RESEARCH

Kristian Ehlers

Echtzeitfähige 3D Posenbestimmung des Menschen in der Robotik

Methoden und Anwendungen

Springer Vieweg

Echtzeitfähige 3D Posenbestimmung des Menschen in der Robotik

Kristian Ehlers

Echtzeitfähige 3D Posenbestimmung des Menschen in der Robotik

Methoden und Anwendungen

Springer Vieweg

Kristian Ehlers
Lübeck, Deutschland

Dissertation Universität zu Lübeck, 2018

ISBN 978-3-658-24821-5 ISBN 978-3-658-24822-2 (eBook)
https://doi.org/10.1007/978-3-658-24822-2

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Danksagung

Ich möchte mich bei jedem Einzelnen bedanken, der mir bei diesem Mammutprojekt auf seine Art und Weise zur Seite gestanden hat.

Mein Dank gilt natürlich meinem Doktorvater Prof. Dr.-Ing. Erik Maehle, der mir durch die Arbeit an seinem Institut die Möglichkeit gegeben hat, mich neben der Lehrtätigkeit wissenschaftlich zu finden und mir die forscherische Freiheit gelassen hat, mich im Gebiet der Posenbestimmung des Menschen in der Robotik zu entfalten.
Auch bei Prof. Dr.-Ing. Erhardt Barth möchte ich mich nicht nur für das Fungieren als Zweitgutachter bedanken, sondern viel mehr dafür, dass ich aufgrund seiner Vorlesung zu meinem Kernthema für die Masterarbeit bei ihm und letztlich für diese Dissertation gefunden habe.
Auch Prof. Dr. rer. nat. Thomas Martinetz gilt an dieser Stelle mein Dank, da er an der Entwicklung der Generalisierten Selbstorganisierten Karte grundlegend beteiligt war.
Bei Prof. Dr.-Ing. Mladen Berekovic möchte ich dafür bedanken, dass er mir nach der Übernahme des Instituts für Technische Informatik ausreichend Freiraum zum Beenden dieser Arbeit gelassen hat.

Einen großen Dank möchte ich an meine Kollegen und ehemaligen Kollegen des Instituts für Technische Informatik der Universität zu Lübeck richten, die stets für ein wohlfühlendes Arbeitsklima sorgten. Besonders hervorheben möchte ich Benjamin, Helge, Alex, Uli, Christopher und Cedric, die mir auch mit Rat und Tat zur Seite standen. Dankeschön.

Im Rahmen meiner bisherigen Tätigkeit am Institut für Technische Informatik hatte ich besonders viel Freude an der Lehre und der Arbeit mit den Studierenden. Ich möchte mich bei all denjenigen bedanken, die im Rahmen ihrer Bachelor- oder Masterarbeiten sowie Robotik-Praktika mit mir das Forschungsgebiet der Mensch-Roboter-Interaktion und insbesondere die Fragestellung der effizienten Posenbestimmung des Menschen und auf ihr basierenden Anwendungen untersucht haben. Besonders hervorheben möchte ich Buddy, Lasse und Thomas, die zusätzlich als Hiwis die ein oder andere fixe Idee meinerseits umsetzen mussten. Danke.

Nicht zu vergessen sind hier all meine Korrekturleser: Tanja, Cedric, Helga & Gerald und meine Mama. Besonders bedanken möchte ich mich bei Kathi & Helge und meiner Frau Christina, die dieses Werk mehrfach lesen durften.

Natürlich möchte ich meinen Eltern, Großeltern und meinem Bruder danken, die mich schon immer unterstützt haben und mir durch erholsame Wochenenden immer neue Energie gaben.

Mein größter Dank gilt jedoch meiner Frau Christina, die mir schon seit meiner Schulzeit jeden Tag zur Seite steht und es mit mir gerade in der Zeit der Doktorarbeit ausgehalten hat. Zusammen mit unseren Kindern Till und Nils hat sie mir immer aufs Neue Kraft gegeben und mich immer wieder neu motiviert. Ihr seid für mich das Wichtigste. Ich liebe Euch!

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Pseudocodeverzeichnis

Abstract

Gestures, as a natural way of communication, have been part of research areas such as human-computer interaction recently. Caused by the dissemination of depth cameras, they have also become increasingly popular in the field of robotics. Depth cameras belong to the standard sensors of humanoid robots such as "Pepper" and they are used for 3D human pose estimation to realize arbitrary applications. The interpretation of human poses should give the robots a kind of understanding of their environment and the behavior of human beings to achieve human-robot interactions.

Real-time human pose estimation provides the basis for the development of corresponding applications and, furthermore, for building robots to collaborate with humans. The algorithms as part of human-robot interaction interfaces need to be so efficient that they can run in parallel to the robot's control system without any kind of negative influence.

Part of this work is the development of three efficient, real-time capable approaches for hand pose estimation as well as human pose estimation based on the 3D point clouds corresponding to the hand or the human body. They are combined to one method uniting all advantages and allow for estimating the pose on standard hardware without GPU or even on low-power hardware such as the Raspberry Pi or an FPGA with frame rates of up to 30 fps. Furthermore, approaches for filtering the required hand respectively body point clouds out of the whole scene are presented.

