Diplomarbeit

# Markerloses 3D-Tracking von Menschen und Bauteil zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems

ausgeführt um Zwecke der Erlangung des akademischen Grades

Diplom-Ingenieur

eingereicht an der

**Technische Universität Wien**
**Fakultät für Maschinenwesen und Betriebswissenschaften**

unter der Leitung von:
**Univ.-Prof. Dr.-Ing Sebastian Schlund**
(Institut für Managementwissenschaftern, Forschungsbereich: Mensch-Maschine-Interaktion)

und
**Patrick Rupprecht, MSc. MSc. MA.**
(Institut für Managementwissenschaftern, Forschungsbereich: Mensch-Maschine-Interaktion)

von
**David Kostolani, BSc.**

Matrikelnummer: 01427234

Diefenbachgasse 9 / 323

1150, Wien

Wien, 07.12. 2020

Vorname Nachname

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Ich habe zur Kenntnis genommen, dass ich zur Drucklegung meiner Arbeit unter der Bezeichnung

# Diplomarbeit

nur mit Bewilligung der Prüfungskommission berechtigt bin.

Ich erkläre weiters Eides statt, dass ich meine Diplomarbeit nach den anerkannten Grundsätzen für wissenschaftliche Abhandlungen selbstständig ausgeführt habe und alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur, genannt habe. Weiters erkläre ich, dass ich dieses Diplomarbeitsthema bisher weder im In- noch Ausland (einer Beurteilerin/einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe und dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Wien, 07.12. 2020

Vorname Nachname

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

The increase in the complexity of the daily activities of workers and operators has a negative effect on the productivity of manual labour. Therefore, the development of new forms of worker assistance has gradually become of significant interest. Lately, systems providing cognitive assistance aiming to reduce the complexity of the workload have been introduced. In this perspective, the integration of augmented reality as a form of cognitive assistance has become more popular. Projection of the work instructions directly on the surface of objects within the working area by the means of augmented reality can potentially benefit the productivity of the labour. In the area of site assembly, this transfers into an opportunity to ergonomically reshape the working environments in a more efficient way.

The concept of a full context-adaptivity of an augmented reality system calls for the knowledge of the information about the environment. This implies the use of recognition and tracking. Emerging technologies, such as depth cameras, provide useful information about the position of objects in 3D. These new technologies pave the way for new applications in the area of augmented reality for cognitive assistance.

This work targets the application of markerless workpiece and human pose estimation, as well as human-machine-interaction, for use in a spatial augmented reality assistance system. The system is based on the data provided by a static camera system consisting of three time of flight cameras. The thesis is structured in the following way. To establish an overview, a literature review focusing on the extraction of the relevant information by machine vision is presented. Afterwards, a concept will be developed according to the system specifications. Subsequently, the system is developed and evaluated with respect to the predefined criteria.

The system fulfils the functionality regarding the pose estimation of the objects of interest. The cluttered industrial environment and the capability to interact in real-time, however, remain a challenge for an industrial application.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Kurzfassung

Die stattfindende Steigerung der Komplexität der Arbeit wirkt sich negativ auf die Arbeitskraft sowie die Produktivität der manuellen Arbeit aus, weshalb das Gebiet der Assistenzsysteme immer mehr an Bedeutung gewinnt. Zu kognitiven Assistenzsystemen, die auf die Reduktion der Belastung durch kognitive Unterstützung zielen, zählen auch Augmented Reality Anwendungen. Im Bereich der Baustellenfertigung weist Augmented Reality das Potential auf, Arbeitsanweisungen direkt auf Objekte innerhalb des Arbeitssystems zu projizieren und somit die Arbeit ergonomischer und effizienter zu gestalten.

Um eine Kontextadaptivität zu gewährleisten, insbesondere durch die Projizierung relevanter Informationen am richtigen Ort und zur richtigen Zeit, wird Tracking relevanter Objekte vorausgesetzt. Neuartige Technologien, wie Tiefenbildkameras und die damit verbundene Kenntnis der Lage der Objekte im dreidimensionalen Raum, öffnen die Tür für neue Anwendungen auf dem Gebiet der Augmented Reality und der kognitiven Assistenzsysteme.

Diese Arbeit befasst sich mit der markerlosen Erfassung der Pose des Mitarbeiters sowie des Bauteils und der Gestaltung einer Mensch-Maschine-Schnittstelle basierend auf dreidimensionalen Daten bereitgestellt durch ein statisches Kamerasystem bestehend aus drei Time of Flight Kameras, wodurch in der Zukunft die Ansteuerung eines Spatial Augmented Reality Assistensysystems ermöglicht werden soll. Basierend auf der Literatur werden zunächst die Möglichkeiten zur Erkennung der relevanten Information aus den Daten mittels Machine Vision untersucht. Anschließend wird ein Konzept entsprechend der Spezifika des statischen Kamerasystems mit Time of Flight Kameras ausgearbeitet, entwickelt und abschließend auf die zuvor definierten Kriterien evaluiert.

Obwohl das System die Funktionalität des Systems in Hinsicht auf die Extraktion der 3D-Informationen erreicht wurde, stellt die geclutterte industrielle Umgebung sowie die Geschwindigkeit der Datenverarbeitung und dadurch die Fähigkeit zur Interaktion in Echtzeit eine Herausforderung für den industriellen Einsatz.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Danksagung

An dieser Stelle möchte ich mein Dank an alle Personen äußern, die mich bei der Umsetzung der vorliegenden Diplomarbeit unterstützt haben.

Zunächst möchte ich mich herzlich bei meinen Betreuern Prof. Sebastian Schlund und Patrick Rupprecht vom Institut für Managementwissenschaften für ihre hervorragende Betreuung, Zeit, konstruktive Kritik und insbesondere die konstante und geduldige Unterstützung während der Anfertigung dieser Arbeit bedanken. Ich bin sehr dankbar, dass ich meine Diplomarbeit zu so einem spannenden Thema verfassen durfte.

Weiters möchte ich meinen Dank an Prof. Markus Vincze, Stefan Thalhammer und Matthias Hirschmanner vom Institut für Automatisierungs- und Regelungstechnik für ihre wertvollen Tipps zur Gestaltung des Systems aussprechen.

Zusätzlicher Dank gilt an meine Studienkollegen, sowie meine Arbeitskollegen, die mir immer hilfsbereit zur Seite standen. Besonders möchte ich mich bei Hans Küffner-McCauley bedanken, dessen Korrekturen dazu beigetragen haben, dass diese Diplomarbeit in dieser Form vorliegt.

Zuletzt möchte ich meiner Familie bedanken, vor allem meinen Eltern, die mir mein Studium durch ihre bedingungslose Unterstützung ermöglicht haben und für mich immer da waren. Auch schulde ich einen Dank meiner Freundin, die mich während der Anfertigung dieser Arbeit tatkräftig unterstützt hat.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Inhaltsverzeichnis

**Abstract** i

**Kurzfassung** ii

**Danksagung** iii

**1 Einleitung** 3
- 1.1 Motivation und Einführung . . . 3
- 1.2 Problemstellung und Methodologie . . . 4
- 1.3 Aufbau der Arbeit . . . 6

**2 Grundlagen** 8
- 2.1 Digitale Assistenzsysteme in der Produktion . . . 8
  - 2.1.1 Mensch-Maschine Interaktion . . . 10
  - 2.1.2 Augmented-Reality basierte Assistenzsysteme in der Produktion . 12
- 2.2 Erfassung von Punktwolken . . . 14
  - 2.2.1 Das Pinhole-Modell . . . 14
  - 2.2.2 Methoden zur Aufnahme von 3D-Punktwolken . . . 15
  - 2.2.3 Time of Flight Kameras . . . 17
  - 2.2.4 Berechnung von Punktwolken . . . 20
  - 2.2.5 Kalibrierung von Kameras . . . 22
- 2.3 Verarbeitung visueller Daten . . . 24
  - 2.3.1 Merkmale der 3D-Computer Vision . . . 25
  - 2.3.2 Machine Learning in Computer Vision . . . 28

