3D Personenerkennung und Verfolgung mit Stereo- und RGB-D Kameras

M. Adduci[1], K. Amplianitis[1], S. Bodas, M. Misgaiski-Haß, R. Reulke

Humboldt-Universität zu Berlin, Institut für Informatik, Computer Vision,
Unter den Linden 6, 10099 Berlin
eMail: reulke@informatik.hu-berlin.de
URL: http://www.informatik.hu-berlin.de/cv

Zusammenfassung. Die Erkennung und Verfolgung von Menschen mit Kamerasystemen ist ein sehr interessantes und sich schnell entwickelndes Forschungsgebiet und spielt gerade für die Sicherheitsforschung eine große Rolle. Bisherige Arbeiten konzentrieren sich auf 2D-Algorithmen, wobei die Erkennung, Extraktion und Verfolgung in 3D ein noch ziemlich unerforschtes Gebiet, vor allem in Bezug auf Multi-Kamera-Systeme ist.
Unser Ansatz konzentriert sich auf die Erkennung und Verfolgung von Personen im öffentlichen Nahverkehr aus den Daten mehrerer Stereo und RGB-D-Systeme (RGB-D bezeichnet die Kombination aus Grau-/Farb- und Distanzinformationen, wie z. B. bei der Microsoft Kinect).
Wesentliche Punkte des hier beschriebenen Ansatzes beziehen sich auf die Synchronisierung mehrerer Aufnahmesysteme und die Bestimmung ihrer Orientierungsparameter im Raum. Darüber hinaus wird mit Hilfe eines Bündelblockausgleichs geometrisch eine einheitliche 3D-Szene erzeugt, die dann einen Ausgangspunkt für die Erkennung und Verfolgung von Menschen im Beobachtungsraum bildet. Dazu werden signifikante Kennzahlen aus den erfassten Datensätzen ermittelt. In dem Beitrag wird eine Übersicht über die von mehreren RGB-D und Stereosensoren erzeugten Punktwolken und daraus abgeleiteten Daten erläutert und diskutiert.

1 Einleitung

Videoüberwachung ist ein wichtiger Teil bei der Gewährleistung der Sicherheit in öffentlichen Transportsystemen. Automatisierung mittels Bild-und Signalverarbeitungsmethoden hilft die Effizienz und Zuverlässigkeit solcher Systeme zu gewährleisten. Aktuelle Entwicklungen von Videoüberwachungssystemen erlauben die automatische Erkennung und Verfolgung von Menschen. Die geschickte Auswertung von Bewegungsmustern erlaubt eventuell Aussagen zur Sicherheit in öffentlichen Transportsystemen zu treffen. Diese Entwicklung ist Teil eines Projekts, das die Konzeption und Umsetzung eines Echtzeit-Tracking-System beinhaltet. Neben der Firma Interautomation, die das Projekt leitet, sind die Firmen Human Factor Consult, ASIS, die HU-Berlin und die TU-Berlin für die Entwicklung, Interpretation und Implementierung zuständig. Eine kurze Übersicht über das Projekt findet sich z.B. in [1].

1 *These authors contributed equally* to this *work*

Diese Arbeit bezieht sich auf die Beobachtung des Verhaltens von Menschen in Wagen des öffentlichen Nahverkehrs (U-Bahn oder S-Bahn). Ziel ist die Erkennung von atypischen oder Gefahrensituationen aus den abgeleiteten Trajektorien. Einen möglichen Ansatz zur Situationsbeschreibung findet man z.B. in [2].

Die Objekterkennung in Innenräumen von Zügen bietet eine Reihe von Herausforderungen:

- Eingeschränkter Beobachtungsbereich
- Verdeckungen durch Passagiere sowie Sitze und Haltestangen im Waggon
- Wenige Kameras und damit ein großer Öffnungswinkel einer Kamera
- Drastische Helligkeitsänderungen innerhalb (Fenster & abgeschattete Bereiche) und zwischen den aufgenommenen Bildern (Bahnhof, Tunnel & sonnige Bereiche)
- Bereiche in denen Bildverarbeitung Fehlerhaft arbeitet (Fenster, TV-Monitore)
- Geringe Deckenhöhe und somit geringer Abstand zu Personen

Daraus ergibt sich ein bestimmtes Herangehen bezüglich der Kameras und der Bildverarbeitung zur (verdeckungsfreien) Überwachung eines Fahrgastraums:

