M. Koch, A. Butz & J. Schlichter (Hrsg.): Mensch und Computer 2014 Workshopband,
München: Oldenbourg Wissenschaftsverlag, 2014, S. 211-219.

Erfassung der Oberkörperpose im Kraftfahrzeug

Matthias Ochs[1], Alexander Schick[1], Rainer Stiefelhagen[2]

Interaktive Analyse und Diagnose, Fraunhofer IOSB[1]
Institut für Anthropomatik und Robotik, Karlsruher Institut für Technologie[2]

Zusammenfassung

Die Bestimmung der Oberkörperpose des Fahrers in einem Kraftfahrzeug stellt einen wichtigen Teil neuer intelligenter Fahrerassistenzsysteme dar, um damit Gefühlslage, Verhalten und Intentionen des Fahrers analysieren zu können. In dieser Arbeit wird ein markerloses System vorgestellt, das basierend auf einem Stereokamerasystem die Oberkörperpose des Fahrers im 3D-Raum erfasst. Hierfür werden zuerst Verfahren zur Extraktion des Oberkörpers aufgezeigt, die sowohl auf 2D-Merkmalen als auch auf den Tiefendaten basieren. Anschließend wird in einem mehrstufigen Prozess auf diesem extrahierten Oberkörpercluster die Pose bestimmt. Mit vier Versuchsteilnehmern erfolgte die Evaluierung des Systems in einem realen Testszenario anhand von zehn Fahreraktivitäten und ca. 12000 Bildern.

1 Einleitung

Ein wichtiger Bestandteil für die Entwicklung neuer intelligenter Fahrerassistenzsysteme stellt die Beobachtung des Fahrers und der Autoinsassen mit Hilfe verschiedener Methoden des maschinellen Sehens im Automobil dar. Derartige Technologien ermöglichen es, sowohl das Fahrerlebnis durch innovative Bedienkonzepte als auch die Sicherheit der Autoinsassen zu steigern, indem Gefühlslage, Verhalten und Intentionen des Fahrers festgestellt werden.

Im Folgenden wird ein System vorgestellt, welches die Oberkörperpose des Fahrers im Auto mittels videobasierter Analyse erfasst. Dies beinhaltet sowohl die Bestimmung der Position des Torsos, des Nackens und des Kopfs als auch der Gelenkpositionen von Händen, Ellbogen und Schultern im 3D-Raum. Im ersten Schritt wird ein geeigneter Sensor, welcher in der schwierigen Autoumgebung operieren kann, und dessen Position ausgewählt. Beispielhaft sind an dieser Stelle die stark variierenden Lichtverhältnisse und die Selbstverdeckungen des Fahrers erwähnt. Anschließend wird eine robuste Segmentierung des Fahrers beschrieben und danach ein Verfahren zur markerlosen Erfassung der Gelenkpositionen vorgestellt, welches mit nur sehr wenigen vereinfachenden Heuristiken und direkt auf dem extrahierten Oberkörper arbeitet.

2 Verwandte Arbeiten

In diesem Kapitel werden zunächst verwandte Arbeiten, die im Auto und einer vom Auto losgelösten Umgebung operieren, vorgestellt und diskutiert. Bisher sind keinerlei Publikationen bekannt, die sich mit der Erfassung der Oberkörperpose im Auto auf Basis eines Stereokamerasystems auseinander setzen. Es existieren jedoch einige Ansätze, welche die Posenbestimmung mit unterschiedlichsten Sensoren, Algorithmen und Modellen beschreiben.

Im markenbasierten Ansatz von Ito und Kanade (Ito & Kanade 2008) wird eine Kamera an der Windschutzscheibe des Autos installiert, um am Fahrer befestigte Marken zu detektieren. Anhand der Position und Orientierung jeder Marke wird durch eine Diskriminanzanalyse die Pose klassifiziert. Ein anderes markerloses System stellt Tran und Trivedi (Tran & Trivedi 2012) basierend auf einem Multikamerasystem im Auto vor. Dabei wird die Oberkörperpose anhand von Hautfarbe und inverser Kinematik erkannt. Demirdjian und Varri (Demirdjian & Varri 2009) benutzen einen Time-of-Flight (ToF) Sensor, um aus der aufgenommenen 3D-Punktwolke die Pose mittels des Iterative Closest Point (ICP) Algorithmus zu bestimmen. Für diesen Algorithmus ist neben geeigneten Modellen für das Matching auch eine Initialisierung notwendig. Dieses Verfahren wird jedoch nur in einer autoähnlichen Laborumgebung eingesetzt.