**3 Tracking und Ermittlung der Pose starrer Körper** 31
- 3.1 Einführung . . . 31
- 3.2 Vorstellung ausgewählter Methoden . . . 33

**4 Tracking und Ermittlung der Pose artikulierter Objekte** 39
- 4.1 Einführung . . . 39
- 4.2 Vorstellung ausgewählter Methoden . . . 42

**5 Implementierung** 47
- 5.1 Konzept des Spatial Augmented Reality Assistenzsystems . . . 47
- 5.2 Beschreibung des Set-Ups . . . 48



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.3 Anforderungen an das System . . . . . 49
5.4 Konzeptionierung und Beschreibung des Vorgehens . . . . . 51
5.4.1 Ausgangssituation . . . . . 51
5.4.2 Teilsystem Bauteil . . . . . 53
5.4.3 Teilsystem Mensch . . . . . 55
5.4.4 Ableitung der Architektur des Assistenzsystems . . . . . 56
5.5 Übersicht verfügbarer Bibliotheken . . . . . 58
5.6 Implementierung des Teilsystems Mensch . . . . . 60
5.6.1 Konzept zur Mensch-Maschine Interaktion . . . . . 67
5.7 Implementierung des Teilsystems Bauteil . . . . . 70
5.7.1 Preprocessing . . . . . 70
5.7.2 Erkennung, Verfeinerung und Verifizierung der Pose . . . . . 74
5.8 Evaluierung und Testing . . . . . 79
5.8.1 Bekannte funktionale Einschränkungen . . . . . 79
5.8.2 Evaluierung auf die nicht-funktionalen Anforderungen . . . . . 81

**6 Zusammenfassung und Diskussion** **83**
6.1 Diskussion der Ergebnisse und des Potentials des Systems . . . . . 84
6.2 Weitere Entwicklung und Optimierung des wahrnehmenden Systems . . . 86
6.3 Entwicklung einer Schnittstelle zur Projektion . . . . . 87

**Abbildungsverzeichnis** **89**

**Tabellenverzeichnis** **93**

**Literaturverzeichnis** **94**

**Anhang** **105**
Mathematische Grundlagen . . . . . 105
Ergänzende Beschreibung der softwaretechnischen Umsetzung . . . . . 107

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# 1 Einleitung

## 1.1 Motivation und Einführung

Die Ära der Industrie 4.0 (oft auch zweites Maschinenalter genannt) wird durch die Begriffe Komplexität und Flexibilität getrieben. Einerseits erwarten Kunden eine hohe Personalisierung der Produkte, was zu Variantenflexibilität führt, auf der anderen Seite erwarten sie aber auch, dass sie diese Produkte sehr schnell bekommen, wodurch auch Terminflexibilität ein wichtiges Thema wird. Dies führt zu einem hohen Komplexitätswachstum der Märkte und bildet für die Unternehmen ein neues Problem, da sie ihre Systeme zur Leistungserbringung gerade so komplex gestalten müssen, damit sie effektiv genug auf dem komplexen Markt unterwegs sein können. Auf der anderen Seite sollen diese aber einfach sein, damit die Komplexitätskosten nicht zu hoch werden. Dieses Problem betrifft alle Ebenen des Unternehmens, von Produkt- und Variantenanzahl, Fertigungs- und Montageprozessen bis zur Organisation und Koordination innerhalb sowie über die Unternehmensgrenzen hinaus. Zeitgleich ist in der Zukunft auch mit dem demographischen Wandel und Mangel an Fachkräften zu rechnen, wodurch Unternehmen eventuell Probleme haben können, die komplexen Anforderungen an ihre Produkte zu bewältigen und die Qualität sowie die Liefertreue zu gewährleisten (vgl. [1]).

In den letzten 30 Jahren wird oft von der sogenannten Humanisierung bzw. Demokratisierung der Arbeit gesprochen. Humanisierung der Arbeit brachte, im Gegensatz zu Taylorismus, eine Erweiterung der Arbeitsaufgaben, insbesondere die Ganzheitlichkeit der Arbeitstätigkeit. Diese besteht nicht nur aus ausführenden, sondern auch organisierenden, planenden und kontrollierenden Tätigkeiten, die ein Mitarbeiter ausführen soll. Ganzheitlichkeit der Arbeitstätigkeit hat eine lernförderliche Arbeitsgestaltung zur Folge, die die Innovationsfähigkeit der Unternehmen stark beeinflusst (vgl. [2]).

Erst in den letzten Jahren wurde die Lücke zwischen diesen zwei Konzepten durch das Gebiet der Assistenzsysteme geschlossen. Durch die zunehmende Digitalisierung soll nun die Komplexität der Arbeit bewältigt werden, und gleichzeitig soll der Mitarbeiter in der Lage sein, durch die digitale Erweiterung seiner kognitiven (und manuellen) Fähigkeiten Probleme einfacher, rascher und effizienter lösen zu können. Im Zusammenhang mit Assistenzsystemen wird in den letzten Jahren der Begriff Operator 4.0 verwendet. Relevant für diese Arbeit ist insbesondere der Begriff Augmented Operator, der eine Untergruppe des Operators 4.0 darstellt. Dieser Begriff bezeichnet die Erweiterung der Realitätswahrnehmung um virtuelle Objekte, die am Arbeitsplatz sonst nicht vorhanden wären, und gehört zu den Forschungsthemen der Industrie 4.0 (vgl. [3]).

Produktivität der manuellen Arbeit ist seit der Gründung des Gebiets des Produktionsmanagements eines der Themen, das im Vordergrund steht. Laut einiger Analysen



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[4] liegt das Potential der Produktivitätssteigerung der manuellen Arbeit durch die Digitalisierung der Informationen bei 45 – 55 %. Digitale Assistenzsysteme stellen dem Mitarbeiter Informationen zur Verfügung und sind daher in der neuen Ära als ein Werkzeug, das zur Produktivitätssteigerung führen soll, zu sehen. Die Forschung rund um Augmented Operator zielt auf eine Steigerung der Produktivität durch Bereitstellung der richtigen Informationen zur richtigen Zeit und am richtigen Ort. Diese Informationen können mittels Tablets und Handys, Datenbrillen oder Projektion bereitgestellt werden. Tablets sowie Handys setzen dabei das Greifen der Geräte zur Informationsbereitstellung voraus, Datenbrillen müssen hingegen vom Mitarbeiter während der Durchführung der Tätigkeiten getragen werden, wodurch für diese zwei Arten die Akzeptanz ein besonders wichtiges Thema wird. Dagegen setzt eine Projektion das Tragen von Sensorik nicht voraus. Dies wirkt sich entsprechend auf die Ergonomie der Arbeit aus, wodurch eine Projektion ein vielversprechendes Tool zur Aufbereitung der Informationen sein könnte (vgl. [5][6]).

## 1.2 Problemstellung und Methodologie

Das Voranschreiten der Automatisierung im Rahmen der dritten industriellen Revolution hat in vielen Gebieten der Industrie den versprochenen Nutzen, Rationalisierung und Wirtschaftlichkeit nicht erreicht. Dazu lassen sich viele Bereiche der Industrie nicht automatisieren, da sie entweder zu komplex sind und die Voraussetzungen für Automatisierung nicht erfüllen, oder der Aufwand ist zu hoch und eine Automatisierung wäre nicht wirtschaftlich. Dazu ist auch wegen der hohen Flexibilität zu erwarten, dass manuelle Arbeitsaufgaben auch weiterhin ein fester Bestandteil der Produktion in der Zukunft sein werden. Zu diesen Bereichen gehören zum Beispiel auch die Kleinserienfertigung und Montage, da die steigende Variantenvielfalt, sich nicht-wiederholende Aufgaben und oftmals auch die Notwendigkeit der höchsten, einem Menschen entsprechenden Präzision, eine Durchführung dieser Aufgaben durch Menschen voraussetzen. Ebenfalls in diesen Bereichen wird nach einer entsprechenden Steigerung der Produktivität gestrebt, die nun mittels digitaler Informationsaufbereitung ermöglicht werden soll (vgl. [2]).