- Vollständige 3D Erfassung des beobachteten Raums durch Punktwolken (Vorteile liegen hier in der Hintergrundschätzung, dem Umgang mit Schatten und mit Verdeckungen)
- Verwendung von Multikamerasystemen zur Vermeidung von Verdeckungen. Ableitung der 3D-Informationen durch Stereokameras und RGB-D Systemen mit großen Öffnungswinkel, schnelles Matching und 3D Punktwolkenhandling
- Bestimmung des 3D Hintergrundes
- Detektion von separaten Objekten (die sich vom 3D-Hintergrund abheben)
- Mathematische Beschreibung der 3D-Objekte durch markante 3D Punkte und 3D Strukturen (z.B. Boxen und Ellipsoide)
- Ableitung von relevanten Attributen aus den einzelnen Kameraansichten
- Tracking und Fusion der unterschiedlichen Ansichten zu einer abstrakten Darstellung der Belegung im Fahrgastraum
- Aus der Analyse der Trajektorien sollen dann Situationen abgeleitet werden

Die Verwendung mehrerer Kameras erfordert die Synchronisierung der Aufnahmesysteme und die genaue Bestimmung der relativen Orientierung der Kameras zueinander.

Der Beitrag ist folgendermaßen aufgebaut: Nach einer kurzen Übersicht über bekannte und vergleichbare Ansätze, wird das Aufnahmesystem und der Arbeitsablauf beschrieben. Anschließend werden einige Beispiele und Situationen erläutert. Der Beitrag schließt mit einer Zusammenfassung und einem Ausblick.

2 3D Objekterkennung und Tracking

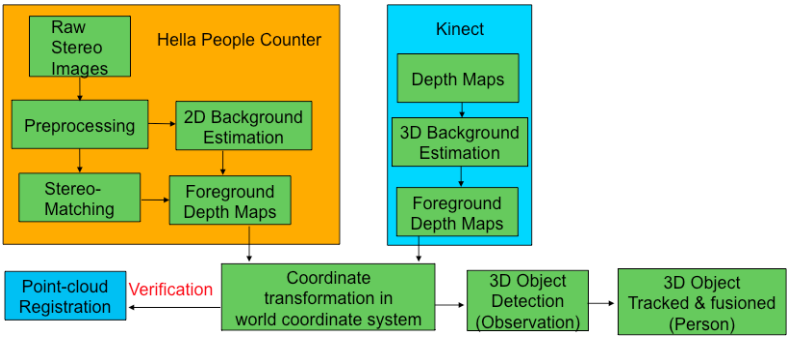

Abb. 1. Workflow (die Erläuterungen sind im Text zu finden).

Man findet in der Literatur eine Reihe von Arbeiten, die sich mit Teilaspekten befassen. In [3] findet man eine Beschreibung der Bestimmung einer Kopftrajektorie. Diese wird hier für die Sturzerkennung verwendet. Dazu wird die 3D Trajektorie mit einer Kamera im Innenraum abgeleitet. Der Kopf wird hier als 3D Ellipsoid modelliert. Die schnelle Höhenveränderung ist dann ein Zeichen für einen Sturz.
Für Posenerkennung von Menschen ist die Armstellung von entscheidender Bedeutung. In [4] wird ein Framework für die Armerkennung vorgestellt. Von besonderer Bedeutung sind dabei singuläre (nicht sichtbare) Bewegungen.
Einen Ansatz für die Detektion und Tracking von mehreren 3D Objekten in einer Szene findet man in [5]. Die verwendete Methode kombiniert Objekterkennung und –verfolgung. Ein einfaches frame-to-frame-Tracking ist zwar weniger rechenintensiv, versagt aber häufiger. Demgegenüber sind robuste Verfahren sicherer aber langsamer. In dem Beitrag wird die Kombination beider Ansätze beschrieben.

3 Hardwarebeschreibung und Softwareimplementierung

In dem folgenden Kapitel werden die Hardwarekomponenten und die implementierte Software beschrieben. Es werden mehrere RGB-D- und Stereosysteme verwendet. Kinect wurde von Microsoft zusammen mit der Firma PrimeSense entwickelt. Neben dem Preis ist die weite Verbreitung dieser Geräte interessant. Damit kann man auf einem gewaltigen Fundus von Software und Anwendungen zurückgreifen. Der Sensor kombiniert einen RGB und einen Tiefensensor. Der Tiefensensor bestimmt den Abstand anhand der lokalen Verschiebung bekannter Muster, die durch einen Szenengenerator im nahen Infrarot generiert werden (Triangulation). Parallel dazu wurde ein Stereokamerasystem der Firma Hella verwendet (People Counter). Insgesamt wurden 8 Stereokameras und vier Kinects verwendet. Alle Systeme wurden vor dem Einsatz kalibriert.

