Fußgängerbezogene Informationsgewinnung zur Situationsanalyse mit einem mobilen Multisensorsystem

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

Zusammenfassung: Fußgänger gehören im städtischen Verkehrsraum nicht nur zu den verwundbarsten Verkehrsteilnehmern, sondern sind auch aufgrund ihrer Größe häufig nur schwer wahrzunehmen. Technische Systeme wie z.B. autonome Fahrzeuge oder Fahrerassistenzsysteme sind daher besonders gefordert diese zu erkennen. Solche Systeme verfügen häufig über eine Vielzahl an Sensoren um eine Abdeckung des kompletten Umfelds zu erreichen. In solchen Fällen ist eine Form von Datenfusion zwischen den Sensoren notwendig. Der vorliegende Beitrag stellt ein Verfahren zur Personendetektion in 3D-LiDAR Daten eines MLS-Systems vor, welches mehrere LiDAR parallel nutzt, um Personen zu detektieren. Es geht dabei mit Verzerrungseffekten um, die durch eine unvollständige zeitliche Synchronisation der eingesetzten Sensoren in deren Überschneidungsbereichen auftreten können. Solche Effekte lassen sich insbesondere bei scannenden Systemen kaum vermeiden. In einer experimentellen Untersuchung der vorgestellten Methode wird gezeigt, dass sie beim vermehrten Auftreten der beschriebenen Verzerrungseffekte bessere Ergebnisse liefert als einfachere Ansätze der Datenfusion.

1 Einleitung

Fußgänger sind eine bedeutende Gruppe von Verkehrsteilnehmern im städtischen Umfeld. Da sie im Vergleich mit anderen Verkehrsteilnehmern nicht nur verwundbarer sind, sondern außerdem aufgrund ihrer kleineren Größe leichter übersehen werden, sollte ihnen ein besonderes Augenmerk zuteilwerden. Hier besteht auch bei technischen Systemen, z.B. bei Fahrerassistenzsystemen oder autonomen Fahrzeugen, ein Bedarf, Fußgänger automatisch und in Echtzeit zu detektieren. Für solche oft mobilen technischen Systeme, die im Umfeld von Personen verwendet werden, hat eine spezifische Detektion von Personen gegenüber einer unspezifischen Hindernis- bzw. Bewegtobjektdetektion verschiedene Vorteile: Es wird zum einen so möglich, auf die Sicherheit von Fußgängern besonderen Wert zu legen, was zum Beispiel in Form eines größeren Sicherheitsabstands erfolgen kann. Außerdem kann so das besondere Bewegungsverhalten von Personen berücksichtigt werden. So kann eine Person z.B.

Abb. 1: Eine Person aufgenommen von zwei LiDAR-Sensoren

[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

viel abrupter die Richtung wechseln, anhalten oder beschleunigen als ein Fahrzeug.
Oben beschriebene technische Systeme verfügen oft über eine Vielzahl unterschiedlicher Sensoren. Dazu gehören neben Kameras für das sichtbare und infrarote Licht vielfach auch Radar und LiDAR-Sensoren (engl. *Light Detection and Ranging*, z.B. Laserscanner). In der Praxis sind solche Systeme oft auch mit mehreren Instanzen der gleichen Sensorart ausgestattet, um z.B. eine bessere Rundumabdeckung zu erzielen, die sonst konstruktionsbedingt oft nicht erreicht werden kann. Im Überschneidungsbereich mehrerer Sensoren wird zudem eine höhere Datendichte erzielt. Durch die Verwendung verschiedener Sensorarten kann von den individuellen Vorteilen der verschiedenen Sensoren profitiert werden. Die Verwendung mehrerer Sensoren stellt aber auch eine zusätzliche Herausforderung dar. So muss eine Form von Datenfusion erfolgen. Selbst bei der Fusion gleichartiger Sensoren kann es aufgrund einer nicht perfekten geometrischen Registrierung oder einer unvollständigen zeitlichen Synchronisation der Sensoren zu Verzerrungseffekten kommen. Wenn sich solche nicht vermeiden lassen, sollte die weitere Datenverarbeitung mit ihnen umgehen können. Abb. 1 stellt das Ergebnis einer direkten Datenfusion zwischen zwei scannenden LiDAR-Sensoren dar, d.h. eine räumliche Überlagerung der aktuellen Punktwolken. Ein Teil der Person wurde vom ersten Sensor erfasst, ein anderer Teil leicht zeitlich versetzt vom zweiten. Da scannende Sensoren ihre Umgebung nach und nach abtasten ist eine wirkliche zeitliche Synchronisation zwischen zweien solcher Sensoren in der Praxis nur schwer zu erreichen.
Die vorliegende Arbeit beschäftigt sich mit einem Verfahren zur Detektion von Fußgängern in MLS-Daten (engl. *Mobile Laser Scanning*) eines mobilen Sensorträgers, der über mehrere LiDAR-Sensoren verfügt. Das Verfahren versucht dabei, die gerade beschriebenen Probleme der Verzerrungseffekte zu lösen.

