Monokulare 3D Positionsschätzung für Mensch-Roboter Interaktionen
Monocular 3D Position Estimation for Human-Robot Interactions

Heiko Renz, Technische Universität Dortmund, Lehrstuhl für Regelungssystemtechnik, 44227 Dortmund, Deutschland, heiko.renz@tu-dortmund.de

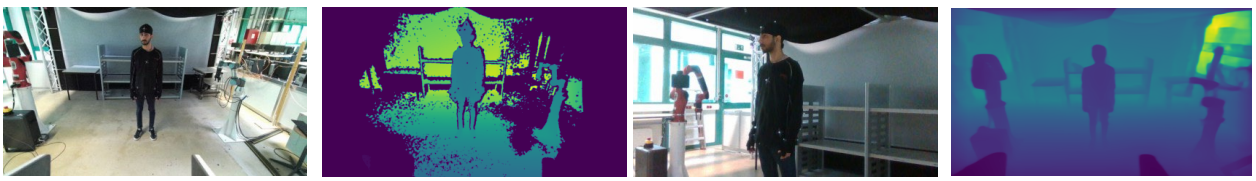
Patrick Palmer, Technische Universität Dortmund, Lehrstuhl für Regelungssystemtechnik, 44227 Dortmund, Deutschland, patrick.palmer@tu-dortmund.de

Aman Kungwani, Ehemaliger Student Technische Universität Dortmund, 44227 Dortmund, Deutschland, aman.kungwani@tu-dortmund.de

Univ.-Prof. Dr.-Ing. Prof. h.c. Dr. h.c. Torsten Bertram, Technische Universität Dortmund, Lehrstuhl für Regelungssystemtechnik, 44227 Dortmund, Deutschland, torsten.bertram@tu-dortmund.de