Der Wunsch nach einer Produktivitätssteigerung betrifft ebenfalls den Bereich der Baustellenfertigung (z.B. Schiff- oder Flugzeugindustrie). Hier ist die Komplexität der Tätigkeiten entsprechend hoch und die Aufgaben wiederholen sich in der Regel nicht. Anleitungen für die Fertigung sowie die Montage erfolgen oft durch Bildschirme, die redundante Bewegungen hin zum Bildschirm und zurück verursachen und mindern dabei die Produktivität der Arbeit (vgl. [7]).

Arbeitssysteme in der Baustellenfertigung zeichnen sich in der Regel durch große räumliche Abmessungen aus, und stellen demnach eine Herausforderung für traditionelle Projektionssysteme dar. Dies kann mittels eines neuartigen Ansatzes unter Anwendung eines dynamischen Projektionssystems bewältigt werden, wodurch die Informationsbereitstellung direkt auf den Oberflächen realer Objekte großer Dimensionen erfolgen kann (vgl. [8]).



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Um die Informationen am richtigen Ort und zum richtigen Zeitpunkt projizieren zu können, werden einerseits die Informationen über das Arbeitssystem, insbesondere das Bauteil, sowie die Position des Mitarbeiters benötigt. Eine der Möglichkeiten zur Aufnahme der Bewegung stellt markerloses Tracking mittels Machine Vision dar. Für den Bereich der Baustellenmontage eignen sich aufgrund der hohen Reichweite besonders die Time of Flight Kameras. Time of Flight Kameras können die Umgebung dreidimensional abbilden und ihre Abbildungsfähigkeit ist im Vergleich zu anderen Kamerasystemen Textur- und Beleuchtungsunabhängig. Eine Kehrseite ist, dass die Daten oft keine Farbwerte besitzen (vgl. [9]). Reine 3D-Informationen ohne 2D-Farbdaten stellen aufgrund der geringeren Informationsdichte, und der oft vorkommender Unstrukturiertheit, eine Herausforderung für klassische Machine Vision Anwendungen dar (vgl. [10]).

Einsatz von Time of Flight Kameras ohne Verwendung der 2D-Farbdaten zur Ansteuerung eines Spatial Augmented Reality-basierten Assistenzsystems ist nach dem besten Wissen des Autors neuartig. Daraus abgeleitet, widmet sich diese Arbeit der softwaretechnischen Analyse der 3D-Aufnahmen des Arbeitsplatzes sowie der Auslegung einer Vision-basierten Mensch-Maschine-Schnittstelle zur Bereitstellung dieser Informationen zur dynamischen Projektion.

Der Forschungsschwerpunkt dieser Arbeit liegt in der Analyse vorhandener Algorithmen auf ihre Eignung zur Entwicklung eines Spatial Augmented Reality-Assistenzsystems basierend auf Punktwolken, sowie der Auswahl und der Konzeptauslegung für die anschließende Umsetzung. Dabei sollen folgende Forschungsfragen beantwortet werden:

**(F1)** Welche Herausforderungen sind zu bewältigen, um markerloses 3D-Tracking von Mitarbeiter und Bauteil mittels Time of Flight Kameras implementieren zu können?

**(F2)** Wie kann die Erkennung und das Tracking mittels Time of Flight Kameras generierten Daten gestaltet werden?

**(F3)** Welche Kriterien bestehen auf das Tracking und wie leitet sich daraus das Potential von 3D-Tracking mit Punktwolken zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems ab?

Abgeleitet aus den Forschungsfragen wird im Rahmen dieser Arbeit eine vertiefende Literaturrecherche durchgeführt, die auf die Auflistung und Evaluierung verschiedener Ansätze zur Erkennung und zum Extrahieren der zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems notwendigen Informationen zielt. Des Weiteren wird das bestehende Set-Up, das in der Pilotfabrik der TU Wien durch den Forschungsbereich Mensch-Maschine-Interaktion entwickelt wurde, auf die technischen Herausforderungen bedingt einerseits durch das Set-Up und die Hardware, andererseits durch den Einsatz der Machine Vision zur Erfassung der menschlichen Arbeit im Produktionsumfeld, untersucht. Aufbauend auf der Literaturrecherche besteht die Aufgabe in der Umsetzung geeigneter Algorithmen entsprechend den gegebenen Herausforderungen. Anhand der zuvor definierten Kriterien, gerichtet insbesondere an die Funktionalität, soll demnach



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

das Potential von Tracking mit Time of Flight Kameras zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems untersucht werden.

Die Relevanz dieser Arbeit lässt sich in zwei Bereiche aufteilen. Die Arbeit leistet einen theoretischen Beitrag zur Konzeption eines 3D-Vision-basierten Augmented Reality Assistenzsystems in der Baustellenfertigung. Hierzu wird, abgeleitet aus der Literaturrecherche, eine Architektur für Augmented Reality Assistenzsysteme mit Time of Flight Kameras vorgeschlagen. Weiters werden die Spezifika der Verarbeitung dreidimensionaler Aufnahmen großer Arbeitsbereiche aufgelistet und ein taugliches Konzept zum Aufbau einer Vision-basierten Schnittstelle zur Erfassung der Pose des Bauteils, des Mitarbeiters sowie Mensch-Maschine-Interaktion für AR-Assistenzsysteme beschrieben.

Die praktische Relevanz der Arbeit besteht in Auslegung der Anforderungen an die Informationsaufbereitung, wodurch ein Beispiel für die künftige Entwicklung 3D-Vision-basierten AR-Assistenzsysteme präsentiert wird. Des Weiteren stellt die Beschreibung der Konzeptumsetzung eine praxisnahe Vorlage zur Umsetzung von Vision-basierten Assistenzsystemen und somit einen Anreiz für künftige Entwicklungen in diesem Bereich, die zur Unterstützung der Mitarbeiter, Erhöhung der Ergonomie sowie der Produktivität der Arbeit beitragen können.

## 1.3 Aufbau der Arbeit

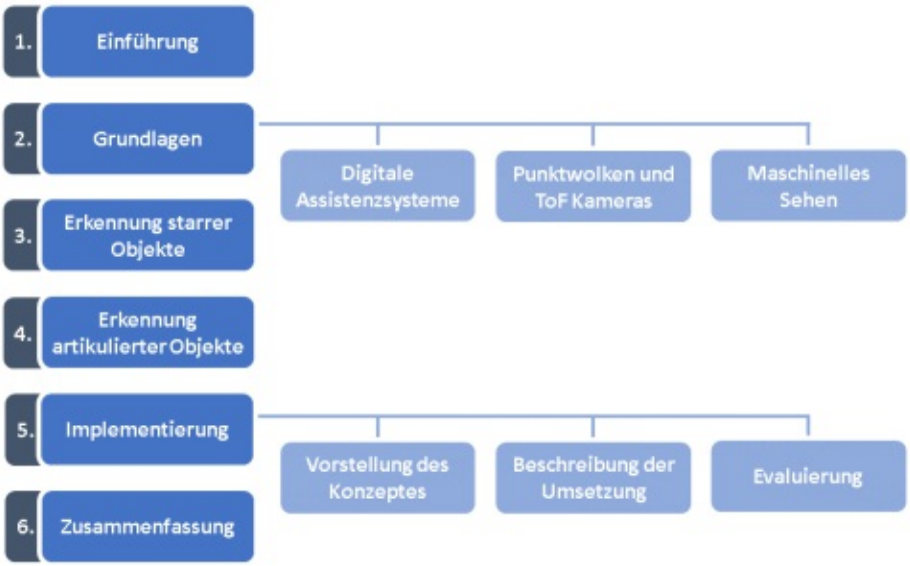



Abbildung 1.1: Veranschaulichung der Struktur der Arbeit

Die Arbeit gliedert sich in sechs thematische Bereiche. Im bestehenden Kapitel wurde zuerst auf die kurze Beschreibung der Ausgangssituation, die Motivation sowie die Ziele der Arbeit eingegangen. Im Anschluss daran werden im zweiten Kapitel theoretische



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Grundlagen und die Taxonomie der digitalen Assistenzsysteme vorgestellt. Der Fokus wird hierbei insbesondere auf Augmented Reality Assistenzsysteme sowie Grundlagen der Mensch-Maschine Interaktion gerichtet. Zunächst werden die Grundlagen der Erfassung von Punktwolken mit Time of Flight Kameras beschrieben. Im Anschluss daran werden die Grundlagen der softwaretechnischen Verarbeitung dieser Daten gelistet.

