Oben links: Eingabedaten. Oben rechts: Bodenebene entfernt. Unten links: Ergebnis der Segmentierung (Segmente sind farbcodiert). Unten rechts: Ergebnis der Filterung (Segmente sind farbcodiert).

5.1 Ermitteln und Entfernen der Bodenebene

Durch das Entfernen der Bodenfläche erfolgt bereits eine deutliche Reduzierung des Datenumfangs. Außerdem erleichtert sie die nachfolgende Segmentierung, da ansonsten während eines Regionenwachstums viele zu unterscheidende Segmente über die Bodenebene miteinander zu einem einzigen Segment verbunden wären. Es müssten dann beim Segmentieren weitere Kriterien wie z.B. die Richtung der Normalen mitberücksichtigt werden, die hierfür jedoch zuvor erst rechenaufwendig bestimmt werden müssten. Für die Entfernung der Bodenebene erzeugen wir zunächst ein niedrig aufgelöstes Bodenraster und bestimmen dann für jeden Punkt seinen Abstand zu der aus diesem Raster gebildeten Bodenebene. Punkte deren Abstand einen bestimmten Grenzwert unterschreiten werden als Teil des Bodens angesehen und entfernt.
Zur Bestimmung des Bodenrasters gehen wir davon aus, dass das Koordinatensystem unserer Eingabedaten über eine Höhenachse verfügt (üblicherweise die *z*-Achse). Außerdem gehen wir davon aus, dass sich der Boden in unseren Daten unten befindet, er also auf der Höhenachse die niedrigsten Koordinaten aufweist (ENU). Zunächst werden alle Punkte anhand der beiden anderen Koordinatensystemachsen in ein zweidimensionales Raster gruppiert. Für jede Rasterzelle werden dann die Koordinaten der *z*-Achse sortiert und der Wert des 0,05 Quantils als Höhenwert der Zelle genutzt. Wir verwenden das 0,05 Quantil und nicht den niedrigsten Wert, um mit möglichen Ausreißern, die sich z.B. aus Fehlmessungen ergeben, umgehen zu können.
Da nicht sichergestellt ist, dass jede Zelle überhaupt über Messwerte auf dem Boden verfügt, erfolgt nun ein Schritt, bei dem in einem rekursiven Vorgehen ausgehend von einer Startzelle versucht wird, alle Zellen des (Boden-)Rasters zu traversieren. Dabei wird ein Kriterium für die maximale Steilheit der Bodenebene angenommen. Wenn eine benachbarte Zelle der gerade aktuellen Zelle einen Höhenwert aufweist, dessen Abweichung vom Höhenwert der aktuellen Zelle dieses Steilheitskriterium nicht verletzt, gilt sie als traversierbar und wird anschließend als aktuelle Zelle verwendet. Zellen, die bei diesem Traversieren nicht erreicht werden können, werden aus dem Bodenraster entfernt. Bei der Ermittlung der Startzelle kommen drei Kriterien zur Anwendung.

1. Höhenwert der Zelle liegt zwischen dem 0,05 und 0,15 Quantil der Höhenwerte aller Zellen. Dies dient ebenfalls zur Abmilderung von Ausreißereffekten.
2. Anzahl der erreichbaren Nachbarzellen. Eine Zelle wird als Startzelle bevorzugt wenn diese Zahl höher ist.
3. Nähe zum Mittelpunkt des Rasters. Eine Zelle wird als Startzelle bevorzugt wenn sie mehr in der Mitte des Rasters liegt.