All diese vorgestellten Publikationen haben gemeinsam, dass entweder stark vereinfachende Heuristiken getroffen werden oder dass sie nur in einer nachgebildeten Laborumgebung zum Einsatz kommen. Für den im Rahmen dieser Arbeit vorgestellten Ansatz werden sehr wenige derartige stark vereinfachende Heuristiken getroffen.

Ein weiterer wichtiger Aspekt ist die Sensorwahl. Nicht jeder Sensor kann im Auto effektiv eingesetzt werden. Beispielhaft angeführt seien hier die Schwierigkeiten bei der Positionierung eines ToF-Sensors im Fahrzeuginnenraum, damit dieser nicht durch die Sonneneinstrahlung gestört wird oder die Probleme eines Structured Light Sensors, wie der Microsoft Kinect, bei diffuser Sonneneinstrahlung. Somit kann der bekannte und auf der Kinect basierende Ansatz von Shotton et al. (Shotton et al. 2011) hier nicht eingesetzt werden, da außerdem noch der Klassifikator nicht für den Einsatz im Auto spezialisiert. Deshalb wird als Sensor ein Stereokamerasystem ausgewählt, da dieses System durch eine Belichtungsregelung oder eine zusätzliche Infrarotbeleuchtung einfach an die stark variierenden Lichtverhältnisse adaptiert und so im Auto positioniert werden kann, dass unerwünschte Selbstverdeckungen des Fahrers minimiert werden. In der Arbeit von Ziegler et al. (Ziegler et al. 2006) wird zwar ebenfalls ein Stereokamerasystem verwendet, um die Körperpose im Büro zu erfassen. Hierbei werden, wie bei Demirdjian und Varri auch, mittels eines ICP-Algorithmus die Gelenkpositionen bestimmt. Ein weiteres Verfahren, um die Körperpose anhand einer Stereokamera in einer Laborumgebung zu bestimmen, wird von Jojic et al. (Jojic et al. 1999) vorgestellt. Dort werden zunächst die Körperteile in 3D-Punktcluster aufgeteilt und danach wird auf die Pose mittels eines bayesschen Netzes geschlossen.

Im Rahmen dieser Arbeit erfolgt die Bestimmung der Oberkörperpose direkt auf der segmentierten 3D-Oberkörperpunktwolke. Diese Punktwolke kann als eine virtuelle Interaktionshülle, wie in der Arbeit von Bittel (Bittel 2013) beschrieben, aufgefasst werden. Anhand

dieser Hülle ist es möglich, Rückschlüsse auf die Lage der einzelnen Körperteile und Gelenke zu ziehen. Die Arbeit von Schick (Schick 2014) baut auf dem Ansatz der Pictorial Structures von Andriluka et al. (Andriluka et al. 2009) auf. Jedoch erfolgt die Bestimmung der Körperpose in 3D und anhand eines Supervoxelgraphens, der durch die Supervoxelsegmentierung gewonnen wird. Dieser beschriebene Supervoxelgraph und die Supervoxelsegmentierung werden in einer abgeänderten Form auch hier zur Bestimmung der Hand- und Ellbogenpositionen eingesetzt.

3 Methoden zur Erfassung der Oberkörperpose

Die Verfahren zur Erfassung der Oberkörperpose im Auto mit einem Stereokamerasystem werden in diesem Kapitel näher erläutert. Als Basis dienen die Tiefendaten eines Stereokamerasystems. Hierfür muss die Stereokamera sowohl intrinsisch als auch extrinsisch kalibriert sein. Danach kann anhand eines geeigneten Stereo-Matching-Algorithmus (Hirschmüller 2008) ein Disparitätsbild und somit die 3D-Daten berechnet werden. Diese Daten sind jedoch gegenüber anderen Sensoren extrem rauschanfällig. Daher werden zunächst Schritte zur Segmentierung und Extraktion des Fahrers aus der 3D-Rekonstruktionsszene vorgestellt. Daraufhin erfolgt die Erfassung der Oberkörperpose direkt auf diesem extrahierten Oberkörpercluster, wobei sehr wenig stark vereinfachende Heuristiken getroffen werden. Zuletzt werden die berechneten Gelenkpositionen mit Hilfe von Tracking-Verfahren zeitlich verfolgt.