Aufbauend auf den Grundlagen wird im Kapitel drei zuerst die Problematik der Erkennung der Pose starrer Körper beschrieben, die insbesondere auf eine allgemeine Taxonomie der Ansätze zielt. Anschließend werden Beispiele für die Ansätze präsentiert, die als Kern zur späteren Auswahl der geeigneten Methode dienen. Analog wird die Beschreibung der allgemeinen Aufteilung der Ansätze zur Erfassung der Pose artikulierter Objekte im Kapitel vier vorgestellt.

Kapitel fünf beginnt mit einer Beschreibung des Set-Ups und der Auslegung der Anforderungen zur späteren Evaluierung des Potentials des entwickelten Systems. Danach erfolgt die Beschreibung der Herausforderungen und die Wahl der Herausforderungen entsprechender Algorithmen und Ansätze zur Umsetzung. Darauf aufbauend erfolgt die Konzeptauslegung und Entwurf der System-Architektur, sowie die Beschreibung der Implementierung. Abschließend wird das System evaluiert.

Im letzten Kapitel werden zuerst die Ergebnisse dieser Arbeit zusammengefasst, und im Anschluss daran einerseits das Potential sowie der Optimierungsbedarf diskutiert.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# 2 Grundlagen

## 2.1 Digitale Assistenzsysteme in der Produktion

Der derzeit stattfindende Wandel in der Industrie, bezeichnet auch als die vierte industrielle Revolution (Industrie 4.0), ist geprägt durch die omnipräsente Vernetzung von Maschinen und Betriebsmitteln, sowie die Integration von Software in die bestehenden Systeme. Der zeitgleich geschehende demografische Wandel und der Mangel an Fachkräften, bei gleichzeitig steigender Komplexität der Arbeitssysteme, die einerseits durch die hohe Flexibilität, andererseits durch die hohe Produktenvielfalt getrieben sind, erfordert das Schaffen eines neuen Arbeitsumfeldes, das den Mitarbeitern ermöglicht, weiterhin produktiv zu arbeiten. Um dies zu gewährleisten, werden Assistenzsysteme eingesetzt. Unter Assistenzsystemen wird die Integration technischer bzw. maschineller Unterstützung in ein Arbeitssystem verstanden, die zur Durchführung der Arbeitstätigkeit im Sinne von Interaktion zwischen dem Mitarbeiter und den Arbeitsmitteln beiträgt [11][12][13].

Eine Unterteilung der Assistenzsysteme nach der Art der Unterstützung erfolgt in physische, wahrnehmende und kognitive Assistenzsysteme. Im Sinne von Industrie 4.0 bilden die wahrnehmenden und kognitiven Assistenzsysteme die Instanz der digitalen Assistenzsysteme. Wahrnehmende Assistenzsysteme verfügen über dedizierte Sensorik zur Erfassung, Auswertung und Interpretation von Informationen und zeichnen sich daher durch die Fähigkeit der maschinellen Wahrnehmung aus. Wahrnehmende Mitarbeiterzentrische Systeme ergänzt um die kognitiven Fähigkeiten, wie Ableiten von Schlussfolgerungen, Problemlösen und Treffen von Entscheidungen, werden als kognitive Assistenzsysteme bezeichnet [14].

Eine Unterteilung digitaler Assistenzsysteme nach dem Zweck der Unterstützung kann wie folgt definiert werden [15]:

- **Aktivierung und Zielsetzung**
  Aktivierung und Zielsetzung bilden eine Voraussetzung zur Erbringung einer Leistung. Assistenzsysteme können ausgehend von aufgearbeiteten Informationen zeitgerecht auf den nächsten Schritt aufmerksam machen.

- **Wahrnehmung**
  Aufbereitung relevanter, für den Mensch oft nicht bekannter, Informationen und Bereitstellung dieser.

- **Integration der Informationen unter Berücksichtigung aktueller Situation**
  Durch die Integration des Domänenwissens, die durch die Vereinigung der Wahrneh-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

mung der Informationen, der aktuellen Situation sowie der Zielorientierung erfolgt, können richtige Informationen zur richtigen Zeit bereitgestellt werden.

- **Fällen von Entscheidungen, Auswahl der Handlung**
  Ableitend aus den Informationen wird eine Reihe von Anweisungen erstellt, mit deren Hilfe die Aufgabe bewältigt werden kann.
- **Ausführung von Aktionen**
  Entweder eine komplett autonome Umsetzung einer Aufgabe oder eine Assistenz bei der Ausführung einer Aufgabe.
- **Feedback zu Aktionen**
  Leistung einer Überprüfung der Aufgabe und ein anschließendes Feedback zu den möglichen Effekten der Aktion.

Insbesondere in Hinsicht auf die Herausforderungen des Zeitalters der Industrie 4.0 können kognitive Assistenzsysteme einerseits die Produktivität der Arbeit steigern, auf der anderen Seite auch den kognitiven Aufwand der Mitarbeiter reduzieren [16]. Trotzdem gilt das Potential digitaler Assistenzsysteme in vielen der Kernkompetenzen als nicht ausgeschöpft. Des Weiteren wird ein Zuwachs der Wichtigkeit einzelner Kernkompetenzen prognostiziert (siehe Grafik 2.1) [17].

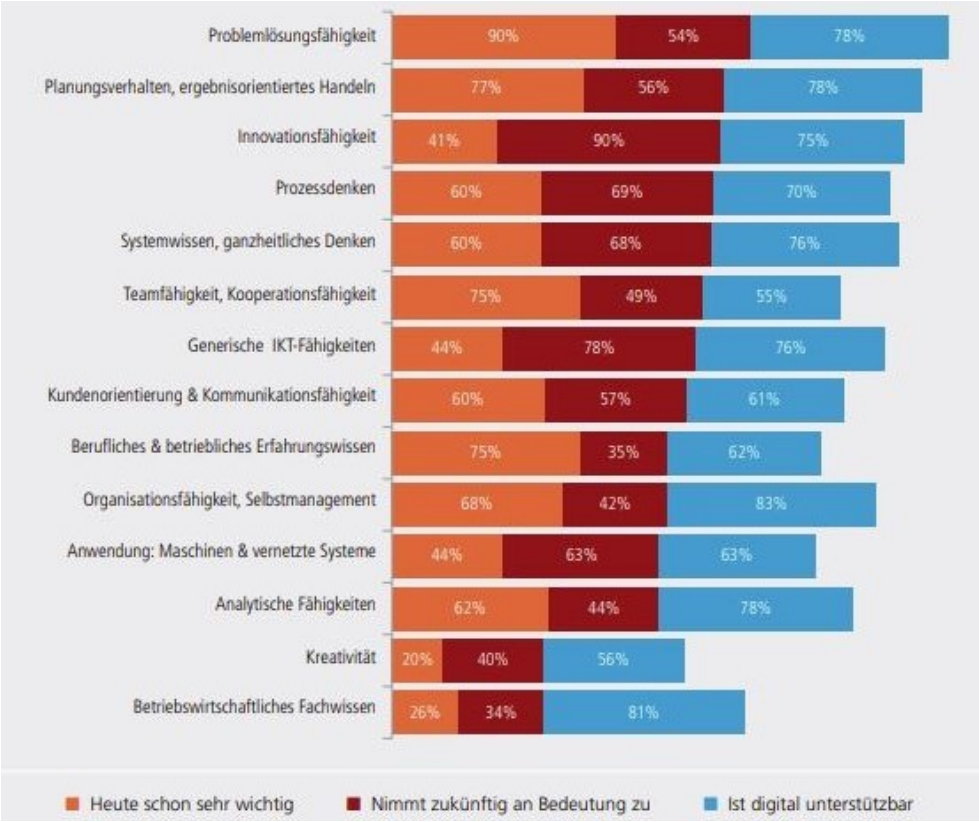



Abbildung 2.1: Umfrage zur Bedeutung verschiedener Kernkompetenzen und das Potential der digitalen Unterstützung [17]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