The first pose estimation approach uses an unsupervised learning neuronal network given by a Self-Organizing Map (SOM). Therefore, specific topologies are designed and used for the estimation of the 3D positions of hand or body features. Furthermore, some kind of control and correction mechanism is developed to improve the results. The topology of the SOM is interpreted as a node-edge model and the distance between SOM and a data sample is given by the smallest Euclidean distance between the data point and the weights of all neurons, seen as the 3D position of the corresponding node.
Generalizing this SOM-data distance by allowing to include the edges given by the connections of the topology leads to a new type of SOM, i.e. the generalized SOM (gSOM). Hand-skeleton-like and human-body-skeleton-like topologies are designed and control and correction mechanisms are developed to allow the pose estimation to use the gSOM.
The third approach uses a self-scaling kinematic model fitted in the 3D point cloud of the hand or the human body by formulating a non-linear optimization problem solved by a Levenberg-Marquardt algorithm. The kinematic model allows for determining the angles

of the hand or body joints.
All three methods are combined making each other's estimated poses available as previous knowledge.

All methods are evaluated using public and private datasets which allows a comparison with the state of the art. Furthermore, the advantages of the approaches in terms of the human-robot interaction are presented.

Several applications are developed on the basis of the pose estimation approaches. There is an application which allows for controlling an industrial robot using arm movements. Furthermore, a human-robot interaction interface for not necessarily mobile robots is presented and provides the interaction with an autonomous moving mobile robot "PeopleBot" based on gesture detection. It is also possible to control a self-built robotic hand by hand and finger movements. A second human-robot interaction interface is designed for the humanoid robot "Pepper" and enables the imitation of arm movements as well as gesture control.

Since the developed approaches are not limited to hand or body pose estimation, they are used for the pose estimation of an industrial robot to enable first tests in terms of a purely visual collision detection.

Kurzfassung

Gesten als intuitive natürliche Form der Kommunikation sind seit längerem Forschungsgegenstand der Mensch-Computer-Interaktion und gewinnen durch die Verbreitung von Tiefenbildkameras im Bereich der Robotik an Bedeutung. Neueste humanoide Roboter zählen diese zu ihren Standardsensoren, die im Bereich der Mensch-Roboter-Interaktion unter anderem für die auf 3D-Daten basierte Posenbestimmung Verwendung finden, um den Robotern nicht nur ein Verständnis ihrer Umgebung, sondern auch für das Verhalten der sich darin befindlichen Menschen zu geben und ihnen die Interaktion miteinander zu ermöglichen. Die Grundvoraussetzung für die Entwicklung entsprechender Anwendungen mitunter im Rahmen der Mensch-Roboter-Kollaboration, in der Roboter unterstützend mit Menschen zusammenarbeiten, bilden effiziente, echtzeitfähige Posenbestimmungsverfahren, die in Form von Mensch-Roboter-Interaktionsschnittstellen parallel zu der Steuerungssoftware der Roboter einsetzbar sein müssen.

Im Rahmen dieser Arbeit erfolgt die Entwicklung von drei effizienten, echtzeitfähigen Methoden für die Posenbestimmung, die zu einem die Vorteile der einzelnen Ansätze vereinenden Gesamtverfahren kombiniert werden, welches die Ermittlung der Posen auf Standardhardware ohne Einsatz einer GPU sowie auf Hardware mit begrenzter Rechenleistung wie einem Raspberry Pi oder FPGA die Pose mit 30 fps ermöglicht. Die Grundlage der Bestimmung der Pose der Hand oder des Körpers bilden die korrespondierenden Punktwolken, für deren Filterung aus der mit Hilfe einer beliebigen Tiefenbildkamera aufgenommenen Gesamtszene verschiedene Ansätze präsentiert werden.

Bei der ersten Methode für die Posenbestimmung handelt es sich um ein unüberwacht lernendes künstliches Neuronales Netz, welches in Form einer Selbstorganisierenden Karte (englisch Self-Organizing Map (SOM)) mit einer entsprechenden hand- respektive körperähnlichen Topologie zusammen mit anwendungsspezifischen Kontroll- und Korrekturmechanismen die Positionsbestimmung definierter Hand- beziehungsweise Körpermerkmale ermöglicht.
Eine Generalisierung dieser Standard-SOM in Form der Erweiterung der Abstandsdefinition eines Datenpunktes zur der als dreidimensionales Knoten-Kanten-Modell interpretierten Topologie auf Basis des minimalen euklidischen Abstandes zwischen dem Datenpunkt und allen Knoten und Kanten führt zu der neuartigen Generalisierten SOM. Diese bildet in Verbindung mit anwendungsspezifischen Topologien und speziellen Kontroll- und Korrekturmechanismen einen weiteren Ansatz für die Posenbestimmung.

Die dritte Vorgehensweise beruht auf einem selbst-skalierenden kinematischen Modell der Hand respektive des Körpers, welches für die Formulierung der Posenbestimmung auf Basis der 3D-Positionen der Merkmale als nicht lineares Optimierungsproblem verwendet wird, dessen Lösung mit Hilfe des Levenberg-Marquardt-Algorithmus in den Winkeln und Positionen der Gelenke resultiert.
Die Kombination dieser drei Basisansätze erfolgt durch die gegenseitige Bereitstellung der ermittelten Merkmalspositionen, die als zusätzliche Informationen in den jeweiligen Posenbestimmungsprozess einfließen.

Die verschiedenen Methoden werden ausführlich unter Verwendung öffentlicher Datensätze evaluiert, mit dem Stand der Technik verglichen und die Vorteile gegenüber anderen Verfahren gerade im Hinblick auf die Anwendung im Bereich der Mensch-Roboter-Interaktion herausgestellt.