In der Abbildung 1 ist der Workflow dargestellt. Aus den Stereodaten wird mittels Stereomatching ein Disparitätenbild abgeleitet, das mit bestimmten geometrischen Informationen in eine Punktwolke umgebildet werden kann. Hier wird eine Implementierung des Semiglobal-Matchers (SGM) auf einer Grafikkarte verwendet (siehe [6]). Durch die effiziente Implementierung können mehrere Bildpaare pro Sekunde verarbeitet werden.
Nach der Ableitung einer Punktwolke von jedem einzelnen Sensor und der Berücksichtigung des Hintergrunds erfolgt die Objektdetektion. Ein Objekt zeichnet sich dadurch aus, dass es sich signifikant vom Hintergrund unterscheidet. Zur Bestimmung der Objekteigenschaften müssen aus den Punktwolken semantische Informationen abgeleitet werden. Es empfiehlt sich eine Anpassung von grob nach fein. Gewöhnlich beginnt man mit umschließenden Körpern (Bounding Boxen). Hier wird aus Teilen der Punktwolke ein Ellipsoid angepasst. Er ist Ausgangspunkt für eine feinere Beschreibung des Objekts. Erfüllen die geometrischen Eigenschaften des Ellipsoids bestimmte Randbedingungen, werden Attribute abgeleitet, die dann im Rahmen des Tracking weiter verarbeitet werden können. Zu den Parametern gehören:

- Hauptachsen (Länge und Position)
- Zentrum des Ellipsoiden
- Höchster Punkt
- Mittlere Helligkeit
- Anzahl der 3D Punkte
- Entfernung zur Kamera

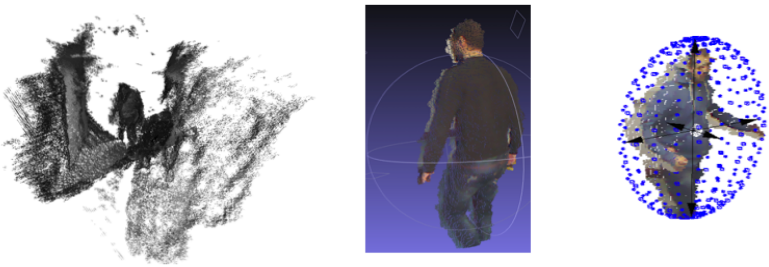

Abb. 2. 3D Verarbeitung (links: 3D Punktwolke, mitte: detektierte Person (Hintergrund ist abgezogen), rechts: Beobachtung mit Ellipsoid).

Im Rahmen des Trackings werden die Informationen von den unterschiedlichen Systemen fusioniert. Eine Person wird dann als fusionierter Ellipsoid aus den einzelnen Beobachtungen beschrieben.

In der Abb. 2 sind die Ergebnisse der einzelnen 3D Operationen dargestellt.

4 Ergebnisse

Die beschriebene Technik wurde labormäßig aufgebaut und konnte für

unterschiedliche Experimente genutzt werden. Gleichzeitig wurde das Equipment

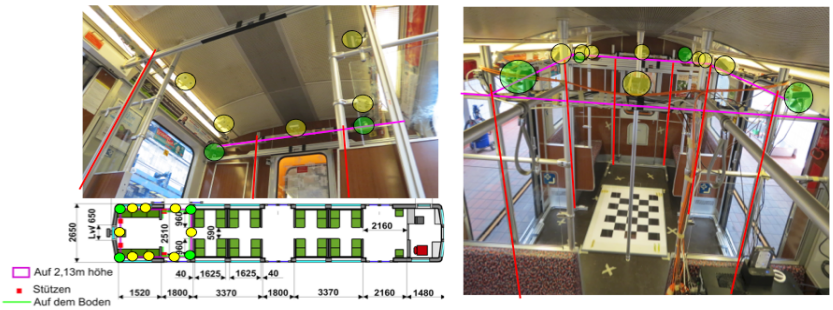

Abb. 3. Aufbauten für den Feldtest. Grün: Position der Kinect, gelb: Aufhängung für die Stereokameras. Unten links: Draufsicht auf den Waggon mit den eingezeichneten Aufbauten.