### 2.1.1 Mensch-Maschine Interaktion

Digitale Assistenzsysteme zeichnen sich durch eine Kommunikation mit dem Menschen aus. Die Geschichte der Interaktion zwischen Menschen und Maschinen im Sinne von digitalen Computern geht bis in die 80er Jahre zurück. Ausgehend von der steigenden Rechenleistung, die einen Wandel von einer limitierten Anzahl möglicher Operationen gesteuert durch einfache Unix-Commandos zu holistischeren Programmen zur Folge hatte, bestand der Bedarf nach einer höheren Usability (Englisch für Benutzerfreundlichkeit). Dieser Bedarf hat die Geburt einer neuen Forschungsdisziplin, die als Mensch-Maschine bzw. Mensch-Computer Interaktion (abgekürzt: MMI) bezeichnet wird, begründet (vgl. [18]).

Im arbeitswissenschaftlichen Kontext wird die Interaktion zwischen Menschen und Computern als ein System von Menschen, Computer und einer dazwischenliegenden Schnittstelle betrachtet (siehe auch Grafik 2.2). Die Schnittstelle wird dabei als derjenige Teil des Systems definiert, der alle Komponenten notwendig zur Kommunikation zwischen dem Menschen und dem Computer umfasst. Interaktion zwischen dem Menschen und der Schnittstelle erfolgt über wahrnehmende Rezeptoren und aktionsausführende Responder. Eine Übersicht der Rezeptoren und Responder relevant für das Gebiet der Mensch-Maschine Interaktion wird in Tabelle 2.1 präsentiert [18].

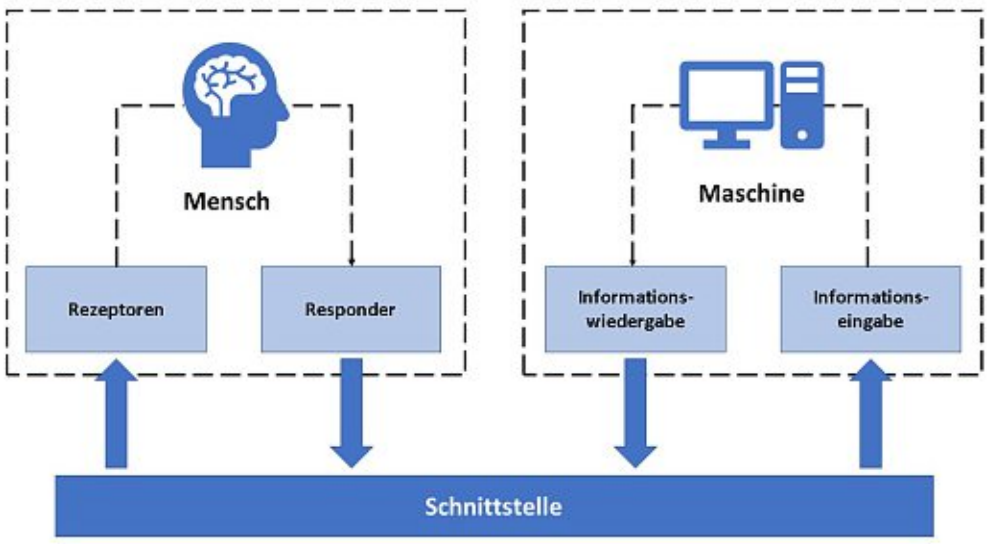



Abbildung 2.2: Veranschaulichung der Komponenten der Mensch-Maschine Interaktion (in Anlehung an [18])



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

| **Rezeptoren** | | **Responder** | |
|---|---|---|---|
| *Art* | *Beispiel* | *Art* | *Beispiel* |
| **Visuell** | Lesen textueller Informationen | **Bewegung** | Gesten, Drücken von Tasten |
| **Auditiv** | Arbeitsanweisungen über Sprachkommandos | **Stimme** | Spracherkennung |
| **Taktil** | Anregung zur Aktion über einen Vibrationsalarm | **Augen** | Eye-Tracking |

Tabelle 2.1: Auflistung der Rezeptoren und Responder des menschlichen Körpers und Beispiele für Anwendungen im Rahmen der MMI (in Anlehung an [18])

Die im Rahmen der Industrie 4.0 stattfindende Informatisierung und der zunehmende Einsatz von CPS (Englisch: Cyber Physical Systems) stellen höhere Anforderungen an die Schnittstellen. Durch die hohe Anzahl an Informationen drängt sich Usability in den Vordergrund. Als eine traditionelle Schnittstelle zur Interaktion mit Computern gelten Benutzerschnittstellen oder sogenannte GUIs (Englisch: Graphical User Interface). Diese Art der Interaktion stellt in Hinsicht auf die in Tabelle 2.1 angeführten Responder des menschlichen Körpers ungenutztes Potential dar. Im Laufe der Zeit hat die Entwicklung im Bereich der Software ermöglicht, Computerfunktionalität über andere Kanäle und in einer für die Menschen natürlicheren Art in die reale Welt zu integrieren[19][20].

Viele der neuen Ansätze zur Computerinteraktion lassen sich unter dem Paradigma Natural User Interface (abgekürzt: NUI) kategorisieren. Natural User Interfaces basieren auf einer direkten Informationseingabe in den Computer über Gesten und Mimik. Die Erkennung der Geste erfolgt dabei entweder visuell (etwa Kameras), sensorisch oder berührend durch z.B. Multi-Touch-Screens [14][19].

In gewissen Bereichen besteht die Möglichkeit zur Interaktion mittels Hände nicht. Dies schließt die Anwendung von GUIs sowie NUIs aus bzw. mindert die Effektivität des Einsatzes einer solchen Schnittstelle. In diesem Fall bietet sich Sprachsteuerung zur Interaktion mit der Maschine. Dies wird auch als Voice User Interface (abgekürzt: VUI) bezeichnet [15][21].

Ausgehend von der Unterscheidung zwischen den perzeptiven und ausführenden Fähigkeiten der Menschen wird die Unterteilung der bestehenden Schnittstellen für Assistenzsysteme in Grafik 2.3 vorgestellt [14].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

(a) Möglichkeiten zur Informationswiedergabe (b) Möglichkeiten zur Bedienung von Maschinen

Abbildung 2.3: Taxonomie der informationstechnischen Schnittstellen (in Anlehnung an [14])

### 2.1.2 Augmented-Reality basierte Assistenzsysteme in der Produktion

Unter Augmented Reality (weiters: AR) wird eine Integration virtueller Elemente in die reale Umgebung bezeichnet. Die Voraussetzungen für AR sind die Einbeziehung der dreidimensionalen Beziehungen der virtuellen und realen Objekte, und deren volle oder teilweise Überlagerung und Interaktionsfähigkeit in Echtzeit. Assistenzsysteme basierend auf Augmented Reality weisen hohes kognitives Unterstützungspotential auf, besonders auf dem Gebiet der Montage sowie der Instandhaltung. Einsatzbereiche von AR-basierten Assistenzsystemen können wie folgt unterteilt werden [22][23][24]:

- Arbeitsplanung und Entwurf von Arbeitsanweisungen
- Montageunterstützung durch Informationswiedergabe
- Training

Eine typische Architektur (siehe auch Abbildung 2.4) eines Augmented Reality Assistenzsystems besteht aus folgenden Bestandteilen (in Anlehnung an [24]):