5.2 Segmentierung und Filtern

Für die Segmentierung wird ein Verfahren zum Regionenwachstum verwendet, bei dem Punkte, die in einem bestimmten Abstand zueinander liegen, als zum selben Segment gehörig angesehen werden. Das einzige Kriterium für die Segmentierung sind also die Koordinaten der Punkte bzw. deren euklidische Abstände.
Nach dem Segmentieren erfolgt ein Filtern, bei dem die Segmente anhand von Merkmalen wie Größe, Seitenverhältnis und Anzahl der Punkte gefiltert werden. Dies dient dazu, Segmente nicht weiter verarbeiten zu müssen, die entweder mit Sicherheit keiner Person zuzurechnen sind bzw. die sich nicht als Person erkennen lassen, da sie beispielsweise zu wenig Punkte aufweisen um für

eine Klassifizierung aussagekräftig genug zu sein. Es wird so also eine weitere Datenreduktion erzielt.

6 Personendetektion

Für die Detektion von Personen verwenden wir eine Variante des „Implicit Shape Model"-Ansatzes. Wir erhoffen uns davon sowohl mit der in Abschnitt 2 erwähnten geringen Datendichte umgehen zu können, als auch den Umfang der benötigten Trainingsdaten zu limitieren.
Die Personendetektion erhält als Eingabe die in der Vorverarbeitung ermittelten segmentierten Punktwolken, wobei wir jedes Segment als mögliche Person betrachten (im Folgenden Personenkandidat genannt). Die Aufgabe besteht also darin, diese Personenkandidaten den Klassen „Person" und „keine Person" zuzuordnen.
Die einzelnen Schritte unseres Verfahrens werden hier zunächst aufgelistet. Anschließend werden in mehreren Unterabschnitten Teilaspekte des Verfahrens näher erläutert. Die Schritte sind:

1. Extrahieren von Merkmalen für die 3D-Punkte des Personenkandidaten.
2. Für jedes extrahierte Merkmal: Suche nach einem dazu passenden geometrischen Wort im Wörterbuch.
3. Abstimmprozess: Jedes ermittelte geometrische Wort gibt eine oder mehrere Stimmen für die Position eines Objektes ab.
4. Gewichten der möglichen Objektpositionen basierend auf den Stimmen für die jeweilige Objektposition bzw. benachbarter Objektpositionen derselben Klasse (aktuell gibt es nur die Klasse „Person". Das Verfahren ist aber in der Lage mit beliebig vielen Klassen umzugehen).
5. Entfernen aller Objektpositionen deren ermitteltes Gewicht unter einem bestimmten Schwellwert liegt.
6. Wenn unter den verbleibenden Objektpositionen derselben Klasse mehrere im geringen Abstand zueinander liegen werden sie zusammengefasst.
7. Ausgabe der verbleibenden Objektpositionen als detektierte Personen.

6.1 Merkmalsextraktion

Die bei der Merkmalsextraktion extrahierten Merkmale dienen zum Beschreiben der geometrischen Wörter des Wörterbuchs und zum lokalen Beschreiben der Punktwolken der verarbeiteten Personenkandidaten. Sie findet daher sowohl im Trainingsprozess als auch im eigentlichen Klassifikationsprozess statt, was im Hinblick auf die schritthaltende Verarbeitung Anforderungen an die Komplexität der Merkmalsextraktion stellt. Es ist natürlich möglich auch komplexe Merkmale zu verwenden, wenn diese dafür im Klassifikationsprozess nur für eine kleinere Auswahl der Daten extrahiert werden. Hierbei stellt sich dann jedoch die Frage, wie eine solche Auswahl der relevanten Daten erfolgt. Wie VELIZHEV et al. (2012) ausführen ist eine solche Auswahl und Ermittlung einiger weniger wohldefinierter Merkmale anfällig für Effekte, die sich durch Rauschen und Verdeckungen ergeben. Deswegen ist es vorzuziehen, Merkmale für sehr viele zufällig ausgewählte bzw. für alle Punkte einer Punktwolke zu extrahieren. In unserem Verfahren verwenden wir die von JOHNSON & HEBERT (1999) vorgestellten *Spin Images* als Merkmalsart, wobei wir planen in Zukunft auch die Verwendung anderer Merkmalsarten zu untersuchen.