Auf Basis der entwickelten Ansätze werden verschiedenste Anwendungen im Bereich der Mensch-Roboter-Interaktion implementiert. Es erfolgen die Steuerungen eines Industrieroboterarms sowie einer selbst konstruierten und gefertigten roboterisierten Hand auf Basis von Arm- respektive Hand- und Fingerbewegungen. Es wird eine Mensch-Roboter-Interaktionsschnittstelle entwickelt, die die Interaktion mit dem sich in seiner Umgebung autonom bewegenden mobilen Roboter „PeopleBot“ über Gesten ermöglicht und auch für stationäre Roboter genutzt werden kann. Eine weitere Mensch-Roboter-Interaktionsschnittstelle wird direkt für den humanoiden Roboter „Pepper“ konzipiert und ermöglicht diesem das Nachahmen von Armbewegungen und die Interaktion über Gesten. Im Rahmen der Entwicklung eines ersten Ansatzes für die Posenbestimmung eines Industrieroboters für erste Untersuchungen im Bereich der rein visuellen Kollisionsvermeidung wird zudem gezeigt, dass die entwickelten Verfahren nicht auf den Einsatz für die Posenbestimmung des Menschen beschränkt sind.

1 Einleitung

> Er spricht mit Händen und Füßen.
>
> —*Redensart — Volksmund*

Diese Redensart mag der eine oder andere bereits über jemanden gesagt oder gar über sich selbst gehört haben. Gemeint ist das sich Ausdrücken unter Verwendung von Gesten, welches auf verschiedenste Weisen zu beobachten ist. Es gibt Menschen, die während des Sprechens beinahe automatisiert und unbewusst durchgehend ihre Arme bewegen, um dadurch ihren Aussagen mehr Nachdruck zu verleihen oder diese zu veranschaulichen. Entsprechende beabsichtigte Gestikulationen können beispielsweise häufig beobachtet werden, wenn man sich als Fremdsprachler in einer unbekannten Umgebung zurechtfinden muss. So wird gegebenenfalls auf Dinge gezeigt, deren Bezeichnungen gerade entfallen sind oder bei einer Wegbeschreibung in eine bestimmte Richtung gedeutet, statt diese rein verbal zu umschreiben.

Unabhängig davon, ob eine Absicht vorliegt oder nicht, behilft man sich bei der Gestikulation einer natürlichen und intuitiven Kommunikationsform, die in unserem alltäglichen Leben allgegenwärtig ist. Beispielsweise grüßen sich Menschen über große Entfernung indem sie winken, Polizisten regeln den Verkehr mit Hilfe von Armbewegungen, Taucher verständigen sich unter Wasser auf Basis von Handzeichen und sogar kleine Kinder deuten auf Dinge, die sie gern haben oder ihren Eltern zeigen möchten. Die Gebärdensprache ist eine komplette rein auf Arm- und Handbewegungen in Kombination mit Handzeichen basierende Sprache für Gehörlose und verdeutlicht, was für eine mächtige und vielseitige Kommunikationsform die Gesten sind.

Aus diesem Grund ist es nicht verwunderlich, dass diese Art der Kommunikation im Bereich der Mensch-Computer-Interaktion (MCI) bereits seit längerem Gegenstand der Forschung ist und mit der Markteinführung der Microsoft Kinect im Jahre 2010 in Verbindung mit der Xbox 360 Spielekonsole in Form einer kommerziell verfügbaren berührungslosen Navigation durch Menüs sowie der Steuerung von Spielen durch Bewegungen des gesamten Körpers der Allgemeinheit verfügbar gemacht wurde. Bei der Kinect handelt es sich um ein Multisensor-Gerät, welches neben Mikrophonen und einer herkömmlichen RGB-Kamera über eine Tiefenbildkamera verfügt. Sie stellt unter Aussendung