Kurzfassung

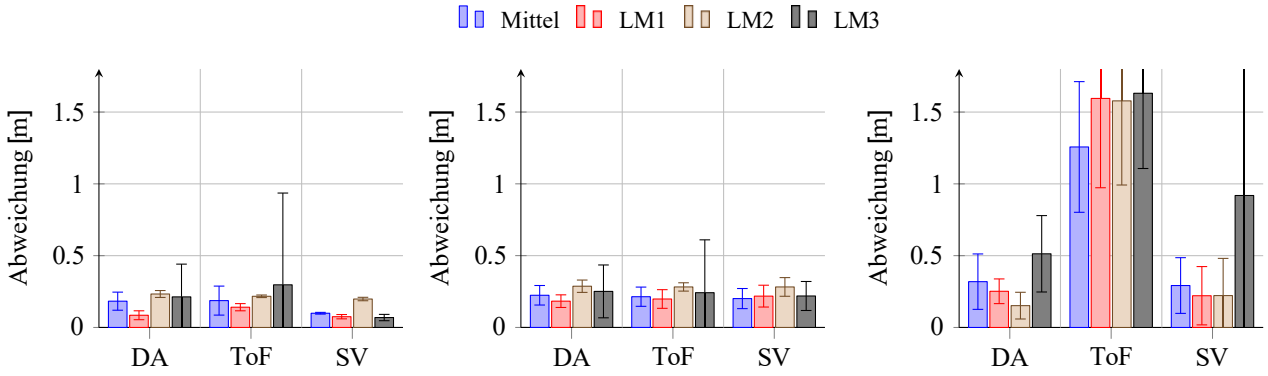
Die Interaktion zwischen Mensch und Roboter gewinnt in Bereichen der Industrie und im täglichen Leben zunehmend an Bedeutung. Eine notwendige Voraussetzung für eine sichere Zusammenarbeit ist dabei die hinreichend genaue Kenntnis der Position des Menschen im Arbeitsraum. Zur Lokalisierung des Menschen im Arbeitsraum gibt es verschiedene Sensorsysteme, wie markerbasierte Motion-Capture Systeme, welche aber oft schwer außerhalb von Testumgebungen einsetzbar und mit hohen Kosten verbunden sind. In dieser Arbeit wird ein zwei-stufiger Prozess zur Positionsschätzung auf einer vergleichsweise günstigen und einfach anwendbaren RGBD Kamera vorgestellt und hinsichtlich seiner Leistungsfähigkeit gegen andere Sensormodalitäten verglichen. Dafür wird die RGBD Kamera am Endeffektor eines Roboterarms montiert und verschiedene statische und dynamische Interaktionsszenarien mit einem Menschen aufgenommen. Zur Positionsschätzung wird ein neuronales Netz zur Tiefenschätzung auf dem RGB Bild, die Tiefendaten aus einem Time-of-Flight Sensor (ToF) und ein Stereo Vision Ansatz (SV) basierend auf zwei Perspektiven mit einem markerbasierten Ansatz gegenübergestellt. Bei der Planung von Robotertrajektorien in Interaktionsszenarien ist zur Kollisionsvermeidung die Kenntnis der aktuellen Position des Menschen notwendig. Verschiedene Ansätze zur Kollisionsvermeidung nutzen dafür markerbasierte Motion-Capture Systeme, da diese die benötigten Daten mit hoher Frequenz und Genauigkeit liefern [1, 2]. Ein alternativer Ansatz ohne Motion-Capture System ist die Positionsschätzung des Menschen unter Verwendung einer RGBD Kamera. Für eine vollständige Positionsschätzung mit einem zwei-Stufen Prinzip erfolgt erst eine 2D Bestimmung des Menschen im Bild und anschließend die Rekonstruktion der Tiefe. Zur Bestimmung der 2D Landmarken im Bild wird im Folgenden, ähnlich zur Literatur [3], ein vortrainiertes neuronales Netz verwendet [4]. Für die monokulare Tiefenschätzung wird anschließend das ebenfalls vortrainierte neuronale Netz DepthAnything (DA) verwendet [5]. Als Vergleichsansatz wird parallel zum RGB Bild das zugehörige Tiefenbild eines ToFs, sowie ein SV mit einer zweiten Perspektive angewendet. Die Genauigkeit der drei Ansätze wird basierend auf der Distanz zwischen den geschätzten Positionen verschiedener Mensch-Landmarker und den als Ground-Truth angenommenen Positionen aus einem markerbasierten Motion-Capture System verglichen. Als Metrik wird dabei der euklidische Fehler einzelner Landmarker, gemittelt über alle Frames eines Szenarios, genutzt. Zusätzlich wird zur weiteren Vergleichbarkeit der Ansätze über alle Landmarken gemittelt. Um auch die Genauigkeit der Ansätze in dynamischen Szenarien zu testen, werden unterschiedliche Szenarien aufgenommen; in denen der Mensch und oder der Roboterarm sich bewegen. Szenario 1 nutzt die Kameradaten des unbewegten Roboterarms bei Armbewegungen des Menschen, während Szenario 2 Daten des bewegten Roboterarms bei Bewegungen des Menschen nutzt. Als drittes Szenario werden zur Testung der Genauigkeit bei partieller Sichtbarkeit Daten aufgenommen, bei denen der Mensch den direkten Sichtbereich der Kamera verlässt. Abbildung 1 zeigt ein RGB Bild der Endeffektor Kamera (1a), das zugehörige Tiefenbild (1b), ein RGB Bild der zweiten Perspektive (1c) für SV, und eine Tiefenschätzung mit DA (1d).
Abbildung 2 zeigt die mittlere Abweichung zwischen markerbasiertem Motion-Capture System und den Ansätzen zur Positionsschätzung der drei Szenarien. Dargestellt ist der Mittelwert der Abweichung von den Ground-Truth Positionen gemittelt über die Länge des gesamten Szenarios für die Hüfte (LM1), die Schulter (LM2) und die Hand (LM3) des rechten Arms. Des Weiteren wird der über alle insgesamt neun Landmarken gemittelte Fehler (Mittel) für alle Ansätze



(a) RGB Bild Perspektive 1 **(b)** Tiefenbild Perspektive 1 **(c)** RGB Bild Perspektive 2 **(d)** DA Schätzung

Bild 1 RGB (1a) und Tiefenbild (1b) der ersten Perspektive, einer zweiten Perspektive (1c), und eine Tiefenschätzung mit DA (1d).

DOI 10.17185/duepublico/82918



(a) Szenario 1 **(b)** Szenario 2 **(c)** Szenario 3

Bild 2 Vergleich der Positionsschätzung mit verschiedenen Tiefenschätzungsansätzen gegen ein markerbasiertes Motion-Capture System. Alle Werte in Metern. Die Fehlerbalken zeigen die Standardabweichung des Mittelwertes des entsprechenden Wertes.