2 Verwandte Arbeiten

Im Bereich der Objekt- bzw. Personendetektion gibt es eine Vielzahl verschiedener Verfahren. Häufig wird auf eine Verarbeitungskette zurückgegriffen, in der die Daten zunächst segmentiert und die Segmente dann klassifiziert werden. Eine Gruppe von Verfahren verwenden als Klassifikator *Support Vector Machines* (SVM). Bei diesen wird in einem Trainingsprozess eine Hyperebene ermittelt, die die Trainingsdaten in einem Merkmalsraum in verschiedene Klassen trennt. Die ermittelte Hyperebene wird dann verwendet um unbekannte Daten zu klassifizieren. NAVARRO-SERMENT et al. (2010) verwenden zwei hintereinander angewendete SVMs zur Klassifizierung von Punktwolkensegmenten. Die erste SVM verwendet dabei eine Reihe geometrischer Merkmale als Eingabe. Ihre Ausgabe wird dann um verschiedene Merkmale aus einem Trackingverfahren ergänzt und als Eingabe der zweiten SVM verwendet um Personen zu erkennen. PREMEBIDA et al. (2014) verwenden eine Datenfusion zwischen LiDAR- und RGB-Sensoren. Bei dieser werden die LiDAR-Sensoren genutzt um für die RGB-Bilder Tiefeninformationen zu liefern. Die SVM wird dann auf diese Tiefenbilder angewendet.
Eine weitere Gruppe von Klassifikatoren greifen auf „Wörterbücher“ zurück, die im Rahmen eines Trainingsprozesses generiert werden. Die „Wörter“ dieser Wörterbücher sind üblicherweise über einen Merkmalsvektor beschrieben und Stimmen über die Klassifizierung der verarbeiteten Daten ab. BEHLEY et al. (2013) verwenden eine Reihe von Wörterbüchern um Segmente

von LiDAR-Daten zu klassifizieren. Die Wörterbücher verwenden dabei unterschiedliche Merkmalstypen bzw. unterschiedlich parametrisierte Merkmale. Dies erlaubt es, mit verschiedenen Eigenschaften der verarbeiteten Segmente umzugehen, z.B. mit unterschiedlicher Datendichte. Es ist denkbar, ein solches Verfahren für die Verarbeitung von verzerrten Daten, wie sie z.B. aus der Fusion mehrerer Sensoren entstehen, zu erweitern und auch hierfür Wörterbücher speziell zu trainieren.
Eine Erweiterung der wörterbuchbasierten Methoden stellen *Implicit Shape Models* (ISM) dar. Diese berücksichtigen nicht nur das Vorhandensein bestimmter Merkmale, sondern auch deren geometrische Verteilung. Sie tun dies, indem die Wörter nicht mehr nur für eine Klasse stimmen, sondern auch für eine Position der Klasse relativ zu der Position des Wortes. Eine Klassifizierung findet dann statt, wenn vermehrt für eine ähnliche Position gestimmt wird. ISM wurden ursprünglich für die Verarbeitung von Bilddaten verwendet (LEIBE et al. 2008), finden aber auch für die Verarbeitung von 3D-Daten Verwendung. KNOPP et al. (2010) verwenden ein 3D-ISM und 3D-SURF Merkmale zur generellen Klassifikation von Objekten in 3D-Daten. VELIZHEV et al. (2012) verwenden eine Verarbeitungskette für Punktwolken, die über eine Bodenextraktion und eine auf dem *Region Growing* (engl. Regionenwachstum) basierende Segmentierung verfügt, um anschließend die Punktwolkensegmente mithilfe von ISM in Objektklassen wie Fahrzeug oder Straßenlaternen zu klassifizieren. Bei den von ihnen verarbeiteten Daten handelt es sich um eine Fusion von mehreren zeitlich aufeinanderfolgend aufgenommenen Punktwolken aus einem urbanen Gebiet. Bewegte Objekte werden daher von ihnen nicht erkannt. Wir selbst haben einen ähnlichen Ansatz verwendet um Personen in Einzelscans eines MLS-Systems zu erkennen (BORGMANN et al. 2017).
SHOTTON et al. (2011 & 2013) verwenden *Random Decision Forests* um einzelne Pixel eines Tiefenbildes zu klassifizieren. Sie nutzen dies um sowohl Personen als auch deren Körperteile zu detektieren und zu verfolgen. Die Tiefenbilder sind dabei verhältnismäßig hoch aufgelöst. Durch die Klassifizierung einzelner Pixel entfällt die Notwendigkeit zuvor eine Segmentierung durchzuführen. Bei Random Decision Forests handelt es sich um eine Ansammlung von mit unterschiedlichem Zufallselement trainierten Entscheidungsbäumen. Durch das Zufallselement und die Verwendung mehrerer Bäume wird das Problem der Überspezialisierung einzelner Entscheidungsbäume gelöst.
In den letzten Jahren werden vermehrt *Convolutional Neural Networks* (CNN) zur Objekterkennung sowohl in RGB-Bildern als auch in Tiefenbildern (SOCHER et al. 2012) und in Volumenrepräsentationen von 3D-Daten (MATURANA & SCHERER 2015, GARCIA-GARCIA et al. 2016) eingesetzt.

3 Methode

In diesem Abschnitt stellen wir unsere Methode zur Detektion von Personen (oder anderen Objekten) in LiDAR-Daten vor. Wir gehen dabei davon aus, dass die entsprechenden Daten durch mehrere LiDAR-Sensoren parallel erfasst werden und für die Sensoren getrennt als zeitlich aufeinanderfolgende Punktwolken vorliegen. Eine Punktwolke umfasst dabei jeweils einen Einzelscan (z.B. eine Rotation) des jeweiligen Sensors. Die Sensoren sind zueinander geometrisch kalibriert, aber wie in der Einleitung dieser Arbeit beschrieben nicht exakt zeitlich synchronisiert