3.1 Segmentierung und Extraktion

Neben dem Stereorauschen enthalten die Daten auch unerwünschte Objekte, wie zum Beispiel die B-Säule oder den Beifahrersitz. Da diese Objekte die spätere Bestimmung verfälschen würden, werden Methoden vorgestellt, welche sowohl auf den 2D-Merkmalen als auch auf der 3D-Rekonstruktion operieren, um den Oberkörper vollständig aus der Szene zu extrahieren.

(a) vor der Filterung *(b) nach 2D-Filterung* *(c) nach 3D-Filterung*

Abbildung 1: Beispiel einer 3D-Rekonstruktion nach Anwendung der Verfahren zur Segmentierung und Extraktion

Bei der 2D-Verarbeitung werden zunächst zur Rauschunterdrückung und zur Glättung von Lücken im Disparitätsbild, welche durch falsche bzw. fehlende Korrespondenzen verursacht

werden, ein Medianfilter und ein morphologischer Closing-Operator auf das Disparitätsbild angewandt. Im Anschluss wird der Floodfill Algorithmus (Heckbert 1990) dazu benutzt, kleine uninteressante Regionen aus dem Disparitätsbild zu entfernen, welche entweder durch Rauschen verursacht werden oder kleine, nicht relevante Objekte darstellen. Im nächsten Schritt wird der Oberkörper von den unerwünschten Objekten segmentiert. Diese Segmentierung erfolgt durch den GrabCut Algorithmus (Rother 2004), welcher sowohl auf dem Disparitätsbild als auch auf dem Aufnahmebild der Stereokamera operieren kann. Die Ausführung auf dem Disparitätsbild hat den Vorteil, dass sie invariant gegenüber Belichtung und dem Erscheinungsbild des Fahrers und des Innenraums ist. Jedoch bestehen Probleme bei Objekten, die in derselben Ebene liegen. In diesem Fall ist eine Segmentierung auf dem Aufnahmebild genauer, wenn sich der Fahrer gut gegenüber dem Fahrzeugsinnenraum abgrenzt.

Wie in Abbildung 1(b) zu erkennen ist, enthalten die Tiefendaten weiterhin Punkte, die vom Stereorauschen stammen und weit entfernt von der tatsächlichen Szene liegen. Daher wird der 3D-Raum auf den Fahrzeuginnenraum begrenzt. Das heißt, es werden nur diejenigen Punkte aufgenommen, die innerhalb eines Quaders liegen, der eine einfache Approximation des vorderen Fahrzeuginnenraums darstellt. Danach wird die Punktmenge der dicht besiedelten 3D-Szene mit Hilfe des Voxelgridfilters deutlich reduziert. Als letzter Schritt der Segmentierung erfolgt die Extraktion des Oberkörpersclusters mittels des euklidischen Clustering Algorithmus von Rusu (Rusu 2009). Dieser Algorithmus zerlegt die Punktwolke in einzelne zusammengehörende Cluster. Jedoch kann dieses Verfahren allein nicht gewährleisten, dass ein Cluster existiert, welcher den kompletten Oberkörper repräsentiert. Aus diesem Grund wird durch den Viola & Jones Facedetektor (Viola & Jones 2004) zunächst dasjenige Cluster gesucht, welches mindestens die Gesichtspunkte enthält. Basierend auf diesem Cluster werden dann alle weiteren personenbeschreibenden Cluster zu einem Oberkörpercluster O verschmolzen. Dies geschieht in einem iterativen Prozess, indem genau diejenigen Cluster C dem Oberkörpercluster O hinzugefügt werden, für die gilt $\| m_i - o \| \leq s$, wobei m_i und o die Schwerpunkte der Cluster C bzw. O sind und s ein Schwellwert, welcher bei jeder Iteration, bei der ein Zusammenschluss stattgefunden hat, angepasst wird. In Abbildung 1(c) ist das Ergebnis der Oberkörpercluster nach der Segmentierung und Extraktion in Grün dargestellt.