K. Ehlers, *Echtzeitfähige 3D Posenbestimmung des Menschen in der Robotik*,
https://doi.org/10.1007/978-3-658-24822-2_1

eines Lichtmusters im infraroten Farbspektrum die Distanzen zu den sich im Sichtfeld befindlichen Objekten in Form eines Tiefenbildes zur Verfügung. Diese Informationen bilden die Grundlage für die Bestimmung der Posen der sich vor der Kamera befindlichen Personen und folglich auch für die MCI mit Hilfe von Gesten.
Im Rahmen der MCI wird die Pose einer Person nicht nur als Position und Orientierung eines ihr zugeordneten Koordinatensystems bezüglich des Kamerakoordinatensystems aufgefasst sondern enthält zudem meist die Positions- sowie gegebenenfalls Orientierungsinformationen einzelner Körpermerkmale wie Hände oder Füße und eventuell sogar die Stellungen beziehungsweise Winkelwerte der einzelnen Körpergelenke.
Es stellte sich schnell heraus, dass der Einsatz der Kinect keineswegs auf den Bereich der Spielekonsolen beschränkt ist und sich dieser Sensor letztlich als kostengünstiger Ersatz für teurere Stereokameras oder Time-of-Flight Kameras im Bereich der Bildverarbeitung weit verbreitet einsetzen lässt [1]. Sie findet unter anderem im Bereich der Objekt-Detektion und -Verfolgung sowie der Objekt-Erkennung und Szenen-Erkennung Verwendung, um bestimmte Objekte im dreidimensionalen Raum zu finden und zu verfolgen oder lediglich zu entscheiden, ob definierte Objekte sich im Raum befinden. Das Problem der Trennung von Objekten und Hintergrund lässt sich in diesem Kontext beispielsweise auf 3D-Informationen zurückführen, statt bisher auf reinen RGB-Bildern zu basieren.
Weitere Anwendungsbereiche für Tiefenbildinformationen im Allgemeinen sind die Analyse von menschlichen Aktivitäten im Bereich der Industrie im Rahmen der Überwachung von Arbeitsräumen von größeren Industrierobotern oder in der Medizin für die Überwachung von Patienten auf Intensivstationen. Die Gestenerkennung basierend auf Tiefenbildern findet für berührungslose Interaktionen mit Geräten zum Beispiel in sterilen Umgebungen wie Operationssälen oder für Industrieanlagen Verwendung. Verlässt man den Menschen als zentralen Gegenstand der Bildverarbeitung, werden 3D-Informationen unter anderem für das Erstellen digitaler Abbilder realer Objekte oder Umgebungen eingesetzt.
Weiterhin sind RGB-D Sensoren wie die Kinect heutzutage in Forschungsgebieten der Robotik als Ersatz teurer Laserscanner verbreitet und bilden zudem die Grundlage für neue Ansätze im Bereich der Simultanen Selbstlokalisation und Kartenerstellung (englisch Simultaneous Localization and Mapping (SLAM)) sowie der Navigation für die mobile Robotik wie beispielsweise graphenbasierte visuelle SLAM Algorithmen [2]. Auch neueste humanoide Roboter, wie der „Pepper“[1] der Firma SoftBank Robotics, zählen Tiefenbildkameras zu ihren standardmäßigen Sensoren. Diese Roboter dringen immer mehr in unser tägliches Leben vor, werden unter anderem als Führer in Museen oder auf Messen eingesetzt und bieten die Möglichkeit einfacher Dialoge und Interaktionen mit dem Menschen. Es existieren sogar vollständig von „Pepper“-Robotern geführte Geschäfte für Mobiltelefone[2]. Während diese Verhalten mehr oder weniger starr einprogrammiert

[1] https://www.ald.softbankrobotics.com/en/robots/pepper, Januar 2018
[2] https://blogs.wsj.com/japanrealtime/2016/01/28/softbank-to-staff-mobile-phone-storewith-pepper-robots, Januar 2018

sind und nur teilweise autonome Reaktionen zulassen, gibt es Anwendungen, in denen autonome Roboter sich nicht nur in ihrer Umgebung lokalisieren und navigieren, sondern komplexere Aufgaben übernehmen müssen und somit ein Verständnis für die Umgebung erforderlich machen. Beispielsweise könnten humanoide Roboter im Bereich der Altenpflege unterstützend eingesetzt werden. Das Forschungsgebiet der Mensch-Roboter Kollaboration beschäftigt sich mit der den Menschen unterstützenden Koexistenz von Mensch und mobilen Roboterplattformen im industriellen Arbeitsumfeld, in dem diese Plattformen beispielsweise den Transport und das Halten schwerer Baukomponenten vornehmen. Zum Bewerkstelligen entsprechender Aufgaben gehört nicht nur Navigation in einer eventuell bekannten Umgebung und die Vermeidung von Kollisionen mit statischen Hindernissen, sondern vielmehr das Beachten von sicherheitskritischen Aspekten. So dürfen die sich in der Umgebung befindenden Personen nicht verletzt werden oder es soll sogar die Möglichkeit der Interaktion mit dem Roboter gegeben sein. Grundlegend zu lösende Probleme für die Realisierung entsprechender Verhalten sind die Detektion von Personen und die Analyse derer Handlungen auf Basis ihrer Posen bis hin zur Deutung ihrer Absichten.

Als ein konkretisiertes Beispiel für die Posenbestimmung und deren Einsatz im Rahmen der MRI sei das in Abbildung 1.1 visualisierte Szenario definiert. Ein autonom in seiner Umgebung agierender Gabelstapler erhält von der sich vor ihm befindlichen Person den Befehl, die Palette aufzunehmen, auf die durch beide Arme gedeutet wird. Die Realisierung dieser Anwendung basiert auf der Posenbestimmung der Person anhand der mit der Tiefenbildkamera aufgezeichneten Bilder. Es erfolgt die Berechnung einer 3D-Punktwolke der Szenerie und die Filterung bezüglich der zur Person korrespondierenden Daten. Diese dient als Grundlage für die Bestimmung der Pose, die im Beispiel als Knoten-Kanten-Modell innerhalb der Daten dargestellt ist und sowohl die Positions- und Orientierungsinformationen der Person und der einzelnen Merkmale als auch die Winkelstellungen der Gelenke repräsentiert. Diese Pose wird bezüglich der definierten Körperhaltung beider Arme analysiert und löst gegebenenfalls das entsprechende Verhalten des Gabelstaplers aus. Zudem erfolgt auf Basis des Schnittpunktes der Ausrichtung der Arme die Bestimmung der aufzunehmenden Palette.

Ein anderer Einsatzbereich der Posenbestimmung ist die Virtuelle Realität, bei der Kameras im Kopfbereich montiert und für die Bestimmung der Handpose meist in Kombination mit einer VR-Brille wie der Oculus Rift[3] verwendet werden[4].