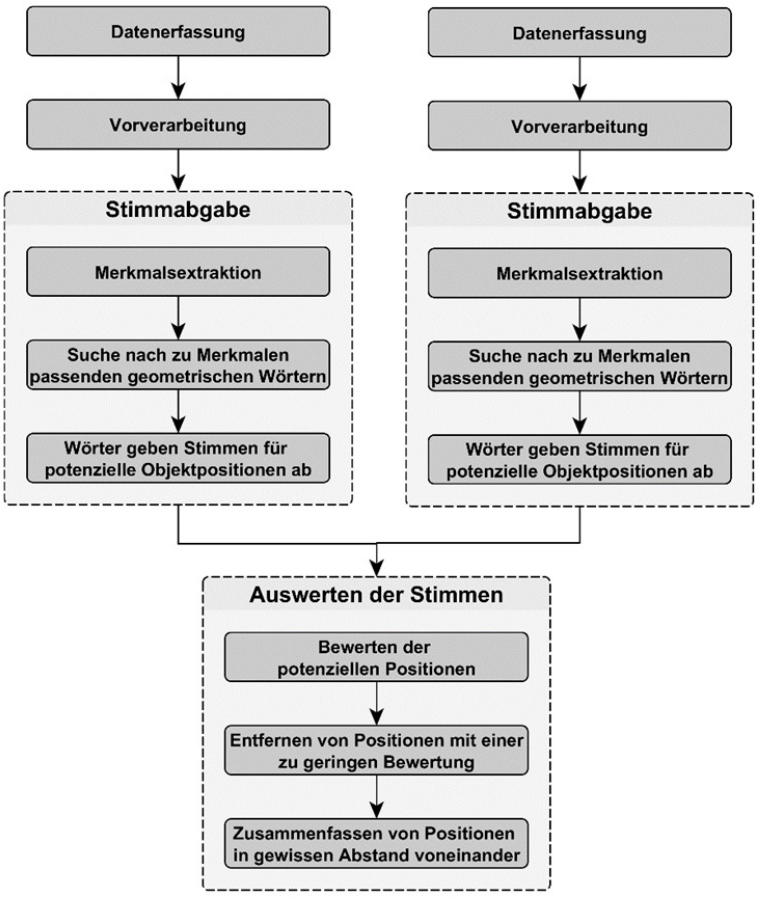


Abb. 2: Schematische Darstellung unseres Verfahrens zur Detektion von Personen in den Daten mehrerer LiDAR Sensoren. Hier dargestellt für zwei Sensoren.

bzw. synchronisierbar. Die Methode muss daher mit den auftretenden temporalen Verzerrungseffekten umgehen.

Der Ansatz stellt eine Weiterentwicklung eines Ansatzes dar, den wir bereits in der Vergangenheit vorgestellt haben (BORGMANN et al. 2017). Wir haben ihn modifiziert und ergänzt. Es handelt sich um ein stimmbasiertes Verfahren welches an 3D *Implicit Shape Models* (ISM) angelehnt ist. Es geht bereits davon aus, dass die abgegebenen Stimmen nicht komplett exakt sind, es also gewisse Unsicherheiten im Stimmraum des Verfahrens gibt. Dies erlaubt es uns eine Datenfusion zwischen mehreren Sensoren in diesem Raum vorzunehmen, ohne dass evtl. auftretende zeitlich bedingte Verzerrungseffekte zwischen den unterschiedlichen Sensoren das Ergebnis stark beeinträchtigen. Eine solche Fusion für die tatsächlichen Punktwolken durchzuführen birgt die Gefahr, dass die Extraktion von Punktmerkmalen durch die Verzerrungseffekte beeinträchtigt wird.

Unser Verfahren ist in Abb. 2 dargestellt und besteht aus drei Komponenten: Einer Vorverarbeitung zur Datenreduktion, der Abgabe von Stimmen und dem Auswerten der abgegebenen Stimmen. Während wir die ersten beiden Komponenten für jeden Sensor getrennt durchführen, erfolgt die dritte Komponente für alle Sensoren gemeinsam. Das Verfahren basiert auf einem Wörterbuch bestehend aus geometrischen Wörtern und verwendet Merkmalsvektoren, um sowohl die

zu verarbeitenden Daten als auch die Wörter im Wörterbuch zu beschreiben. Im Folgendem werden wichtige Komponenten unseres Verfahrens näher beschrieben.

3.1 Merkmalsextraktion

Wir verwenden eine Merkmalsextraktion um Merkmale zu erhalten, welche die lokale Nachbarschaft eines Punktes der gerade verarbeiteten Daten beschreiben. Die Annahme hierbei ist, dass die jeweilige lokale Nachbarschaft Rückschlüsse auf das Objekt zulässt, zu dem der Punkt gehört und es so erlaubt dieses zu klassifizieren bzw. zu detektieren. Für 3D-ISM Verfahren werden im Hinblick auf die Merkmalsextraktion zwei Strategien verfolgt. Eine ist es, komplexere Merkmale für eine kleine Gruppe von wohlausgewählten Punkten zu bestimmen (KNOPP et al. 2010). Die andere Strategie ist es, „weniger leistungsfähige“ Merkmale für eine größere Anzahl zufällig ausgewählter (VELIZHEV et al. 2012) oder gar aller Punkte zu ermitteln. Diese Variante ist robuster gegen verschiedene Störeffekte und Verdeckungen. In unserem Verfahren verwenden wir die zweite Strategie und extrahieren *Fast Point Feature Histograms* (FPFH) (RUSU et al. 2009) für alle Punkte die wir verarbeiten. Wir erhoffen uns von den FPFH Merkmalen, dass diese auch bei verhältnismäßig geringen Datendichten gute Ergebnisse liefern.