3.2 Erfassung der Oberkörperpose

In diesem Kapitel werden die einzelnen Methoden zur Positionsbestimmung des Torsos, Kopfs, Nackens, der Schultern, Ellbogen und Hände beschrieben, die direkt auf dem zuvor extrahierten Oberkörpercluster erfolgen. Jedoch kann im Allgemeinen nicht davon ausgegangen werden, dass die extrahierte Punktwolke keine Rauschanteile und keine störenden Objekte enthält. Daher muss durch die nachfolgenden Methoden sichergestellt werden, dass trotzdem eine korrekte Erfassung der Oberkörperpose realisiert wird.

Die Basis der Erfassung der Gelenkpositionen bildet eine abgeänderte Form der Interaktionshülle (Bittel 2013). Es wird angenommen, dass die extrahierte Punktwolke einer 3D-Normalverteilung $\mathcal{N}_3(\boldsymbol{\mu}, \Sigma)$ unterliegt. Somit beschreibt dann der Funktionswert der Dichtefunktion jedes Punkts die Lage der Körperteile adäquat. Dies bedeutet, dass Punkte, welche zum Torso gehören, einen höheren Funktionswert aufweisen, als Punkte, welche die Extremitäten

approximieren. In Abbildung 2(a) wird der Funktionswert jedes Punkts anhand einer Heatmap dargestellt. Die Torsoposition entspricht nun dem Schwerpunkt derjenigen Punkte, deren Funktionswert oberhalb eines definierten Schwellwerts liegt. Des Weiteren wird anhand dieser klassifizierten Punkte mittels Hauptkomponentenanalyse (PCA) ein Quader konstruiert, welcher den kompletten Torso approximiert.

(a) Heatmap und Torso *(b) Kopf* *(c) Supervoxel inkl. Pfad zur Hand*

Abbildung 2: Beispiele für die Verfahren zur Positionsbestimmung

Die Kopfposition wird mit Hilfe des Viola & Jones Facedetektor und des Min-Cut Algorithmus (Golovinskiy & Funkhouser 2009) bestimmt. Der Min-Cut Algorithmus wird zunächst dazu benutzt, die Kopfpunkte zu klassifizieren, somit den Schwerpunkt zu berechnen, welcher die gewünschte Kopfposition repräsentiert. Dabei separiert der Algorithmus die Punktwolke anhand eines konstruierten Graphs an genau derjenigen Stelle in zwei Teile, an welcher der größte Fluss herrscht. Dieser maximale Fluss sollte dabei immer auf der Höhe des Halses liegen. In Abbildung 2(b) ist die Teilung dargestellt. Des Weiteren wird durch diese klassifizierte Kopfpunktwolke der Kopfradius bestimmt, indem aus der Menge der euklidischen Distanzen das 0,9-Quantil zwischen diesen klassifizierten Punkten und der Kopfposition berechnet wird.

Die Gelenkpositionen der Schultern sowie die Lage des Nackens werden anhand geometrischer Beziehungen bestimmt. Aus den bekannten Kopf- und Torsopositionen wird die Longitudinalachse des Körpers ermittelt, auf welcher sich, ausgehend von der Kopfposition, mit dem Abstand des Kopfradius die Nackenposition befindet. Die Sagittalachse kann mit Hilfe des zuvor berechneten Quaders bzw. der PCA definiert werden. Auf Basis dieser beiden Achsen kann durch das Kreuzprodukt die Transversalachse berechnet werden, auf welcher die Schultergelenke liegen. Die Schulterposition wird ausgehend von der Nackenposition und der Distanz, welche durch die Seitenlänge des konstruierten Torsoquaders gegeben ist, ermittelt.

Zur Bestimmung der Hand- und Ellbogenpositionen wird die Punktwolke zunächst in Supervoxel (Schick 2014) aufgeteilt. So wird zum einem der mögliche Suchraum reduziert und zum anderen nimmt auch die Komplexität für die Suche der Handgelenke ab. Anschließend wird ausgehend von der bekannten Schulterposition iterativ mittels Minimierung einer Energiefunktion ein Pfad entlang des Arms zu den Handgelenken gesucht, dern Abbildung 2(c) ist ein solcher Pfad exemplarisch dargestellt. Dabei setzt sich die Energiefunktion folgendermaßen zusammen: $E(c, c_i) = \alpha \cdot winkel() + \beta \cdot distanz() + \gamma \cdot gauss() + \delta \cdot tracking()$, wobei die Parameter α, β, γ und δ Gewichtungsfaktoren entsprechen, c das besuchte