Der Großteil der genannten Anwendungsszenarien für stationäre Tiefenbildkameras oder in Kombination mit mobilen Robotern sind aktueller Forschungsbestandteil und erfordern möglichst effiziente, echtzeitfähige Verfahren für die Bestimmung und Analyse der Pose des Menschen als Ganzes oder je nach Anwendung der Pose der Hand. Entsprechende Methoden müssen parallel zu bestehender, für den Betrieb des jeweiligen Systems notwen-

[3] https://www.oculus.com/rift/#oui-csl-rift-games=mages-tale, Januar 2018
[4] https://developer.leapmotion.com/orion,http://nimblevr.com, Januar 2018

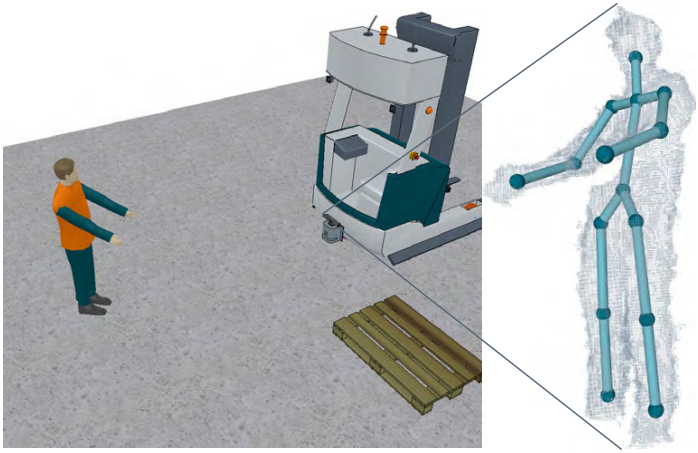

Abbildung 1.1: Beispielszenario für den Einsatz der Posenbestimmung im Bereich der MRI für die Steuerung eines sich autonom in seiner Umgebung bewegenden Gabelstaplers. Das auf die Palette Deuten der Person mit beiden Armen wird vom Roboter als Befehl für die Aufnahme dieser Palette interpretiert.

diger Software wie zum Beispiel der Navigation und Steuerung eines mobilen Roboters verwendbar sein, ohne diese negativ zu beeinflussen oder gar die kompletten Ressourcen des Systems aufzubrauchen.

Diese Arbeit ist in den entsprechenden aktuellen Forschungsgebieten der 3D-Bildverarbeitung, MCI und vor allem der MRI anzusiedeln, denn als Hauptgegenstand sind die Entwicklung von Verfahren für die Bestimmung der dreidimensionalen Pose des Menschen im Raum auf Basis von Tiefenbildern für die hauptsächliche Verwendung im Bereich der mobilen Robotik sowie die Implementierung entsprechender MRI Anwendungen zu definieren. Es werden je drei entwickelte Vorgehensweisen für die Bestimmung der Pose der Hand und die des gesamten menschlichen Körpers präsentiert und zu einem Gesamtverfahren kombiniert. In einem ersten, aus dem Gebiet der künstlichen neuronalen Netze stammenden Ansatz erfolgt die Verwendung einer Standard-Selbstorganisierenden Karte (englisch Standard Self-Organizing Map (sSOM)) mit einer hand- beziehungsweise körperähnlichen Topologie für die Bestimmung von Posen in Form der Positionen definierter Hand- beziehungsweise Körpermerkmale. In einem zweiten Vorgehen findet die hergeleitete Generalisierte Selbstorganisierende Karte (englisch Generalized Self-Organizing Map (gSOM)) als neuartige Selbstorganisierende Karte (englisch Self-Organizing Map (SOM)) Verwendung, bei der die Abstandsdefinition zwischen der Topologie in Form der Interpretation der Gewichtungsfaktoren der Neurone und deren Verbindungen als Positionen

und Strukturen im dreidimensionalen Raum generalisiert wird. Auch dieser Ansatz ermöglicht die Positionsbestimmung spezifischer Merkmale. Das dritte Vorgehen basiert auf einem selbst-skalierenden kinematischen Modell und ermöglicht die Bestimmung der Hand- und Körperposen in Form der Positionen der Merkmale im Raum sowie als vollständige kinematische Beschreibung, welche als die zu den Gelenken des kinematischen Modells korrespondierenden Gelenkstellungen gegeben ist. Diese drei Ansätze werden je Anwendungsbereich zu einem effizienten Gesamtverfahren kombiniert, welches die Posenbestimmung in Echtzeit ermöglicht, die im Rahmen dieser Arbeit in Anlehnung an die Nutzbarkeit für reale Anwendungen und nach Oikonomidis et al. [3] als mindestens 15 fps respektive einer verzögerungs- und verlustfreien Verarbeitung der Tiefenbilder und folglich der Bereitstellung der Posen mit der Bildwiederholfrequenz der Kamera von meist 30 fps definiert ist. Als Grundlage dienen stets die mit einer Tiefenbildkamera aufgenommen Informationen einer Szene. Lösungsansätze für grundlegende Probleme wie die Detektion der Hand oder die Bestimmung der zu dem betrachteten Objekt korrespondierende Punktwolke werden ebenfalls dargelegt. Des weiteren erfolgt die Implementierung verschiedenster Anwendungen im Bereich der MRI wie der Gestenerkennung als Standardanwendung, einer Schnittstelle zur Steuerung eines Industrieroboters auf Basis von Armbewegungen sowie zweier Mensch-Roboter-Interaktionsschnittstellen (MRIS) für die Interaktion mit nicht notwendigerweise mobilen sich autonom in einer Umgebung bewegenden Robotern wie dem „PeopleBot“ oder dem humanoiden Roboter „Pepper“. Ferner wird eine Möglichkeit zur Schätzung der Pose eines Industrieroboters auf Basis der entwickelten Posenbestimmungsverfahren präsentiert, die unter anderem für eine rein visuelle Kollisionserkennung genutzt werden kann. Die korrekte Funktionsweise der einzelnen Verfahren und Anwendungen wird mit Hilfe entsprechender Evaluationen untermauert.