3.2 Wörterbuch

Das Wörterbuch enthält geometrische Wörter, welche die Stimmen abgeben die zur Detektion von Objekten verwendet werden. Es ist das Ergebnis des Trainingsprozesses unserer Methode. In Abb. 3 wird die Struktur des Wörterbuchs dargestellt. Es besteht aus mehreren Wörtern, die jeweils durch einen Merkmalsvektor beschrieben sind und mehrere Stimmen abgeben können. Jede Stimme wiederum stimmt für das Vorhandensein eines Objektes einer bestimmten Klasse an einer bestimmten Position (relativ zum Wort) ab. Als Position ist hierbei der Mittelpunkt des Objektes definiert. Jede Stimme hat außerdem ein Stimmgewicht, welches im Trainingsprozess ermittelt wird.

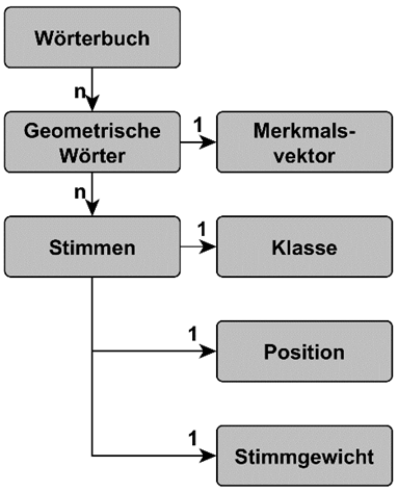


Abb. 3: Struktur des verwendeten Wörterbuchs

Das Training unserer Methode erfolgt in zwei Schritten und erhält als Eingabe Trainingsdaten, bei denen es sich um vorklassifizierte Punktwolkensegmente handelt, die jeweils ein Objekt bzw.

eine Person umfassen. Im Fokus dieses Beitrags haben wir Trainingsdaten der Klassen „Person“ und „keine Person“ verwendet. Das Training mit solchen Negativbeispielen hat sich als vorteilhaft gegenüber einem reinen Training mit Positivbeispielen herausgestellt, da es hierdurch möglich wird, personentypische Merkmale von solchen zu unterscheiden, die bei unterschiedlichsten Objektklassen auftreten. Im ersten Schritt des Trainings erfolgt für jeden Punkt jedes Trainingsdatensatzes eine Merkmalsextraktion. Die dabei ermittelten Merkmale sowie die relative Position des jeweiligen Punktes zum Objektmittelpunkt des Trainingsdatensatzes werden dann genutzt um neue Wörter mit jeweils einer Stimme zu initialisieren.
Nachdem alle Trainingsdaten verarbeitet wurden erfolgt der zweite Schritt des Trainings. Hierbei werden im Merkmalsraum ähnliche Wörter mithilfe eines *k*-means Clustering zusammengefasst und der Umfang des Wörterbuchs so reduziert. Bei diesem Prozess erben die zusammengefassten Wörter zunächst alle Stimmen der ursprünglichen Wörter. Anschließend werden für jedes Wort und jede Klasse getrennt voneinander ähnliche Stimmen, also Stimmen bei denen die Position ähnlich ist, ebenfalls zusammengefasst. Zuletzt wird das Stimmgewicht jeder Stimme ermittelt, wofür folgende Formel Anwendung findet:

$$G_S(w_i) = \frac{1}{N_S(w_i)}$$

$G_S(w_i)$: Gewicht jeder Stimme vom Wort w_i
$N_S(w_i)$: Anzahl der Stimmen des Wortes w_i

Hierdurch wird das Gesamtgewicht der Stimmen eines Wortes normiert. Außerdem bevorzugen wir so Stimmen von aussagekräftigen Wörtern (Wörter, die nur wenige unterschiedliche Stimmen abgeben) gegenüber denen von weniger aussagekräftigen Wörtern (Wörter, die viele verschiedene und widersprüchliche Stimmen abgeben).