Supervoxel im aktuellen Iterationsschritt und c_i ein Supervoxel aus der Menge aller Supervoxeln ist. Der Term *winkel()* berechnet den Winkel zwischen c und c_i gegenüber einem iterativen Referenzvektor ausgehend von der Transversalachse. Der zweite Term *distanz()* besteht aus der inversen euklidischen Distanz zwischen c und c_i. Die Funktion *gauss()* entspricht der bekannten Dichtefunktion, welche zur Erkennung des Torsos eingeführt wurde. Im letzten Term *tracking()* wird ein Tracking-Verfahren berücksichtigt, das im nächsten Abschnitt näher beleuchtet wird. Nachdem in diesem iterativen Prozess ein Supervoxel gefunden wurde, welches das Handgelenk approximiert, kann nun mit Hilfe eines modifizierten A*-Algorithmus (Hart et al. 1968) und des Supervoxelgraphens (Schick 2014) ein Supervoxel für das Ellbogengelenk gefunden werden. Dieses muss notwendigerweise auf dem Pfad zwischen Hand und Schultern liegen.

3.3 Tracking

Um zusätzliche Informationen durch den zeitlichen Verlauf zu gewinnen und somit die Bestimmung der Gelenkpositionen zu verbessern, werden zwei Tracking-Verfahren eingesetzt. Zunächst wird eine Likelihood-Funktion, die auf einer Reward-Funktion basiert und für die zuvor definierte Energiefunktion benötigt wird, eingeführt, damit jedem einzelnen Supervoxel aufgrund der Historie eine Wahrscheinlichkeit zugeordnet werden kann, mit welcher es als Handgelenkposition klassifiziert wird. Diese Reward-Funktion ist ähnlich der eines Markov-Decision Process (Thrun 2005) aufgebaut und ist wie folgt definiert: $R_T = E[\sum_{\tau=1}^{T} \gamma^{\tau} f_{T-\tau}()]$. Der Erwartungswert E wird durch die Funktion $f_{T\text{-}\tau}()$ berechnet, welche die Wahrscheinlichkeiten der Handgelenkposition innerhalb der Zeitschritte τ und $T\text{-}\tau$ modelliert. Der Faktor γ^{τ} entspricht dem Diskontfaktor. Das zweite Tracking-Verfahren wendet auf jede erfasste Position das Kalman-Filter (Kalman 1960) an. Ziel einer derartigen Filterung ist es, aufgrund der zeitlichen Historie, mögliche fehlerbehaftete Erfassungen zu eliminieren bzw. zu korrigieren oder, sollte in einem Zeitschritt keine Bestimmung möglich sein, eine Position vorherzusagen.

4 Experimente und Ergebnisse

Das entwickelte System wurde in einem realen Testszenario in einem Auto evaluiert. Dazu wurde eine Stereokamera mit Weitwinkelobjektiven, welche nur eine geringe Verzeichnung aufweisen, an der Sonnenblende des Fahrers befestigt. Diese Sensorposition minimiert zum einen die Selbstverdeckungen des Fahrers und zum anderen wird ein großer Bereich des Fahrzeuginnenraums erfasst. Anschließend wurde von vier Versuchsteilnehmern ein Testszenario durchgeführt, das mit zehn Fahreraktivitäten eine reale Autofahrt simulieren soll. Hierbei wurden ca. 12000 Bilder aufgenommen, wovon für die Evaluation jedes 30. Bild manuell für die Ground Truth annotiert wurde.