Nachfolgend wird der Stand der Technik bezüglich der Posenbestimmung präsentiert und der wissenschaftliche Beitrag der entwickelten Ansätze unter dessen Berücksichtigung dargestellt. Im Anschluss erfolgt die Darlegung der Struktur der Arbeit unter Kennzeichnung der eigenen Beiträge des Autors.

1.1 Stand der Technik

Dieser Abschnitt präsentiert den aktuellen Stand der Technik im Bereich der Verfahren für die Posenbestimmung des menschlichen Körpers und der Hand sowie der Gestenerkennung als zentrale Bestandteile der Bildverarbeitung, der MCI und spezieller der MRI als sehr aktuelle Forschungsgebiete. Da das Themengebiet der Posenbestimmung sehr weitreichend ist, erfolgt die Darstellung wichtiger Ansätze in Form eines Überblicks, der in die Posenbestimmung des Körpers und der Hand unterteilt wird.

Wang und Popović [4] nutzen für die Bestimmung der Handpose einen Handschuh mit einem speziellen Farbmuster und einem zuvor aufgenommenen Datensatz von gerasterten Bildern mit dem Farbhandschuh in verschiedenen natürlichen Handposen. Das Problem der Posenbestimmung entspricht folglich der Suche nach dem passendsten Bild bezüglich einer Metrik. Schröder et al. [5] erweitern dieses Verfahren für die Steuerung einer Roboterhand, indem sie ein kinematisches Handmodell zum Erstellen eines synthetischen Datensatzes von Handschuhbildern verschiedenster Posen nutzen. Weitere Ansätze basieren auf an die Hand angebrachten Markern [6, 7]. Entsprechende Ansätze sind als klassisch zu bezeichnen, da diese zusätzliche Materialien wie Handschuhe und Marker benötigen, die von den Personen korrekt getragen und angebracht werden müssen. Diese Notwendigkeiten sind für reale Anwendungen nicht praktikabel und bilden meist eine Hürde für viele Nutzer. Aus diesem Grund orientiert sich die Forschung mehr auf rein visuelle Verfahren.

Gemäß Taylor et al. [8] lassen sich die Methoden für die Bestimmung der Pose der Hand in drei Klassen unterteilen, die den folgenden Definitionen genügen. Es gibt „*discriminative*“ Ansätze, deren Grundgedanke die direkte Ermittlung der Handpose auf Basis von Merkmalsextraktion unter Verwendung von Klassifikations- oder Regressionsverfahren ist [9, 10] und die nachfolgend als merkmalbasierte Absätze bezeichnet werden. Diese sind somit vom zeitlichen Ablauf der Handbewegungen unabhängig und bestimmen die Pose der Hand für jedes Bild separat. Sie basieren häufig auf zuvor berechneten Datenmengen und sind somit nicht präzise auf jede Handproportion abgestimmt. Im Gegensatz dazu existieren die *generativen* oder auch modellbasierten Verfahren, deren Posenbestimmung auf dem sequentiellen zeitlichen Verlauf von Handbewegungen und der direkten Verwendung und Anpassung eines Handmodells basiert [3, 11]. Die Genauigkeit dieser Methoden hängt meist von der korrekten Initialisierung des Modells beziehungsweise der korrekt bestimmten Pose im vorherigen Bild ab. *Hybride* Verfahren nutzen häufig einen merkmalbasierten Ansatz für eine eventuelle Initialisierung und modell-basierte Ansätze für die Einbeziehung zeitlicher Abläufe [12–15].

Nachfolgend werden einige merkmalbasierte Ansätze präsentiert. Ren et al. bestimmen die Pose der Hand in Form von Fingergesten mit Hilfe der Definition einer Finger-Earth-Mover-Distanz als Metrik für den Unterschied von Handposen [16, 17]. Die Finger werden auf Basis der durch das Tiefenbild einer Kinect gegebenen Handform bestimmt und die Gestenerkennung mit Hilfe von Template-Matching auf einem zuvor aufgezeichneten Datensatz realisiert. Athitsos und Sclaroff [18] formulieren ein Indexing-Problem auf einer Bilddatenbank, um plausible 3D-Handkonfigurationen zu ermitteln. Zu diesem Zweck berechnen sie einen großen Datensatz synthetischer Handbilder mit Hilfe eines gelenkigen Handmodells für verschiedenste Posen. Auch in diesem Fall beruht die Posenbestimmung auf dem Finden des ähnlichsten Bildes unter Verwendung von Kanteninformationen und des Chamfer-Abstandes. Der Ansatz von Keskin et al. [19] erweitert das ursprünglich für die Posenbestimmung des Körpers genutzte Verfahren von Shotton et al. [20] und

verwendet zuvor auf Handdaten trainierte randomisierte Entscheidungswälder (englisch random decision forests (RDFs)) für die Bestimmung von Handposen. In [9] erweitern sie ihr Verfahren zu mehrschichtigen RDFs.