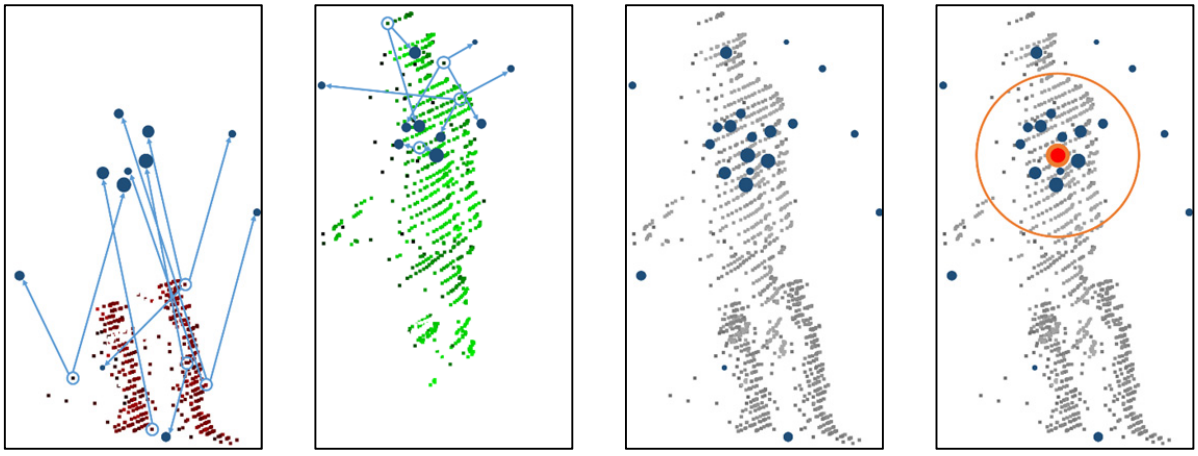
3.3 Vorverarbeitung

Die Vorverarbeitung stellt den ersten Schritt der Datenverarbeitung unserer Methode dar und dient dazu, das Laufzeitverhalten der Methode durch eine Datenreduktion zu verbessern. In der Vergangenheit bestand unsere Vorverarbeitung aus mehreren Schritten und umfasste eine Bodenextraktion, eine Segmentierung und ein Ausfiltern von Segmenten, die aufgrund einfacher geometrischer Merkmale keine Person sein konnten (BORGMANN et al. 2017). Ähnliche Vorverarbeitungsketten finden sich auch in vergleichbaren Methoden (VELIZHEV et al. 2012).
Obwohl dies meist gut funktioniert, kann es insbesondere bei Punktwolken aus Einzelscans eines MLS-Systems zu Problemen kommen. Solche Punktwolken weisen eine verhältnismäßig geringe Datendichte und viele Verdeckungen auf. Dies erschwert die korrekte Segmentierung der Daten, sodass es vermehrt zu einer Über- oder Untersegmentierung kommt. Solche Segmentierungsfehler führen nachfolgend zu Schwierigkeiten beim Filtern. Insbesondere im Falle einer Übersegmentierung wird zudem die eigentliche Detektion erschwert, da der Detektor ebenfalls auf den Punktwolkensegmenten arbeitet. Im Falle einer zu starken Segmentierung einer Person sind dann im einzelnem Segment nicht mehr genügend Informationen für eine erfolgreiche Detektion vorhanden.

Im Rahmen dieser Arbeit haben wir daher die Vorverarbeitung auf eine Bodenextraktion beschränkt. Durch diese lässt sich für unsere Anwendungsfälle bereits eine Datenreduktion von ca. 45% erreichen (BORGMANN et al. 2017). Damit der Detektor auch bei unsegmentierten Daten korrekt funktioniert sind einige Anpassungen nötig, welche im Abschnitt 3.5 erläutert werden.

3.4 Stimmabgabe

Zur Abgabe von Stimmen erfolgt zunächst die Merkmalsextraktion. Die ermittelten Merkmale werden dann verwendet um im Wörterbuch das zum jeweiligen Merkmal am besten passende Wort zu suchen. Für diese Suche im Wörterbuch greifen wir auf einen Verfahren zur schnellen approximativen Nachbarschaftssuche zurück welches von MUJA & LOWE (2009) vorgestellt wurde. Dieses wenden wir auf den Merkmalsraum der Wörter des Wörterbuchs an. Die Stimmen der gefundenen Wörter werden dann basierend auf der Stimmposition und der Position des Wortes im 3D-Raum in diesen projiziert.
In Abb. 4a) und b) wird die Stimmabgabe beispielhaft für die Verarbeitung von Messdaten zweier Sensoren dargestellt. Es ist zu beachten, dass anders als in der Darstellung in unserer Methode tatsächlich für alle Punkte Stimmen abgegeben werden. Wie bereits erläutert erfolgt die Stimmabgabe für beide Sensoren getrennt. Dies verhindert gegenüber der direkten Fusion der Punktwolken Probleme bei der Merkmalsextraktion, die durch zeitliche Verzerrungseffekte verursacht werden. Der Nachteil hierbei ist, dass wir bei der Merkmalsextraktion nicht von einer potenziell höheren Datendichte im Überlappungsbereich profitieren können. Wir gehen jedoch davon aus, dass die Vorteile hier die Nachteile überwiegen.



a) Beispiele von Stimmen für Daten von Sensor 1

b) Beispiele von Stimmen für Daten von Sensor 2

c) Fusion der Stimmen beider Sensoren

d) Bewerten einer Position basierend auf Stimmen der Umgebung

Abb. 4: Illustration unseres Detektors für die Verarbeitung von Daten zweier LiDAR-Sensoren

3.5 Bewerten der Stimmen

Nach der Stimmabgabe erfolgt die weitere Verarbeitung aller abgegebenen Stimmen aller Sensoren gemeinsam. Der Zustand zu Beginn dieses Schritts ist beispielhaft in Abb. 4c) dargestellt. Wir versuchen nun zu entscheiden, bei welchen der abgegebenen Stimmen es sich tatsächlich um

eine Person handelt. Hierfür betrachten wir jede Stimme als eine potenzielle Position eines Objektes und nehmen eine Bewertung von diesen vor, und zwar basierend auf ihrem eigenen Gewicht und dem Gewicht von anderen potenziellen Positionen derselben Klasse in der Nachbarschaft. Als Nachbarschaft definieren wir dabei einen gewissen Radius um die gerade bewertete Position. Die Bewertung ist in Abb. 4d) illustriert und basiert auf der Gaußschen Normalverteilung. Sie addiert einen Teil des Gewichts benachbarter Positionen zu dem Gewicht der gerade bewerteten Position. Die Annahme hierbei ist, dass ein großes Stimmgewicht auf das Gebiet um die tatsächliche Position eines Objektes fallen wird. Wir verwenden für die Bewertung folgende Formel:

$$B_p = \sum_{k \in K} W_k \times RW_{\text{norm}} \times e^{-\frac{D_{pk}^2}{2 \times \sigma^2}}$$

B_p : Bewertetes Gewicht von Position p
K : Alle potenziellen Positionen mit derselben Klasse wie Position p im Bewertungsradius
W_k : Nicht bewertetes Gewicht der Position k
RW_{norm} : Normierungsfaktor für den Bewertungsradius R
D_{pk} : Euklidische Distanz zwischen Positionen p und k
σ : Breite der Normalverteilung.