- Datenaufnahme – Erfassung visueller Daten durch Kameras
- Processing – Ableiten relevanter Merkmale aus den Daten
- Tracking – Erfassung der Bewegung auf Basis der extrahierten Merkmale
- Mensch-Maschine Interaktion – Erfassung des menschlichen Inputs (i.d.R. visuell, haptisch etc.)
- Informationsmanagement – Datenmanagement aller zu widergebenden Komponenten
- Rendering – Projizierung und Überlagerung der Informationen in die reale Umgebung



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

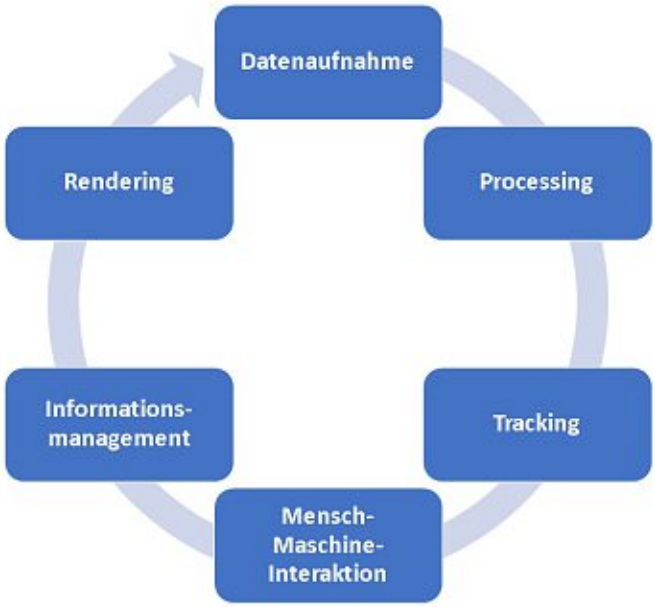



Abbildung 2.4: Bestandteile der Architektur von Augmented Reality Assistenzsystemen (in Anlehung an [24])

AR-basierte Assistenzsysteme werden weiterhin nach der Art der Aufnahme in statische (Szene-orientierte) und dynamische (z.B. First-Person-View) unterschieden. Eine statische Aufnahme weist den Vorteil auf, dass der Range-of-Interest (weiters: ROI) statisch bleibt. Dadurch wird eine vollständige Abbildung des Arbeitssystems garantiert. Des Weiteren ist dadurch der ganze Körper des Mitarbeiters sichtbar. Die Nachteile der statischen Kamerasysteme sind die notwendige Rekonfiguration im Falle einer Änderung des ROI, sowie die abnehmende Auflösung mit dem steigenden Abstand von dem ROI bzw. einem Objekt im ROI. Dynamische Arbeitssysteme zeichnen sich dagegen durch ein gutes Abbild der Händebewegungen, sowie durch eine Close-Up-Aufnahme der Objekte im Zentrum der Aufmerksamkeit [25].

Die Endgeräte zur Informationswidergabe können weiters wie folgt klassifiziert werden [26]:

- Handys / Tablets
- Bildschirme
- Augmented Reality Brille
- Projektion (2D oder 3D)

Ein projektionsbasierter Ansatz zeichnet sich im Vergleich zu anderen Methoden als besonders geeignet aus, da einerseits der Arbeitskomfort nicht beeinträchtigt wird (z.B. durch die Reduktion des Sichtfeldes), andererseits erfordert dies das Mittragen eines Gerätes nicht. Darüber hinaus weist sich die Projizierung der Informationen direkt auf



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

die Bauteile in Bezug auf die Produktivität der Arbeit vorteilhafter aus. Des Weiteren bieten sich projektive, Laser-basierte Assistenzsysteme auch zum Einsatz zur Qualitätssicherung an, wodurch Kosten für zusätzliche Prüfungssysteme entfallen [23][27].

Zu den einschränkenden Faktoren der Projektion gehören insbesondere die Reflektivität der Oberflächen, auf die die Projizierung erfolgt, sowie die Qualität des Abbilds. Die Abbildqualität ist im Allgemeinen von dem Einfallwinkel der Projektion abhängig, wodurch Pixelverzerrungen vorkommen können [28].

## 2.2 Erfassung von Punktwolken

### 2.2.1 Das Pinhole-Modell

Das Pinhole-Modell ist ein vereinfachtes optisches Modell zur mathematischen Beschreibung der Aufnahme von 3D-Objekten und der anschließenden Projektion auf eine 2D-Ebene (siehe Abbildung 2.5). Die Koordinaten des 3D-Raums, ausgehend von dem Koordinatenursprung der Kamera, werden hier als $x$, $y$, $z$ bezeichnet, hingegen werden die Koordinaten der 2D-Ebene der Pixel mit $u$, $v$ bezeichnet. Die Kamera befindet sich im optischen Zentrum zwischen der dreidimensionalen Szene und dem zweidimensionalen Abbild auf der optischen Achse, die senkrecht auf die Kameralinse steht. Der Abstand zwischen der Kamera und der Abbildebene wird hier als $b$ bezeichnet (vgl. [29][30]).

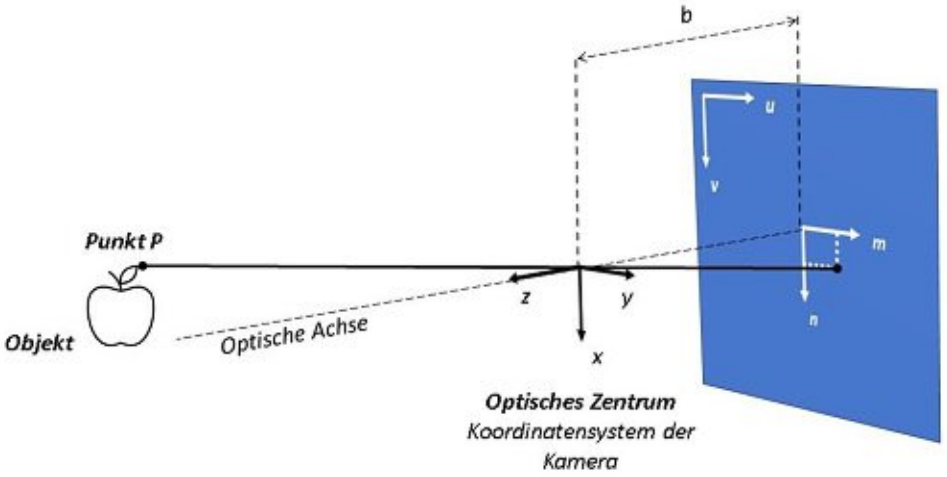



Abbildung 2.5: Darstellung des Pinhole-Modells (vgl. [29][31])

Die Transformation aus dem euklidischen Raum kann wie folgt formuliert werden: als erstes wird ein Hilfskoordinatensystem $m$, $n$ in der zweidimensionalen Pixelebene eingeführt, dessen Koordinatenursprung auf der optischen Achse liegt. Dann gilt für einen in das $m$, $n$ Koordinatensystem zu projizierenden Punkt $P$ mit den ursprünglichen Koordinaten $x$, $y$, $z$ [29]:



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$m = -b \cdot \frac{x}{z} \tag{2.1}$$

$$n = -b \cdot \frac{y}{z} \tag{2.2}$$

Dabei werden alle Punkte liegend auf der Geraden zwischen $P$ und seiner Projektion in das Koordinatensystem $m$, $n$ in den gleichen Punkt auf der 2D-Ebene projiziert. Dies impliziert, dass ihre ursprüngliche Tiefeninformation bei der Transformation verloren geht [29].