6.2 Struktur des Wörterbuchs und Training

Das Wörterbuch umfasst alle geometrischen Wörter die im Klassifikationsprozess verwendet werden. Es stellt somit das Ergebnis des Trainingsprozesses dar. Die geometrischen Wörter selbst verfügen zum einen über ein Merkmalshistogramm, durch welches sie beschrieben werden, zum anderen über ein bzw. mehrere Stimmen für Objektpositionen. Diesen Stimmen wiederum ist neben einem Vektor vom Wort zur Objektposition auch ein Gewichtungsfaktor zugeordnet, es haben also nicht alle Stimmen dasselbe Gewicht.
Im Trainingsprozess erfolgen zunächst für die einzelnen Trainingsdatensätze eine Merkmalsextraktion und eine Bestimmung des Mittelpunkts. Anschließend wird für jedes extrahierte Merkmal ein neues Wort mit einer Stimme für die Objektklasse des Trainingsdatensatzes erzeugt. Dabei wird der ermittelte Mittelpunkt als Objektposition der Stimme genutzt und ihr Gewicht mit 1 initialisiert. Nachdem alle Trainingsdatensätze verarbeitet wurden, erfolgt ein *k*-means Clustering der erstellten Wörter, bei dem ihrer Merkmalsausprägung nach ähnliche Wörter zusammengefasst werden. Hierdurch wird der Umfang des Wörterbuchs reduziert.
Nach dem Zusammenfassen verfügen die hieraus resultierenden fusionierten Wörter über alle Stimmen der ursprünglichen Wörter aus denen sie entstanden sind. Die Stimmen eines Wortes werden anschließend gruppiert nach Objektklassen anhand der Ähnlichkeit der Positionen, für die sie stimmen, ebenfalls zusammengefasst. Hierbei gibt es einen Parameter der definiert bis zu welchem Abstand ähnliche Stimmen zusammengefasst werden. Die danach verbleibenden Stimmen eines Wortes werden so neu gewichtet, dass das Gesamtgewicht aller Stimmen des Wortes wieder 1 ergibt. Ein Wort, welches viele verschiedene Stimmen abgibt, wird hierdurch weniger stark für eine bestimmte Klasse und Position stimmen als ein Wort, welches wenige bzw. nur eine Stimme abgibt. Hierdurch wird das Stimmgewicht von Wörtern, die weniger aussagekräftig sind, gegenüber den aussagekräftigeren Wörtern reduziert.

6.3 Gewichten der möglichen Objektpositionen

Nach dem eigentlichen Abstimmprozess gibt es in unserem Verfahren eine Vielzahl möglicher Objektpositionen, da jedes gefundene geometrische Wort mindestens eine Stimme für eine Position abgegeben hat, meist jedoch sogar deutlich mehr als eine. Aus diesen sollen nun diejenigen bestimmt werden, bei denen es sich tatsächlich um die Position eines Objektes handelt. Dabei ist davon auszugehen, dass Stimmen, die für eine tatsächliche Objektposition abgegeben wurden, für Positionen gestimmt haben, die in einem kleinen Umkreis um die tatsächliche Objektposition liegen. Falsche Stimmen sollten hingegen für zufällig im Raum verteilte Positionen gestimmt haben. Daher wird ein hohes Stimmgewicht auf den Umkreis der tatsächlichen Objektpositionen fallen. Bei der Gewichtung der möglichen Objektpositionen betrachten wir daher das Gewicht anderer möglicher Objektpositionen in der Umgebung, wofür das folgende Vorgehen zur Anwendung kommt.
Zunächst findet eine Normalisierung der Stimmgewichte aller abgegeben Stimmen dahingehend statt, dass ihre Summe 1 ergibt. Dies erlaubt es uns, später die Klassifizierungsentscheidung mit einfachen Schwellwerten zu treffen und dabei robust gegen die schwankende Anzahl abgegebener Stimmen zu sein. Es stellt jedoch auch einen Nachteil dar, wenn ein Personenkandidat tatsächlich