Die Formel verfügt über einen Normierungsfaktor, welcher es uns erlaubt mit unterschiedlichen Datendichten umzugehen und dabei trotzdem später einen einfachen Schwellwert als Detektionsentscheidung anwenden zu können. Anders als in unseren früheren Arbeiten basiert der Normierungsfaktor in dieser Arbeit auf der Anzahl an potenziellen Positionen im Bewertungsradius. Früher basierte er auf der Anzahl an potenziellen Positionen in dem gerade verarbeiteten Segment der Daten. Dieser Änderung wurde notwendig um auf eine Segmentierung verzichten und um mit Stimmen aus unterschiedlichen Punktwolken umgehen zu können. Diese ergeben sich aus der parallelen Verarbeitung für mehrere Sensoren. Der Normierungsfaktor wird wie folgt berechnet:

$$RW_{\text{norm}} = \frac{1}{N(P_R)}$$

RW_{norm} : Normierungsfaktor für den Bewertungsradius
$N(P_R)$: Anzahl Positionen im Bewertungsradius R

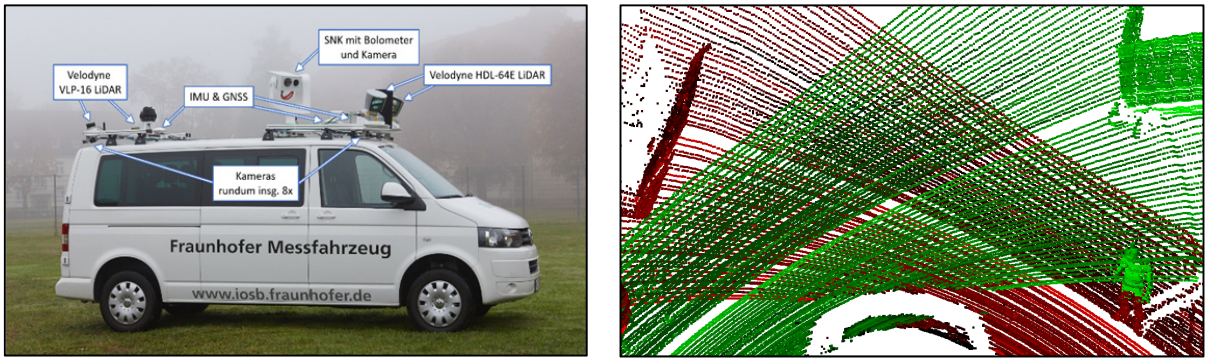
Diese Variante der Normierung ist anfällig in Regionen, in denen es nur sehr wenige Stimmen in der Nachbarschaft gibt. Daher gibt es einen Schwellwert für die Mindestanzahl an Nachbarn, die eine potenzielle Position in ihrem Bewertungsradius haben muss, um weiter verarbeitet zu werden.

4 Experimentelle Untersuchung

In diesem Abschnitt beschreiben wir die von uns durchgeführten Experimente. Der Fokus liegt dabei auf der Verarbeitung von Daten mehrerer LiDAR-Sensoren. Wir beschreiben zunächst den Aufbau und Ablauf der durchgeführten Experimente. Danach stellen wir deren Ergebnisse vor und analysieren diese.

4.1 Ablauf

Für unsere Experimente benötigten wir 3D-Messdaten von Personen, welche von mehreren LiDAR-Sensoren erfasst werden. Derartige Daten haben wir mit unserem am Institut vorhandenen Messfahrzeug *MODISSA* erstellt. Dieses ist in Abb. 5a) dargestellt und verfügt unter anderem über zwei Velodyne HDL-64E Laserscanner, welche wir für unsere Experimente verwendet haben. Bei den Sensoren handelt es sich um scannende LiDAR-Sensoren mit einem rotierenden Aufnahmekopf. Sie sind jeweils in der Lage 1,3 Millionen Messungen pro Sekunde durchzuführen, welche sich auf 64 Scanzeilen verteilen. Durch den rotierenden Kopf ist der horizontale Sichtbereich der Sensoren 360°, der vertikale ist 26,9°. Für die Experimente wurden die LiDAR-Sensoren mit 10 Umdrehungen pro Sekunde betrieben und für jeden Sensor alle 0,1 Sekunden eine Punktwolke generiert. Jede Punktwolke umfasst also ca. 130.000 Messungen. Die tatsächliche Anzahl der Punkte ist jedoch geringer, da Teile der Messungen zu keinem Ergebnis führen. Dies ist z.B. der Fall, wenn der Laserpuls auf kein Objekt trifft („Messungen in den Himmel“). Durch das am Fahrzeug vorhandene GNSS/IMU-System ist eine direkte Georeferenzierung der aufgenommenen Daten möglich. Diese erlaubt es die Eigenbewegungen des Fahrzeugs zu kompensieren. Die verbleibenden Verzerrungen in den Daten sind daher auf die Bewegungen der aufgenommenen Objekte und Personen zurückzuführen.