auch in einem Feldtest eingesetzt.
Da bis zu diesem Zeitpunkt die notwendige Zahl der Aufnahmesysteme nicht bekannt war, wurde mit der maximal möglichen Zahl von Aufnahmesystemen gearbeitet. Im Ergebnis der Feldtest konnte gezeigt werden, dass die Abbildungsbedingungen trotz geringer Deckenhöhe ausreichend sind, um Personen sicher zu detektieren. Das betrifft auch die Beleuchtungsbedingungen im Waggon. Allerdings konnten Verdeckungen durch Personen in einer Menschentraube nicht ausgeschlossen werden. Nachvollziehbar ist, dass die Erzeugung von Punktwolken an reflektierenden Objekten (Scheiben, metallische Haltestengen, usw.) fehlerhafte Ergebnisse liefern.

Außerdem wurde ein eigener Aufbau für Labortests erstellt. Damit konnten einzelne Objekte detektiert und verfolgt werden. Außerdem wurden damit menschliche Interaktionen untersucht. Für das Laborsystem wurde außerdem ein Referenzsystem (ARTTRACK2) verwendet.

5 Schlussfolgerungen und Ausblick

In dem Beitrag wurde ein System zur Erkennung und Verfolgung von Personen im Innenraumbereich vorgestellt. Die besonderen Herausforderungen ergeben sich aus Einbaubedingungen und Beobachtungsbereich, Verdeckungen, Lichtverhältnisse, etc.
Es konnte gezeigt werden, dass mit dem beschriebenen Ansatz den Anforderungen genügt werden kann. Allerdings müssen noch weitere Untersuchungen erfolgen, um die Form der Ellipsoide zu stabilisieren. Dies soll auch durch eine bessere Vorverarbeitung erreicht werden (Rauschen minimieren, usw.).
Für die Situationsanalyse muss die Datengrundlage deutlich vergrößert werden. Außerdem müssen Szeneninhalte auf die wesentlichen Personen und die

Interaktion zwischen ihnen reduziert werden.

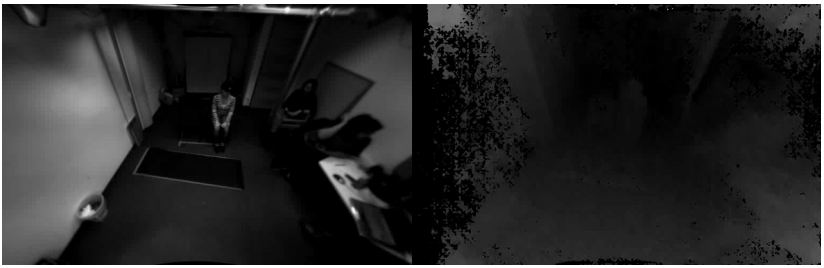

Abb. 4. Kameraansicht (links) mit Punktwolke (rechts).

Ein größerer Aufwand ist noch bei der Echtzeitfähigkeit und dem Handling vieler Personen bei der Situationserkennung notwendig.

Literatur

1. T. Jürgensohn, „Videokameras in Bus und U‑Bahn, Automatische Bildauswertung in der Entwicklung“, SIGNAL 5/2013, GVE-Verlag
2. R. Reulke, F. Meysel, and S. Bauer, "Situation Analysis and Atypical Event Detection with Multiple Cameras and Multi-Object Tracking," in *Robot Vision*. vol. 4931, G. Sommer and R. Klette, Eds., ed: Springer Berlin / Heidelberg, 2008, pp. 234-247.
3. C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "3D head tracking for fall detection using a single calibrated camera," *Image and Vision Computing,* vol. 31, pp. 246-254, Mar 2013.
4. F. Guo and G. Qian, "Singularity detection and consistent 3D arm tracking using monocular videos," *Image Analysis and Recognition,* vol. 3656, pp. 844-851, 2005.
5. Y. Park, V. Lepetit, and W. Woo, "Extended Keyframe Detection with Stable Tracking for Multiple 3D Object Tracking," *Ieee Transactions on Visualization and Computer Graphics,* vol. 17, pp. 1728-1735, Nov 2011.
6. I. Ernst and H. Hirschmüller, “Mutual Information based Semi-Global Stereo Matching on the GPU” In: Proceedings of the International Symposium on Visual Computing. ISVC08, Las Vegas, Nevada, USA.