aus zwei oder mehr Personen besteht bzw. sehr viel Rauschen in der Punktwolke des Personenkandidaten vorhanden ist. In diesem Fall würde sich das zur Verfügung stehende Stimmgewicht weiter verteilen was die Gefahr birgt, dass der Schwellwert nicht mehr erreicht werden kann.
Nach der Normalisierung werden für jede mögliche Position die benachbarten möglichen Positionen gesucht und deren Stimmgewichte zu dem Stimmgewicht der Position selbst hinzuaddiert. Hierbei wird die Entfernung zwischen den beiden Positionen mithilfe der Gauß-Funktion zur Normalverteilung berücksichtigt. Je weiter die benachbarte Position entfernt ist, umso geringer ist der Anteil ihres Stimmgewichtes, der addiert wird.
Nach diesem Gewichtungsschritt der möglichen Objektpositionen findet die Klassifizierungsentscheidung anhand eines Schwellwertes statt. Alle Positionen, deren resultierendes Gewicht unter dem Schwellwert liegt, werden entfernt. Da anschließend häufig mehrere Objektpositionen in engem Umkreis zueinander verbleiben, werden diese dann noch zusammengefasst, wobei die Position mit dem höchsten Stimmgewicht dann die verbleibende ist.

7 Experimentelle Untersuchung

Unsere Verfahren bestehend aus der Vorverarbeitung und der Personendetektion wurden sowohl im Hinblick auf ihr Laufzeitverhalten als auch in Bezug auf die Güte ihrer Ergebnisse untersucht. Für die Experimente haben wir auf zwei Datensätze zurückgegriffen, die wir mit dem Messfahrzeug „MODISSA“ aufgezeichnet haben, welches in Abb. 4 dargestellt ist. Dieses Fahrzeug verfügt u.a. über vier LiDAR-Sensoren wobei es sich um zwei Velodyne HDL-64E und zwei Velodyne VLP-16 handelt. Bei unseren Experimenten verwendeten wir nur Messdaten der HDL-64E.

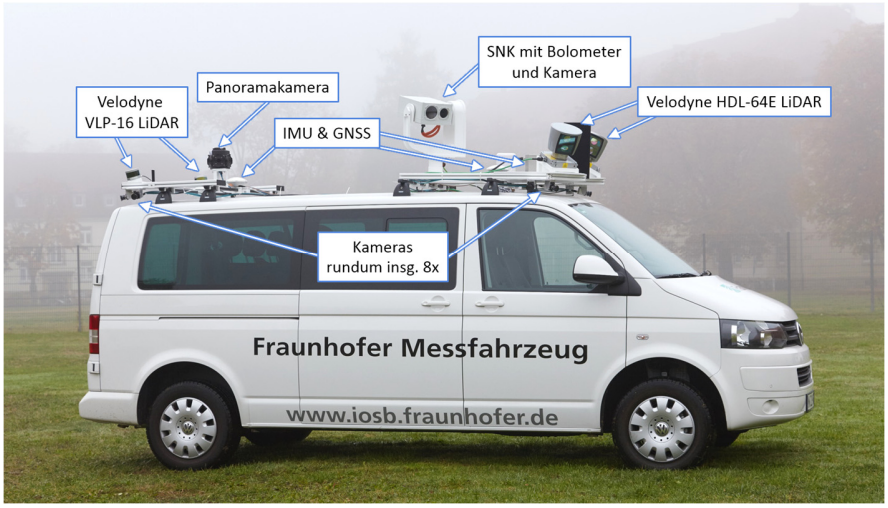

Abb. 4: Sensorfahrzeug „MODISSA“ von Fraunhofer IOSB. Dieses verfügt über verschiedene Sensoren, zu denen neben mehreren Kameras auch vier LiDAR-Sensoren gehören.