und Szenarien aufgeführt. Der gezeigte Fehler in Abbildung 2 ist dabei die Standardabweichung des Mittelwertes des entsprechenden Experiments. Zu beachten ist, dass bei fehlenden Tiefenwerten aufgrund des begrenzten Kamera-Sichtfeldes oder nicht erkannter Landmarker der letzte bekannte Tiefenwert angenommen wird. Basierend auf den Ergebnissen zeigt sich, dass SV in Szenarien, in welchen der Mensch vollständig erfasst wird, den geringsten mittleren Fehler aufweist. Grund hierfür ist die höhere Genauigkeit der Tiefenschätzung durch die Verwendung von zwei Kameras und die Reduktion der Genauigkeit durch die fehlende Sichtbarkeit von Landmarken in nur einer Perspektive bei Verlust der vollständigen Sicht. Der mittlere Fehler zwischen DA und ToF ist in den Szenarien 1 und 2 ähnlich, während in Szenario 3 der mittlere Fehler von ToF deutlich höher ist. Ein Grund hierfür ist, dass das Sichtfeld des Tiefenbildes geringer ist und mehrere Landmarken nicht sichtbar sind. Außerdem ist insbesondere in den Randbereichen des Sichtfeldes das Tiefenbild verrauscht. Die Tiefenschätzung mit DA wiederum nutzt das RGB Bild und hat ein größeres Sichtfeld, wodurch mehr Landmarken sichtbar sind und die mittlere Genauigkeit der Tiefenschätzung höher ist. Weitere Unterschiede zwischen den Ansätzen sind in den Standardabweichungen der mittleren Fehler erkennbar. Dabei ist zunächst zu betrachten, dass die Standardabweichung bei Körperteilen mit hoher Bewegung, wie der Hand (LM3), höher ist als bei Körperteilen mit geringer Bewegung, wie Hüfte oder Schulter. Ein weiteres ableitbares Merkmal ist, dass eine geringe Standardabweichung bei gleichbleibendem mittleren Fehler auf eine systematische Abweichung eines Ansatzes schließen lässt. Dies ist beispielsweise bei Szenario 1 und 2 für LM1 und LM2 bei allen Ansätzen zu erkennen. Diese systematische Abweichung entsteht durch die unterschiedlichen Aufnahmen und Modellierungen des Menschen zwischen den vorgestellten Ansätzen und dem Motion-Capture System. Das Motion-Capture System nutzt einen Lösungsansatz, der den entsprechenden Körperpunkt an die Stelle des Gelenkes im Körper legt. Kamerabasierte Ansätze können den Landmarker nur auf den äußerlich sichtbaren Körper abbilden, nicht aber in die tatsächliche Gelenkposition im Körper, da eine Rundumsicht fehlt. Ein weiterer Aspekt zur Bewertung der Eignung der Positionsschätzungen in der Mensch-Roboter Interaktion ist die Laufzeit, welche in allen Fällen über 100 ms liegt und eine Einschränkung für dynamische Interaktionen darstellt. Die verwendete Hardware ist mit einem Intel Core i5-3550 Prozessor, 32 GB RAM und einer NVIDIA Titan X Grafikkarte mit 12 GB VRAM ausgestattet.

Insgesamt zeigt sich, dass die monokulare Tiefenschätzung eine Alternative zu markerbasierten Motion-Capture Systemen darstellt, um die Position des Menschen im Arbeitsraum zu schätzen. Für enge und dynamische Interaktionen zwischen Menschen und Robotern sind die vorgestellten Ansätze ohne weitere Optimierung jedoch nicht geeignet, da die notwendige Genauigkeit von wenigen Zentimetern nicht dauerhaft oder echtzeitfähig erreicht wird.

Die Autoren bedanken sich für die finanzielle Unterstützung des Projektes durch die Deutsche Forschungsgemeinschaft (DFG, Projektnummer 497071854).

Literatur

[1] S. R. Schepp, J. Thumm, S. B. Liu, and M. Althoff, „SaRA: A Tool for Safe Human-Robot Coexistence and Collaboration through Reachability Analysis“, in *2022 Int. Conf. on Robotics and Automation*, ICRA ’22, 2022.

[2] H. Renz, M. Krämer, and T. Bertram, „Moving Horizon Planning for Human-Robot Interaction“, in *Proceedings of the 2024 ACM/IEEE Int. Conf. on Human-Robot Interaction*, HRI ’24, 2024.

[3] S. Mehraban, V. Adeli, and B. Taati, „MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network“, in *2024 IEEE/CVF Winter Conf. on Applications of Computer Vision*, WACV ’24, 2024.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, „Deep High-Resolution Representation Learning for Human Pose Estimation“, in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, CVPR ’19, 2019.

[5] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, „Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data“, in *2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, CVPR ’24, 2024.

DOI: 10.17185/duepublico/82918
URN: urn:nbn:de:hbz:465-20250219-112630-1