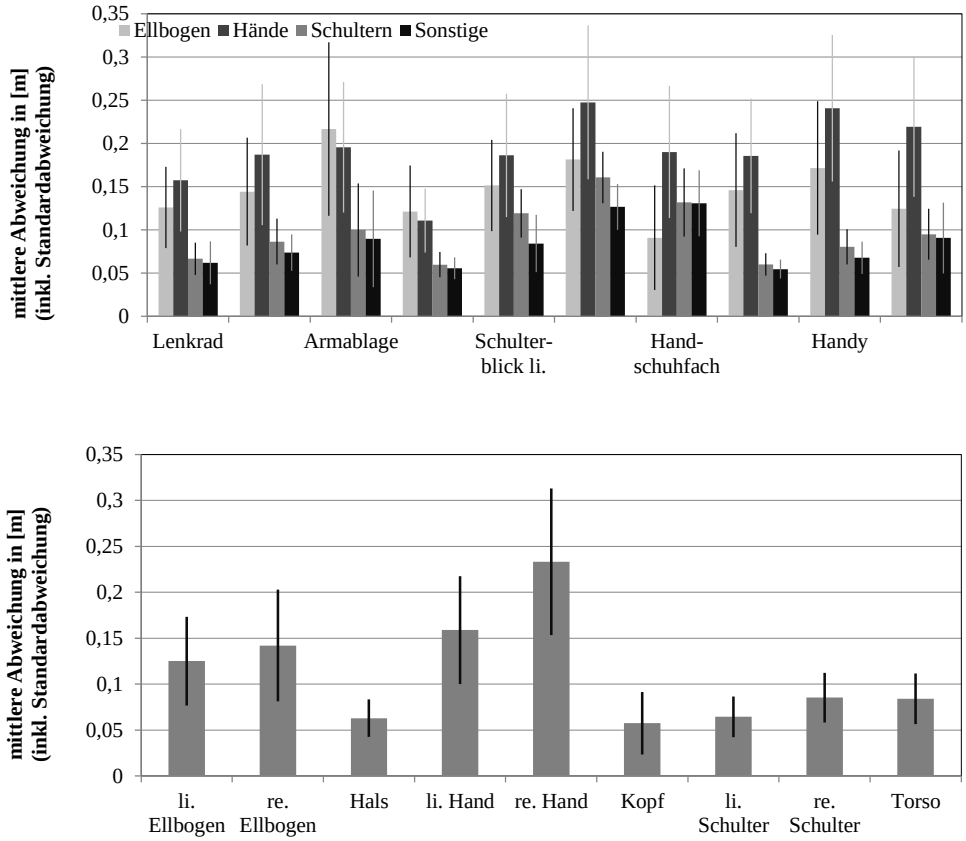

Abbildung 3: Mittlere Abweichung und Standardabweichung bzgl. Aktivitäten (a) und Gelenken (b)

Die in Abbildung 3 und in Tabelle 1 gezeigten Abweichungen entsprechen dem mittleren euklidischen Abstand zwischen der Soll- und Ist-Position in 3D über alle zur Auswertung verfügbaren Daten. Die Ergebnisse zeigen, dass das System insbesondere bei den Kopf-, Nacken-, Torso- und Schulterpositionen über alle Aktivitäten hinweg sehr gute Ergebnisse erzielt. Aber auch die sehr schwer zu erkennenden Hand- und Ellbogenpositionen erreichen bei vielen Aktivitäten sehr gute Resultate. Bei einigen wenigen schwierigen Aktivitäten, wie den Essen- und Handygesten, liegt die Abweichung noch unter 25 cm. Dennoch reicht diese erzielte Genauigkeit des Systems aus, dass es die Basis für eine weitergehende Aktivitäts-klassifikation bilden kann und mindestens vergleichbare Ergebnisse zu den verwandten Arbeiten erzielt. Das System ist derzeit noch nicht echtzeitfähig, jedoch kann dies durch weitergehende Optimierungen und GPU-Programmierung erreicht werden.

	Kopf	li. Hand	re. Hand	li. Ellbogen	re. Ellbogen
Mittelwert	5,7 cm	14,1 cm	19,6 cm	9,4 cm	14,1 cm
Standardabweichung	6,5 cm	11,5 cm	15,6 cm	12,5 cm	11,9 cm

Tabelle 1: Übersicht über quantitative Ergebnisse der Oberkörpererfassung über alle Aktivitäten