Die Datensätze bestehen jeweils aus mehreren Scans beider LiDAR-Sensoren. Als „Scan“ bezeichnen wir dabei eine Umdrehung des Sensors, wobei jeder Scan als einzelne Punktwolke vorliegt. Da die HDL-64E ca. 1,3 Millionen Messungen pro Sekunde durchführen und wir sie mit 10

Umdrehungen pro Sekunde betreiben umfasst ein einzelner Scan ca. 130.000 Messungen. Die daraus resultierenden Punktwolken sind jedoch kleiner, da nicht jede Messung zu einem verwertbaren Ergebnis führt (z.B. fehlende Pulsechos im Bereich des Himmels).
Bei dem ersten verwendeten Datensatz handelt es sich um Daten einer Messkampagne, welche im April 2016 auf dem Stammgelände der Technischen Universität München durchgeführt wurde. Dieser Datensatz enthält 3D-Messdaten einer Vielzahl an Personen im urbanen Umfeld (Campus der TU München). Für den zweiten Datensatz wurden Szenen mit jeweils einer Person gestellt. Dieser wurde primär im Rahmen des Trainings der Personendetektion verwendet. Weitere Trainingsdaten stammten jedoch auch aus dem ersten Datensatz.

7.1 Training und Parametrisierung der Personendetektion

Zur Erstellung der für den Personendetektor verwendeten Trainingsdaten wurden ganze Scans der Datensätze zunächst durch die Vorverarbeitung unseres Verfahrens segmentiert und gefiltert. Anschließend wurden die daraus entstehenden Segmente interaktiv durch visuelle Unterscheidung den Klassen „Person“ und „Keine Person“ zugeordnet, um dadurch die Trainingsdaten zu erzeugen. So entstanden 243 Trainingsdatensätze für die Klasse „Person“ und 1.095 für die Klasse „keine Person“, welche dann im Training verwendet wurden. Es erfolgte also ein Training sowohl mit Positiv- als auch mit Negativbeispielen, was sich gegenüber einem Training nur mit Positivbeispielen als vorteilhaft in Bezug auf die Qualität der Ergebnisse herausgestellt hat.
Wie im vorherigen Abschnitt erwähnt wird im Verlauf des Trainings zunächst für jedes ermittelte Merkmal in den Trainingsdatensätzen ein Wort und eine Stimme generiert, die dann entsprechend ihrer Ähnlichkeit zusammengefasst werden. Für den verwendeten Umfang an Trainingsdaten wurden so zunächst 353.782 Wörter und Stimmen generiert. Bei der anschließenden Zusammenfassung von Wörtern und Stimmen wurde die Verwendung unterschiedlicher Parameter empirisch untersucht. Dabei zeigt sich, dass ein zu starkes Zusammenfassen ähnlicher Wörter und damit ein zu kleines Wörterbuch negative Auswirkungen auf die Laufzeit während der Detektion haben. Dies liegt aber darin begründet, dass bei einem zu starken Zusammenfassen einzelne Wörter durchschnittlich über mehr Stimmen verfügen. Ein effizientes Durchsuchen auch eines größeren Wörterbuchs ist gut mit Suchstrukturen wie *k*-d-Bäumen möglich. Die Berücksichtigung einer größeren Anzahl abgegebener Stimmen im Detektionsprozess kann sich daher stärker auf die Laufzeit auswirken, als die Suche nach einem passenden Wort in einem größeren Wörterbuch. Gute Ergebnisse konnten wir mit einer Wörterbuchgröße von 10 % des ursprünglichen Umfangs erzielen, was sich in der Literatur auch für ähnliche Verfahren bestätigt (KNOPP et al. 2010). Für die einzelnen Wörter wurden Stimmen für Positionen, die bis zu 40 cm voneinander entfernt sind zusammengefasst. Im Ergebnis erhielten wir so ein Wörterbuch bestehend aus 35.378 Wörtern und 163.788 Stimmen, welches wir für die weiteren Untersuchungen verwendet haben.
Die im Hinblick auf die Detektionsrate und zum Teil auch Laufzeit hin wichtigsten Parameter unseres Verfahrens sind der Schwellwert für eine erfolgreiche Erkennung und der Umkreis, in dem benachbarte Stimmen bei der Gewichtung der möglichen Objektpositionen berücksichtigt werden. Hierfür haben wir in dem Experiment zuvor empirisch ermittelte Werte verwendet.