Die generativen Ansätze sind ebenfalls weit verbreitet. Horaud et al. [21] bestimmen die Pose der Hand, indem sie ein gelenkiges Handmodell mit Hilfe von Punktregistrierung basierend auf Expectation-Conditional-Maximization-Ansätzen an korrespondierende 3D-Daten anpassen. Gorce et al. [22] rekonstruieren die 3D-Handpose auf Basis der Optimierung einer Zielfunktion mit Hilfe eines quasi Newton Ansatzes und eines parametrisierbaren Handmodells. Die Zielfunktion vereint Texturen und Schattierungen, um das Problem von Selbstverdeckungen zu behandeln und im Rahmen der Posenbestimmung wird ein der Realität möglichst entsprechendes synthetisches Handbild mit Hilfe eines gelenkigen Handmodells und RGB-Informationen erzeugt. Die Ansätze von Oikonomidis et al. bestimmen das Skelett der Hand mit Hilfe eines speziellen Handmodells und Partikelschwarmoptimierung (PSO) durch Minimierung einer Fehlerfunktion basierend auf Merkmalen der Haut und Kanten (Feature Maps) sowie hypothetischer Posen [3, 23]. In [24] wird ein dreistufiger Iterative Closest Point (ICP) Ansatz für die Handposenbestimmung mit Hilfe eines gelenkigen 3D-Handnetzmodells präsentiert. Für die Bestimmung der globalen Pose der Hand werden die Modelldaten mit den aktuellen Handdaten mit einer ICP auf diesen und den sichtbaren Bereichen der Modelloberfläche übereinandergelegt. Eine Detektion der Fingerspitzen gefolgt von einer inversen kinematischen Approximation resultiert in einer ersten Schätzung des Modells. Auf Basis einer finalen ICP wird die Modelloberfläche mit den realen Daten möglichst in Einklang gebracht. Schröder et al. [25] nutzen die inverse Kinematik für das Einpassen eines virtuellen Handmodells in die von der Kamera aufgezeichneten 3D-Daten. Das Modell besteht aus einem triangulierten Netz mit einem darunterliegenden kinematischen Handskelett. Die Anpassung des Modells beruht auf einer kleinste-Quadrate Optimierung basierend auf dem Datenpunkt-Dreieck-Abstand. Dieser Ansatz ähnelt stark dem in dieser Arbeit für die Posenbestimmung genutztem kinematischen Modell. Die wichtigsten und entscheidenden Unterschiede sind der Daten-Modell Abstand, der in dieser Arbeit auf einen gemittelten Daten-Knoten Abstand reduziert wird, sowie die Fähigkeit der automatisierten Größenskalierung des Modells.

Die folgenden Methoden bilden Vertreter der hybriden Verfahren. Die Methode aus Sridhar et al. [14] verwendet eine Detection-guided Optimierungsstrategie und kombiniert diese mit einer generativen Optimierung eines Handmodells auf Basis von einer Repräsentation der Tiefendaten unter Verwendung von gemischten Gauß-Modellen und eines auf RDFs fußenden Detektionsschrittes. Ballan et al. [12] gehen einen Schritt weiter und passen ein gelenkiges Handoberflächenmodell an die realen Daten der Hände an während diese interagieren und zum Beispiel einen kleinen Ball halten oder miteinander verschränkt werden. Um den Informationsverlust durch gegenseitige Verdeckungen zu reduzieren, verwenden sie mehrere Kameras. Es existieren weitere Ansätze, die Pro-

blematiken der Mensch-Objekt Interaktion wie das Greifen von Gegenständen, andere Manipulationen oder die Handposenbestimmung für spezielle computergestützte Entwurfsaufgaben adressieren [26–28] oder auf die Daten mehrerer Kameras zurückgreifen [13]. Eines der bekanntesten hybriden Verfahren ist das von Sharp et al. [15], da es beeindruckende Resultate für die Handposenbestimmung in Echtzeit auf Standard Hardware liefert. Als Nachteil ist allerdings die Verwendung einer GPU anzusehen. Um dem Verlust der Hand während der Bewegungsverfolgung entgegenzuwirken, kombiniert das System einen Neuinitialisierungsschritt für jedes Bild mit einem Anpassungsschritt eines Modells an die Daten basierend auf den zeitlichen Informationen. Zu diesem Zweck erfolgt die Formulierung einer „goldenen" Zielfunktion, die mit einem stochastischen Optimierungsansatz den Fehler zwischen der rekonstruierten Handpose in Form eines Handmodells und den realen Daten minimiert. Taylor et al. [8] kombinieren verschiedene Energiefunktionen zu einer Zielfunktion, die nach einem Neuinitialisierungsschritt mit Hilfe einer Gauss-Newton Optimierung für die Bestimmung der Handpose in Echtzeit auf einer Standard-CPU ohne Grafikkarte genutzt wird. Die Zielfunktion vereint verschiedenste Ansätze und betrachtet beispielsweise die Bedingungen, dass jeder Datenpunkt möglichst nahe an der Oberfläche des Handmodells liegen soll und seine Normale der des dichtesten Punktes ähnlich ist [29]. Andere Energieterme sorgen dafür, dass die Pose einer möglichst realen menschlichen Handpose gleicht und Gelenkbeschränkungen nicht missachtet werden [11, 30, 31]. Auch der zeitliche Verlauf wird berücksichtigt, indem eine Energiefunktion dafür Sorge trägt, dass sich aufeinanderfolgende Posen ähneln [11]. Weiterhin soll sich das Modell nicht selbst schneiden [12] und jede Fingerspitze in den Daten sollte möglichst in der Nähe einer Fingerspitze des Modells liegen [11, 13]. Dieser Ansatz ist ein Beispiel dafür, dass die hybriden Verfahren weit verbreitet sind und beinahe einen Standardansatz bilden. Sie ergänzen sich gegenseitig und sind miteinander kombinierbar.