Abbildung 4: Erfassung der Oberkörperpose bei zehn Fahreraktivitäten. Die Punkte zeigen die erfassten Positionen, die zurück von der 3D-Position in das Aufnahmebild projiziert wurden.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein System zur Erfassung der Oberkörperpose im Kraftfahrzeug vorgestellt. Das Verfahren bestimmt hierbei die neun relevanten Gelenkpositionen des Oberkörpers eines Fahrers in 3D in einer realen Fahrzeugumgebung mit Hilfe eines Stereokamerasystems. Zuerst wurden Methoden zur Extraktion des Oberkörpers aus den sehr verrauschten Stereotiefendaten gezeigt und anschließend die Bestimmung der Gelenkpositionen, welche direkt aus dem extrahierten Oberkörper erfolgte, beschrieben. Die Evaluierung des Systems erfolgte in einem realen Testszenario mit verschiedenen Fahreraktivitäten. Die bisher erzielte Genauigkeit reicht völlig aus, um eine Klassifikation der Aktivitäten des Fahrers vorzunehmen. Jedoch besteht durchaus noch Potenzial, die erreichten Resultate durch wietere Optimierungen und Verbesserungen zu steigern. Beispielsweise könnte die Segmentierung durch die Verwendung eines detaillierten realen Modells des Fahrzeuginnenraums verbessert werden. Somit können anhand einer im nächsten Schritt folgender Klassifikation der Gesten und Aktionen des Fahrers zukünftige intelligente Fahrerassistenzsystem weiterentwickelt und verbessert werden.

Kontaktinformationen

Matthias Ochs, E-Mail: ochs.matthias@googlemail.com
Alexander Schick, E-Mail: alexander.schick@iosb.fraunhofer.de
Prof. Dr.-Ing. Rainer Stiefelhagen, E-Mail: rainer.stiefelhagen@kit.edu

Literaturverzeichnis

Andriluka, M., Roth, S. & Schiele, B. (2009). Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. *In: Proc. Conference on CVPR*, S. 1014-1021.

Bittel, S. (2013). *Virtuelle Interaktionshülle für druckbasierte Gestenerkennung*. Bachelorarbeit, KIT.

Demirdjian, D. & Varri, C. (2009). Driver pose estimation with 3D Time-of-Flight sensor. *In: Proc. CIVVS Workshops*, S. 16-22.

Golovinskiy, A. & Funkhouser, T. (2009). Min-Cut Based Segmentation of Point Clouds. *In: Proc. ICCV Workshops*, S. 39-46.

Hart, P. E., Nilsson, N. J. & Raphael, B. (1968): *A Formal Basis for the Heuristic Determination of Minimum Cost Paths*. IEEE Transactions on Systems, Science, and Cybernetics, 4(2), S. 100-107.

Heckbert, P. (1990). A Seed Fill Algorithm. In Glassner, A. S. (Hrsg.): *Graphics Gems*. Boston: Academic Press. S. 275-277 & 721-722.

Hirschmüller, H. (2008). *Stereo Processing by Semiglobal Matching and Mutual Information.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), S. 328-341.

Ito, T. & Kanade, T. (2008). Predicting driver operations inside vehicles. *In: Proc. Int. Conference on Automatic Face & Gesture Recognition*, S. 1-6.

Jojic, N., Turk, M. & Huang, T. S. (1999). Tracking Self-Occluding Articulated Objects in Dense Disparity Maps. *In: Proc. ICCV*, S. 123-130.

Kalman, R. E. (1960). *A New Approach to Linear Filtering and Prediction Problems*. Transactions of the ASME-Journal of Basic Engineering, 82(Series D), S. 35-45.

Rother, C., Kolmogorov, V. & Blake, A. (2004). *GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts*. ACM Transactions on Graphics, 23(3). S. 309-314.

Rusu, R. B. (2009): Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. Dissertation. TUM.

Schick, A. (2014). *Human Pose Estimation with Supervoxels.* Dissertation. KIT.

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 Single Depth Images. *In: Proc. Conference on CVPR*, S. 1297-1304.

Thrun S., Burgard, W. & Fox, D. (2005). *Probabilistic Robotics*. Cambridge, Mass.: MIT Press.

Tran, C. & Trivedi, M. M. (2012): 3-D Posture and Gesture Recognition for Interactivity in Smart Spaces. *IEEE Transactions on Industrial Informatics, 8(1)*, S. 178-187.

Viola, P. & Jones, M. J. (2004). *Robust real-time face detection*. Int. Journal of CV, 57(2). S. 137-154.

Ziegler, J., Nickel, K. & Stiefelhagen, R. (2006). Tracking of the Articulated Upper Body on Multi-View Stereo Image Sequences. *In: Proc. Conference on CVPR*, S. 774-781.