Aufgrund der Verzerrungen und Abweichungen, die aufgrund der nicht-idealen Geometrie der Optik unvermeidbar sind, weichen die Koordinaten $m$, $n$ von den unter den idealen Bedingungen transformierten Koordinaten ab. Das Verhältnis zwischen den realen und den idealen Koordinaten (bezeichnet durch ein Tilde) wird durch das Heikkila-Modell beschrieben [31]:

$$\begin{bmatrix} m \\ n \end{bmatrix} = \psi^{-1} \cdot \begin{bmatrix} \hat{m} \\ \hat{n} \end{bmatrix} \tag{2.3}$$

$$\begin{bmatrix} m \\ n \end{bmatrix} = \begin{bmatrix} \hat{m} \cdot (1 + k_1^2 + k_2^4 + k_3^6) + 2d_1 \cdot \hat{m} \cdot \hat{n} + d_2 \cdot (r^2 + 2\hat{m}^2) \\ \hat{n} \cdot (1 + k_1^2 + k_2^4 + k_3^6) + d_1 \cdot (r^2 + 2\hat{n})^2 + 2d_2 \cdot \hat{m} \cdot \hat{n} \end{bmatrix} \tag{2.4}$$

$$r = \sqrt{(\hat{m} - c_x)^2 + (\hat{n} - c_y)^2} \tag{2.5}$$

mit $c_x$ und $c_y$ als Koordinatentransformation zwischen $u$, $v$ und $m$, $n$:

$$\begin{bmatrix} m \\ n \end{bmatrix} = \begin{bmatrix} u - c_x \\ v - c_y \end{bmatrix} \tag{2.6}$$

Die Parameter $k_i$ und $d_i$ stehen dabei jeweils für die radiale und die tangentiale Verzerrung.

### 2.2.2 Methoden zur Aufnahme von 3D-Punktwolken

Wie im Kapitel 2.2.1 beschrieben, gehen die ursprünglichen Tiefenkoordinaten bei der Abbildung auf eine zweidimensionale Ebene verloren. Eine Aufnahme von Daten, aus denen die ursprünglichen Koordinaten nachbildet werden können, erfordert daher die Verwendung spezieller Methoden. Zu den Verfahren für die Aufnahme von Tiefendaten, aus denen anschließend eine Punktwolke gewonnen werden kann, zählen folgende Methoden [9][32]:

- Passive Triangulation
- Aktive Triangulation
- Laser-basierte Systeme
- Radar-basierte Systeme



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Passive Triangulation ist ein Verfahren zur Gewinnung der Tiefendaten aus mehreren zweidimensionalen Abbildungen durch den Vergleich der Bildkoordinaten u, v der jeweiligen Aufnahme. Der Vergleich der Koordinaten erfolgt dabei softwaretechnisch, im ersten Schritt werden Punkte bzw. Features in den Aufnahmen erkannt, anschließend werden korrespondierende Features zwischen den Aufnahmen gepaart, und ihre Koordinaten verglichen [9][31][32].

Eine Voraussetzung für passive Triangulation ist eine gleichzeitige Aufnahme durch zwei bzw. mehrere Kameras oder eine Mehrfachaufnahme aus unterschiedlichen Perspektiven. Des Weiteren müssen die physischen Orte, aus denen die Aufnahme erfolgt, mit dem aufzunehmenden Objekt ein Dreieck schließen (siehe Grafik 2.6). Die Reichweite bei passiver Triangulation ist von dem Abstand der Orte der Aufnahme abhängig. Weiters ist die Qualität des Abbilds aufgrund der mathematischen Koordinatenabhängigkeit von der Auflösung der Bilder abhängig, sowie von der Beleuchtung und der Genauigkeit der Matching-Algorithmen zum Vergleich der Features [9][31][32].

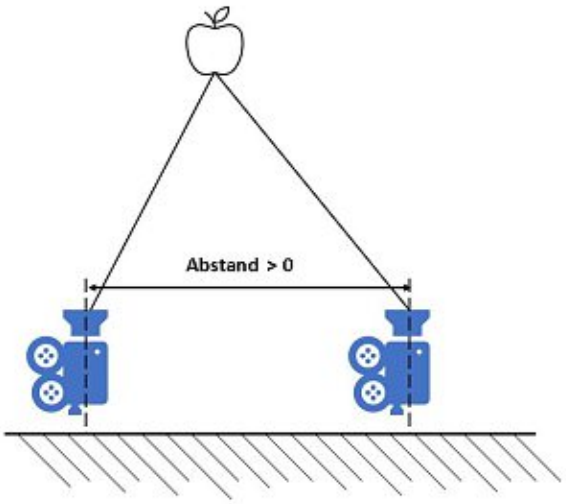



Abbildung 2.6: Veranschaulichung des Prinzips der passiven Triangulation (in Anlehnung an [31])

Aktive Triangulation erfordert im Vergleich zu der passiven Triangulation nur eine Kamera zusammen mit einer Lichtquelle, die die Umgebung beleuchtet. Aufgrund der Struktur der Beleuchtung, die sich durch Muster aufzeichnet, wird das Verfahren auch als Streifenlichtscanning bezeichnet (Englisch: structured-light scanning). Ähnlich wie bei passiver Triangulation, erfordert die aktive Triangulation das Schließen eines Dreiecks durch die Kamera, das Objekt und die Lichtquelle (siehe Grafik 2.7). Abgeleitet aus den Koordinaten auf der 2D-Pixeleben, unter der Voraussetzung, dass der Abstand der Lichtquelle sowie der gleichbleibende Einfallwinkel der Lichtquelle bekannt sind, kann der Abstand von der Kamera direkt berechnet werden [9][33].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

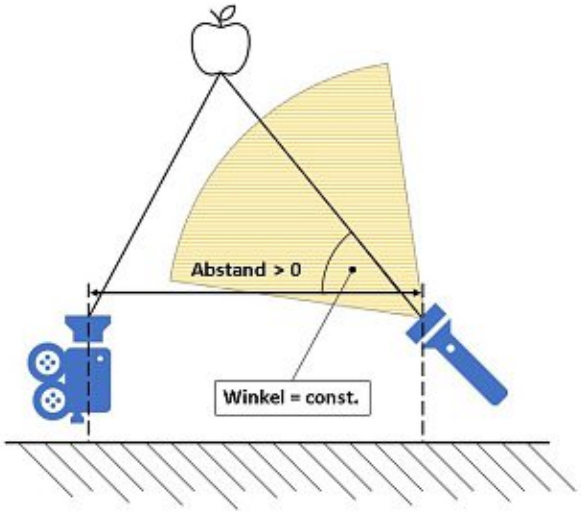



Abbildung 2.7: Veranschaulichung des Prinzips der aktiven Triangulation (in Anlehnung an [33])

Zu den typischen Vertretern der Laser-basierten Verfahren zählen LiDAR sowie Time of Flight Kameras. Hierbei wird zur Messung des Abstandes ein optisches Signal verwendet, das in der Regel moduliert wird. Das Prinzip basiert auf der Zeitmessung zwischen der Emission des Strahls und der Reflexion durch die Bestrahlung der Objekte. Dadurch aufgenommene Punktwolken beinhalten in der Regel keine spektrale Information, d.h. die Punkte sind monochrom. Die Abbildungsgenauigkeit sowie die Reichweite sind bei diesen Verfahren in der Regel hoch [9][32][34].

Die Radar-basierten Verfahren werden insbesondere für geoinformatische Zwecke verwendet. Ein typischer Vertreter ist hierbei SAR (Englisch: Synthetic Aperture Radar). Das Verfahren basiert auf der Reflexion der zuvor ausgestrahlten Wellen im Radiofrequenzbereich. Im Vergleich zu anderen optischen Verfahren ist SAR, aufgrund der abweichenden Wellenlänge der Radiowellen, unabhängig von der Beleuchtung, sowie von den Wettereinflüssen [32][35].