Weitere aktuelle Forschungen gehen in die Richtung der Echtzeitfähigkeit ohne unterstützende Verwendung von GPUs und der Personalisierung der verwendeten Modelle für die Hand [32, 33], um die Genauigkeit und Robustheit zu erhöhen.

Es sind bereits kommerzielle Lösungen für die Handposenbestimmung und entsprechende Anwendungen verfügbar. Das Intel® Perceptual Computing SDK 2013[5] bildet in Kombination mit dem Creative Senz3D Time-of-Flight Kamera System eine Schnittstelle für die Bestimmung der Handpose bis zu einer Distanz von einem Meter. Dabei wird kein komplettes 3D-Modell der Hand bestimmt, sondern lediglich die Positionen der Fingerspitzen und verschiedener anatomischer Landmarken wie der Handfläche oder des Ellbogens geliefert. Auch die neueste Variante in Form des Intel Perceptual Computing SDK[6] in Kombination mit der RealSense Kamera bietet lediglich das Handtracking im Nahbereich. Das Leap Motion[7] System liefert in seiner ersten Version eine Bestimmung der Finger-

[5] https://software.intel.com/en-us/perceptual-computing-sdk, Januar 2018
[6] https://software.intel.com/realsense, Januar 2018
[7] https://www.leapmotion.com/product, Januar 2018

spitzen im sehr nahen Bereich des Sensors und wurde später um eine Art gelenkiges Modell erweitert. Der Sensor wurde komplett neu konstruiert und auf den Einsatz für die Befestigung am Kopf optimiert. Es ermöglicht die Bestimmung komplexer Posen auch während der Interaktion und findet als Orion[8] SDK im Bereich der virtuellen Realität Verwendung.

Obwohl die Bestimmung der Handpose bei der Entwicklung entsprechender Verfahren im Vordergrund steht, lässt die Vielseitigkeit der Ansätze mit nur wenigen Adaptionen die Bestimmung der Pose des menschlichen Körpers zu. Aus diesem Grund werden nachfolgend relevante Vertreter von Verfahren für die Körperposenbestimmung benannt. Taylor [34] bedient sich bekannter Gelenkpositionen in 2D Bildern, um die Konfigurationen eines zuvor definierten gelenkigen Skeletts des Menschen auf Basis der Korrespondenz von Punkten und Gelenken zu bestimmen. Bregler und Malik [35] führen das Produkt sogenannter Exponential Maps und Twist Motions ein, um die Pose des Körpers mit Hilfe von linearen Gleichungssystemen zu bestimmen und auf ein 3D-Modell zu übertragen. Andere Ansätze stammen aus dem Gebiet des maschinellen Lernens. Menschliche Posen werden als zweidimensionale Anordnungen von Gelenkpositionen definiert, die statistisch in Cluster mit ähnlichen Posen eingeteilt werden können. Anschließend finden Verfahren des maschinellen Lernens Verwendung, um Cluster spezifische Funktionen zu bestimmen, die das Mappen von Features zu jedem Cluster erlauben und eine Bestimmung der Pose ermöglichen [36]. Shakhnarovich et al. [37] benutzen visuelle Features für die Bestimmung der Pose auf Basis von zuvor trainierten Hashing-Funktionen. Alle bisherigen Ansätze basieren auf Daten von 2D Kameras, wohingegen [38] die Bildinformationen einer Time-of-flight Kamera und die lokale Suche innerhalb eines auf kinematischen Ketten basierenden Modells nutzen, um die 3D-Positionen von Gelenken des Körpers zu bestimmen. Shotton et al. [20] reduzieren das Problem der Positionsbestimmung einzelner Körpermerkmale auf ein pixelbasiertes Klassifizierungsproblem unter Verwendung der Daten einer Kinect Kamera. Haker et al. [39] behilft sich für die Posenbestimmung des Oberkörpers eines künstlichen neuronalen Netzes in Form einer SOM mit einer dem Oberkörper des Menschen ähnelnden Topologie. Im Rahmen dieser Arbeit bildet der Ansatz von Haker et al. [39] die Grundlage für die Entwicklung des Verfahrens für die Posenbestimmung mit Hilfe einer sSOM.

Gerade die Posenbestimmung des Körpers ist kommerziell weit verbreitet. So ist sie wesentlicher Bestandteil der Steuerung von Avataren oder Spielen auf Konsolen wie der Xbox360 und der Xbox One. Auch die zu der Kinect für Xbox 360 und Kinect für Xbox One veröffentlichen SDK[9] bieten die Möglichkeit der Posenbestimmung unter Windows auf Standard-PCs. Ein weiterer Vertreter für die Posenbestimmung am PC ist das SoftKinetic iisu SDK[10].

[8] https://developer.leapmotion.com/orion, Januar 2018
[9] https://developer.microsoft.com/de-de/windows/kinect, Januar 2018
[10] www.softkinetic.com, Januar 2018