7.2 Laufzeit

Zur Untersuchung der Laufzeit wurden insgesamt 100 Scans verarbeitet und die dafür benötigte Zeit gemessen. Für diese Untersuchung wurde ein Computer mit einem Intel Core i7-3930K Prozessor verwendet. Die implementierten Verfahren liefen auf einer virtuellen Maschine, der sechs virtuelle Kerne des Prozessors zur Verfügung standen. Hierbei zeigte sich, dass das Ziel einer schritthaltenden Verarbeitung zumindest mit der verwendeten Hardware von uns aktuell noch nicht komplett erreicht wird. Im Schnitt benötigte die Verarbeitung eines Scans 1.287 ms, sie schwankt dabei jedoch stark zwischen 268 ms und 6.194 ms. Dieses schwankende Laufzeitverhalten muss von uns noch näher untersucht werden, liegt aber zum Teil in der jeweiligen Komplexität der aufgenommenen Szene begründet. In einigen Szenen lässt bereits die Vorverarbeitung nur noch wenige Daten übrig. In anderen, z.B. wenn viel Vegetation vorhanden ist, müssen mehr Segmente durch das Verfahren zur Personendetektion verarbeitet werden.

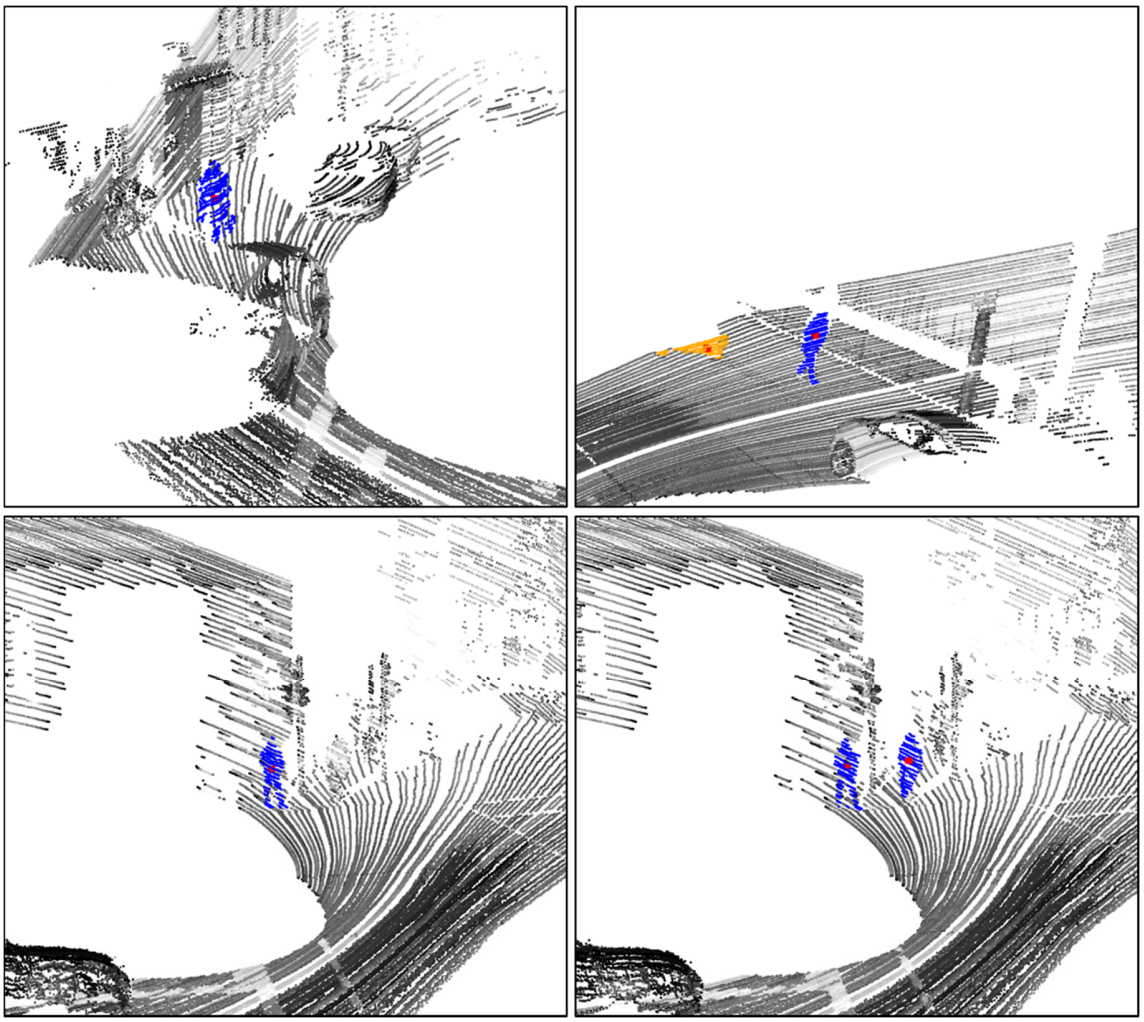

Abb. 5: Detektionsergebnisse: Oben links: Korrekt detektierte Person. Oben rechts: Korrekte (blau) sowie fehlerhafte Detektion (orange), wobei es sich um ein teilweise aufgenommenes Fahrzeug handelt. Unten: Eine Person wird in zwei aufeinanderfolgenden Scans nur im zweiten erkannt.

7.3 Qualität der Personendetektion

Zur Bewertung der Detektionsleistung wurden Ausschnitte aus den auch bei der Laufzeituntersuchung verarbeiteten Daten untersucht, um Auffälligkeiten bei den Resultaten festzustellen. Auch sollte näher untersucht und bestimmt werden, wie viele korrekt bzw. falsch erkannte Personen es in den Datensätzen gibt.