a) Das verwendete Messfahrzeug MODISSA b) Beispiel für die verwendeten Daten

Abb. 5: Experimentalsystem und Ablauf der Experimente

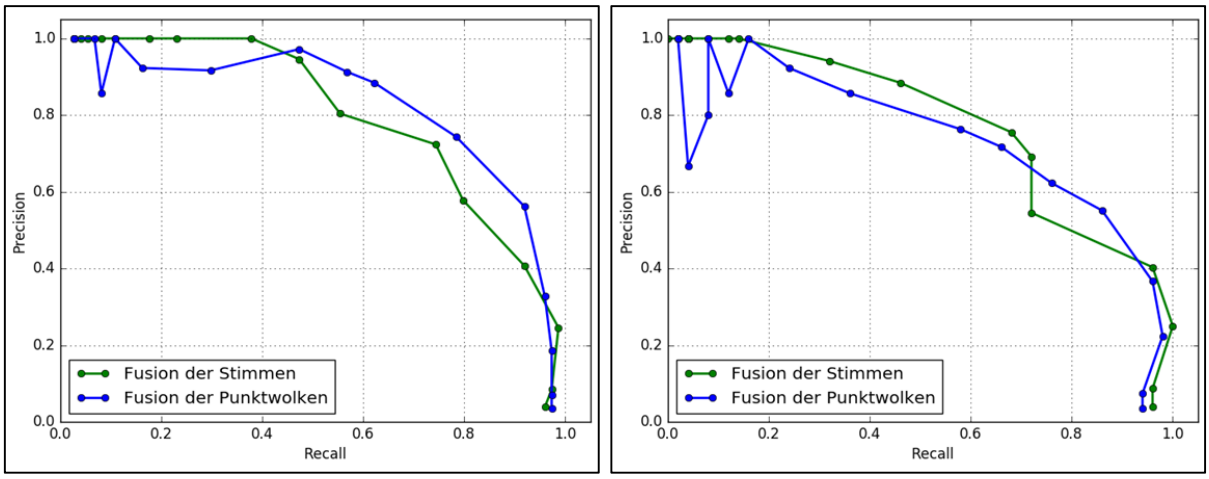
Für die Experimente wurde zunächst ein Wörterbuch trainiert. Für dieses Training wurden 993 per Hand annotierte Punktwolken verwendet. Die Annotationen umfassen dabei sowohl Positivbeispiele der Klasse „Person“ als auch Negativbeispiele, bei denen es sich um keine Personen handelt. Die für das Training verwendeten Daten stammen aus verschiedenen Messkampagnen bzw. gestellten Szenen.

Für die Experimente wurde zwei Sequenzen aufeinanderfolgender Punktwolken beider Sensoren verwendet. Die erste Sequenz umfasst 74 Punktwolken pro Sensor, insgesamt also 148. Die zweite Sequenz umfasst 50 Punktwolken pro Sensor. Beide Sequenzen wurden ebenfalls annotiert um eine Grundwahrheit zu generieren. In Abb. 5b) wird ein Beispiel dieser Daten gezeigt, wobei die einzelnen Sensoren farblich hervorgehoben sind. Die Sequenzen unterscheiden sich darin, wie sehr sich die aufgenommene Person im Überschneidungsbereich der beiden Sensoren bewegte. Bei Sequenz 1 gibt es nur wenige solcher Bewegungen, bei Sequenz 2 jedoch mehr. Es

kommt daher in dieser öfter zu den beschriebenen Verzerrungseffekten. Außerdem gibt es in der Sequenz 2 einen längeren Zeitraum, in dem die Person trotz gemeinsamer Verwendung zweier Sensoren nur teilweise zu sehen ist.
Um unser Verfahren der Datenfusion zwischen den Sensoren im Stimmraum mit einem einfachen Ansatz der Fusion der Punktwolken zu vergleichen, haben wir die Testdaten mit unserer Methode in beiden möglichen Varianten mit unterschiedlichen Detektionsschwellwerten verarbeitet. Es wurde ausgewertet, ob Detektionen in den verarbeiteten Daten mit Detektionen in der Grundwahrheit übereinstimmen. Dies wurde dann genutzt um Richtig-Positiv, Falsch-Positiv und Falsch-Negativ zu ermitteln. Diese Kennzahlen nutzen wir um *Precision* und *Recall* zu ermitteln und um damit die Leistung der Methode zu quantifizieren. Precision gibt den Anteil der korrekten Detektionen bezogen auf alle Detektionen an. Recall den Anteil der detektierten Personen bezogen auf die in der Grundwahrheit tatsächlich vorhandenen Personen.

4.2 Ergebnisse

In Abb. 6 werden die Ergebnisse unserer Experimente als Precision-Recall-Kurve dargestellt. a) stellt dabei die Ergebnisse für die Verarbeitung der ersten Testsequenz dar und b) für die zweite. Es wird deutlich, dass die Fusion im Stimmraum verglichen mit der Fusion der Punktwolken in der zweiten Sequenz bessere Leistungen liefert als in der ersten Sequenz. Dies ist mit dem größeren Anteil an Verzerrungseffekten zwischen den Sensoren in der zweiten Sequenz zu erklären. Treten solche Effekte kaum auf, profitiert die Fusion der Punktwolken davon, dass sie eine größere Datendichte für die Merkmalsextraktion verwenden kann. Umgekehrt hat sie jedoch mehr Fehler bei der Merkmalsextraktion, wenn solche Verzerrungseffekte häufiger vorkommen.