In den untersuchten Daten konnten wir ca. 76 % der vorhandenen Personen detektieren (Recall). Dabei lag der Anteil der korrekten Detektionen bei ca. 66 % (Precision). In Abb. 5 sind verschiedene Beispiele für Ergebnisse unseres Personendetektors dargestellt. Oben rechts sieht man ein Beispiel für eine Fehldetektion, bei der es sich um ein nur teilweise in der Punktwolke sichtbares Fahrzeug handelt. Diese Fehldetektion ist wohl u.a. auch auf Schwächen in unseren Trainingsdaten zurückzuführen, da in den verwendeten Negativbeispielen kaum Fahrzeuge vorhanden waren.
Im unteren Teil der Abbildung ist ein Phänomen beispielhaft dargestellt, welches wir bei der Untersuchung der Detektionsergebnisse mehrerer aufeinanderfolgender Scans häufiger beobachten konnten. Fehldetektionen aber auch Nichtdetektionen traten oft nur in einzelnen Scans auf. Hier besteht die Möglichkeit, durch die Berücksichtigung mehrerer im Zeitverlauf hintereinander aufgenommener Scans die Detektionsergebnisse in diesen gegeneinander zu plausibilisieren. So könnten Personen z.B. erst dann als detektiert gelten, wenn sie in mehreren Scans erkannt werden. Vorübergehende Nichtdetektionen ließen sich auf ähnliche Weise überbrücken, was insbesondere im Hinblick auf vorübergehende Verdeckungen wünschenswert ist. An dieser Stelle spielt natürlich auch das geplante Tracking der erkannten Personen eine Rolle.