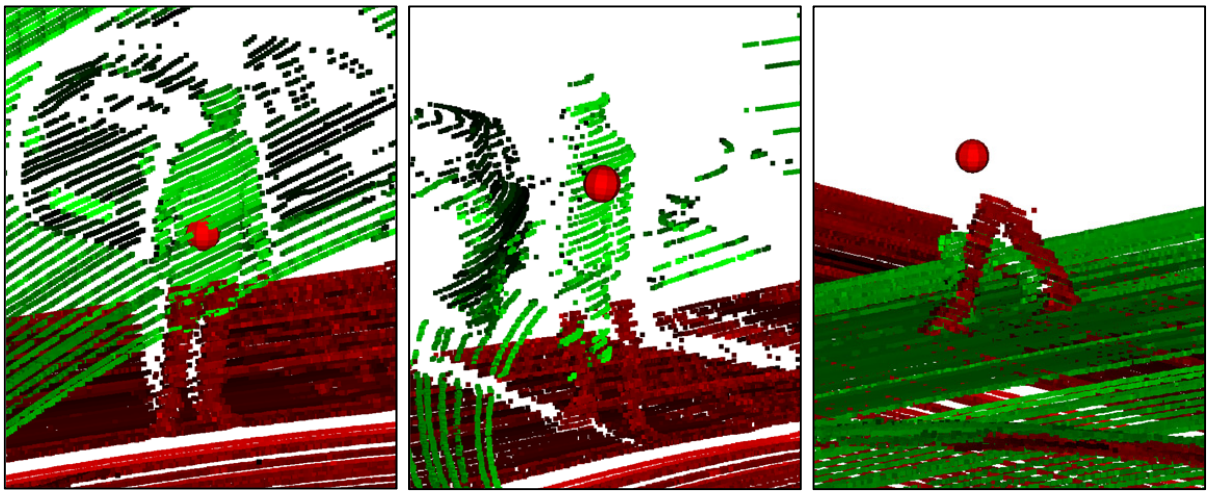


a) Precision-Recall Kurve für die erste Testsequenz b) Precision-Recall Kurve für die zweite Testsequenz

Abb. 6: Auswertung der Ergebnisse

Die Leistung ist insgesamt in der zweiten Sequenz geringer als in der ersten. Dies liegt am höheren Schwierigkeitsgrad der zweiten Sequenz. Sie hat nicht nur einen größeren Anteil Verzerrungseffekte, sondern auch Bereiche in denen die Person überhaupt nur teilweise zu sehen ist.

Abb. 7 zeigt Beispiele für die Ausgabe unserer Methode ohne Verzerrungen a) und mit Verzerrungen b). In Abb. 7c) ist zu sehen, wie unsere Methode auch noch eine korrekte Personenposition liefert, wenn der Mittelpunkt der Person gar nicht in den Daten enthalten ist.



a) Person ohne Verzerrungseffekte b) Person mit Verzerrungseffekten c) Person nur teilweise zu sehen

Abb. 7: Beispiele für die Ausgabe unserer Methode

5 Fazit & Ausblick

In diesem Beitrag haben wir unseren auf 3D-ISM basierten Ansatz auf eine Art und Weise erweitert, die ihn Robust gegenüber Verzerrungseffekte macht, wie sie auftreten, wenn bewegte Objekte von mehreren nicht komplett zeitlich synchronisierten LiDAR-Sensoren erfasst werden. Wir verwenden dafür eine Datenfusion im Stimmraum des ISM Verfahrens. Es wurde experimentell untersucht, dass diese Methode bei steigender Anzahl der beschriebenen Verzerrungseffekte gegenüber einer einfachen Fusion der Eingangsdaten bessere Ergebnisse liefert.
Die vorgestellte Methode ist für die Detektion von Personen bzw. mit entsprechendem Training auch anderer Objektklassen nicht auf eine vorherige Segmentierung der Daten angewiesen und unterscheidet sich darin von vielen anderen Methoden mit gleicher Zielsetzung. Sie ist dadurch robust gegen Segmentierungsfehler.
In der Zukunft planen wir den Ansatz der Datenfusion im Stimmraum weiter auszubauen und ihn auch für die Integration eines Trackingverfahrens und ggf. anderer Sensortypen zu nutzen. So kann die Ausgabe eines Trackingverfahrens ebenfalls als potenzielle Objekt Position mit einem gewissen Gewicht in den Stimmraum überführt werden. Die Ausgabe eines Kamerabasierten Detektors könnte als Strahl in diesen überführt werden.

6 Literaturverzeichnis

BEHLEY, J., STEINHAGE, V. & CREMERS, A.B., 2013: Laser-based segment classification using a mixture of bag-of-words. 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 4195-4200.

BORGMANN, B., HEBEL, M., ARENS, M. & STILLA, U., 2017: Detection of persons in MLS point clouds using implicit shape models. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W7, 203-210.

GARCIA-GARCIA, A., GOMEZ-DONOSO, F., GARCIA-RODRIGUEZ, J., ORTS-ESCOLANO, S., CAZORLA, M. & AZORIN-LOPEZ, J., 2016: PointNet: A 3D Convolutional Neural Network for real-time object class recognition. 2016 International Joint Conference on Neural Networks (IJCNN), 1578-1584.

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, Springer, 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.

MATURANA, D. & SCHERER, S., 2015: VoxNet: A 3D Convolutional Neural Network for real-time object recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 922-928.

MUJA, M. & LOWE, D. G., 2009: Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. International Conference on Computer Vision Theory and Application VISSAPP'09, INSTICC Press, 331-340.

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.

PREMEBIDA, C., CARREIRA, J., BATISTA, J. & NUNES, U., 2014: Pedestrian detection combining RGB and dense LIDAR data. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 4112-4117.

RUSU, R.B., BLODOW, N., & BEETZ, M., 2009: Fast Point Feature Histograms (FPFH) for 3D Registration. Proceedings of the 2009 IEEE International Conference on Robotics and Automation, IEEE Press, Piscataway, NJ, USA, 1848-1853.

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. CVPR 2011, IEEE.

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.

SOCHER, R., HUVAL, B., BATH, B., MANNING, C.D. & NG, A.Y., 2012: Convolutional-recursive deep learning for 3d object classification. Advances in Neural Information Processing Systems, 656-664.

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.