8 Fazit und Ausblick

Dieser Beitrag stellt zunächst ein Design einer Methode vor, deren Ziel es ist personengefährdende Situationen mithilfe von LiDAR-Sensoren zu detektieren. Anschließend werden bereits realisierte Teile dieser Methode näher untersucht. Hierbei zeigen wir auf, dass die Detektion von Personen mithilfe von „Implicit Shape Models“ auch in 3D-Punktwolken mit verhältnismäßig geringer Datendichte grundsätzlich möglich ist. Wir zeigen jedoch auch, dass noch weitere Arbeiten und Verbesserungen an unserem Verfahren notwendig sind, um unser Ziel der schritthaltenden Verarbeitung zu erreichen und um die Rate von falsch bzw. nicht erkannten Personen zu reduzieren. Hierfür planen wir zunächst eine umfangreichere Untersuchung der verschiedenen Parameter, die in unser Verfahren eingehen, um für diese bessere Werte zu bestimmen. Außerdem planen wir die Verwendung anderer Merkmalstypen zu untersuchen sowie den Umfang der verwendeten Trainingsdaten zu vergrößern.
Wir erwarten, dass sich die Laufzeit unseres Verfahrens durch verschiedene Maßnahmen in der Zukunft noch verbessern lässt. Potenziale liegen dabei z.B. im Bereich der Parallelisierung. Diese wird zwar bereits genutzt, jedoch zeigt sich im Rahmen einer Untersuchung der Prozessorauslastung, dass es hier noch Verbesserungspotenzial gibt.
In zukünftigen Schritten planen wir zunächst das entwickelte Verfahren zur Personendetektion um die Detektion von Körperteilen und der Körperhaltung zu erweitern und anschließend auch die anderen Komponenten der entworfenen Methode zu realisieren.

9 Literaturverzeichnis

GANDHI, T. & TRIVEDI, M. M., 2007: Pedestrian Protection Systems: Issues, Survey, and Challenges. IEEE Transactions on Intelligent Transportation Systems **8**(3), 413-430.

JOHNSON, A. E. & HEBERT, M., 1999: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(5), 433-449.

JÜNGLING, K. & ARENS, M., 2011: View-invariant person re-identification with an Implicit Shape Model. 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 197-202.

KELLER, C. G., DANG, T., FRITZ, H., JOOS, A., RABE, C. & GAVRILA, D. M., 2011: Active Pedestrian Safety by Automatic Braking and Evasive Steering. IEEE Transactions on Intelligent Transportation Systems **12**(4), 1292-1304.

KNOPP, J., PRASAD, M., WILLEMS, G., TIMOFTE, R. & VAN GOOL, L., 2010: Hough Transform and 3D SURF for Robust Three Dimensional Classification. Proceedings of the 11th European Conference on Computer Vision: Part VI 2010, Springer-Verlag, Heraklion, Crete, Greece, 589-602.

LEIBE, B., LEONARDIS, A. & SCHIELE, B., 2008: Robust Object Detection with Interleaved Categorization and Segmentation. International Journal of Computer Vision **77**(1), 259-289.

NAVARRO-SERMENT, L. E., MERTZ, C. & MARTIAL, H., 2010: Pedestrian Detection and Tracking Using Three-dimensional LADAR Data. The International Journal of Robotics Research **29**(12), 1516-1528.

SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP T., FINOCCHIO M., MOORE, R., KIPMAN, A. & BLAKE, A., 2011: Real-Time Human Pose Recognition in Parts from a Single Depth Image. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, 1297-1304.

SHOTTON, J., GIRSHICK, R., FITZGIBBON, A., SHARP, T., COOK, M., FINOCCHIO, M., MOORE, R. KOHLI, P., CRIMINISI, A., KIPMAN, A. & BLAKE, A., 2013: Efficient Human Pose Estimation from Single Depth Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2821-2840.

SPINELLO, L., ARRAS, K. O., TRIEBEL, R. & SIEGWART, R., 2010: A Layered Approach to People Detection in 3D Range Data. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 1625.

VELIZHEV, A., SHAPOVALOV, R. & SCHINDLER, K., 2012: Implicit shape models for object detection in 3D point clouds. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, **I**-3, 179-184.