### 2.2.3 Time of Flight Kameras

Time of Flight Kameras (weiters: ToF) verwenden Wellenreflexion zur Messung der Tiefeninformation, ähnlich wie LiDAR oder SAR. ToF Kameras bestehen aus einem Transmitter, der den optischen Strahl emittiert, und einem Sensor, der die Reflexion des Strahles von der Oberfläche eines Objekts erfasst (siehe Abbildung 2.8) [31].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

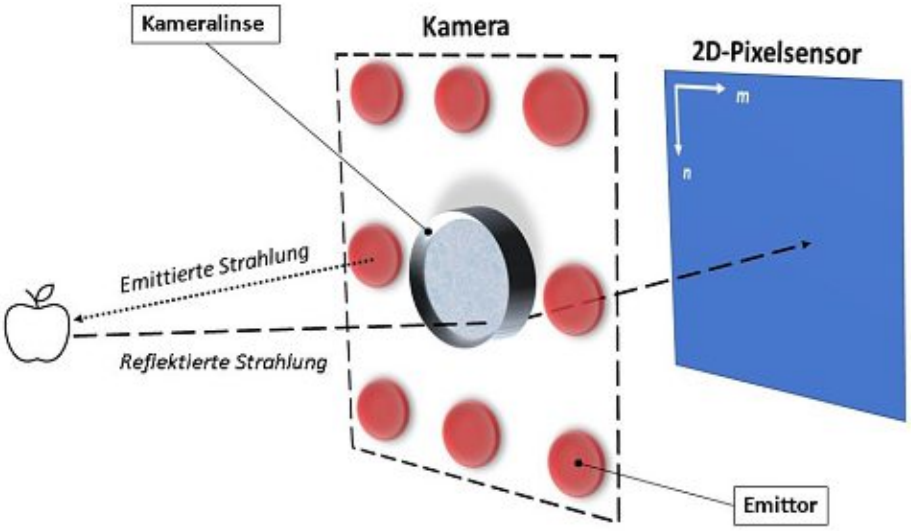



Abbildung 2.8: Veranschaulichung des Prinzips von Time of Flight Kameras (in Anlehung an [31])

Der Sensor einer ToF Kamer besteht aus $M$ x $N$ Zellen entsprechend der Anzahl der Pixel. Dadurch kann der Abstand von der Oberfläche der Objekte je Pixel erfasst werden, wodurch sogenannte Depth-Maps entstehen (siehe auch Grafik 2.9). Die Wellenlänge der emittierten Strahlung liegt im NIR-Bereich (Nah-Infrarot), in der Regel wird zusätzlich ein optischer Band-Pass-Filter um diesen Bereich angewendet, um die reflektierte Strahlung von Störsignalen zu bereinigen. Im Idealfall sollten der Transmitter und der Sensor parallel bzw. ko-positioniert sein. In der Praxis heißt dies, dass der Transmitter und der Sensor möglichst nah aneinander positioniert sind (siehe Grafik 2.8). Diese Aufstellung meidet Okklusionen, die bei einer Triangulation oft auftauchen können und werden durch Blockierung der Sicht einer der Kameras (etwa durch Ecken der Objekte) auf die Features verursacht. Dies meidet demnach das Matching von Features und taucht in Form von Löchern bzw. fehlerhaften Daten in Depth-Maps auf [31][36].

Abbildung 2.9: Darstellung der Amplituden(links), Intensitäten(Mitte) und der Depth-Map(rechts) [31]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Das Messprinzip berührt auf der Erfassung der Zeit, die vergeht, bis die Reflexion des ausgestrahlten Signals vom Sensor erfasst wird. Die Messung erfolgt direkt oder indirekt. Bei direkten Messverfahren wird der Phasenunterschied und somit die Zeit direkt durch den Sensor erfasst. Bei indirekten Verfahren wird der Phasenunterschied nicht direkt erfasst, sondern aus dem Korrelationsintegral des erfassten Signals und des Referenzsignals berechnet [31].

Übliche Formen der emittierten Signale sind gepulste Modulation sowie CW-Modulation (Englisch: Continuous Wave), dies wird in Grafik 2.10 dargestellt. Gepulste Modulation weist den Vorteil auf, dass Pulse höhere Signal-zu-Rauschen-Verhältnisse erreichen können. Die Pulse sind somit gegenüber äußeren Einflüssen, wie z.B. Hintergrundbeleuchtung, weniger anfällig und erlauben geringere Intensitäten des Strahls. Die Form des reflektierten Strahls ist von der Dämpfung, die den Strahl auf dem Weg begegnet, abhängig [31][37].

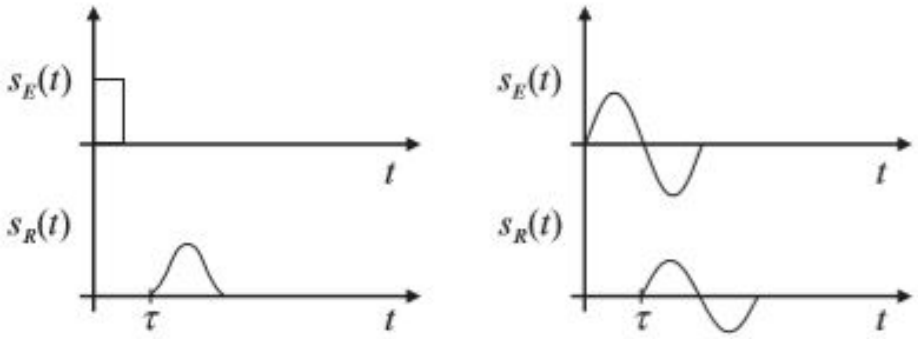



Abbildung 2.10: Emittiertes und reflektiertes Signal bei gepulster Modulation(links) und bei CW-Modulation(rechts) [31]

Time of Flight Kameras erfassen neben Tiefendaten (in Form von Depth-Maps) ebenfalls Intensitäten und Amplituden (siehe auch Grafik 2.9). Zur Aufklärung dieser Begriffe wird nun die üblicherweise verwendete indirekte Methode vorgestellt. Dazu wird eine Sinus-Modulation betrachtet [31]:

- die Modulation des emittierten Signals $m_E$

$$m_E(t) = A_E \cdot (1 + \sin(2\pi_m + \psi_m)) \tag{2.7}$$

- das bereinigte reflektierte Signal $m_R$

$$\begin{aligned} m_R(t) &= A_R \cdot (1 + \sin(2_m + \psi_m + \Delta\psi) + B_R) \\ \Rightarrow m_R(t) &= A \cdot \sin(2\pi_m + \psi_m + \Delta\psi) + \frac{B}{2} \end{aligned} \tag{2.8}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Dabei wird $A = A_R$ als die Amplitude des empfangenen Signals bezeichnet, das um die Absorption der Oberfläche gemindert wurde, und $B = \frac{A_R+B_R}{2}$ als die Intensität bezeichnet, die zusätzlich um $B_R$ als die Größe der Interferenz, die z.B. durch die Hintergrundbeleuchtung verursacht ist, gemindert wird. Weiters wird die Referenzfunktion definiert [31]:

$$g_R(t) = \frac{2}{T_m} \cdot (1 + \cos(2\pi_m + \psi_m) \tag{2.9}$$

Wird das reflektierte Signal mit der Referenzfunktion über die Modulationsperiode $T_m$ kreuzkorreliert, folgt [31]:

$$\begin{aligned} c_R(t) &= \int_0^{T_M} m_R(t) \cdot g_R(t+T) \\ \Rightarrow c_R(t) &= A \cdot \sin(\Delta\psi - 2\pi_m) + B \end{aligned} \tag{2.10}$$

Mit der Kenntnis der Modulationsfrequenz $f_m$ verbleiben noch drei Unbekannte: $A$(die Amplitude), $B$(die Intensität) und der Phasenunterschied $\Delta\psi$ auf der rechten Seite der Gleichung 2.10. Daraus abgeleitet muss bei indirekten Methoden die Korrelationsschaltung die Kreuzkorrelation vier Mal auswerten, um die Gleichung aufzulösen und den Phasenunterschied, und somit den Zeitabstand zu berechnen. Diese Berechnung liefert als Nebenprodukt gleichzeitig die Größe der Amplituden und die Intensitäten. Abschließend, nach dem der Zeitabstand zwischen dem emittierten und dem reflektierten Signal vorliegt, kann der radiale Abstand pro Pixel wie folgt ermittelt werden [31]:

$$r_P = \frac{c \cdot \Delta\psi}{2} \tag{2.11}$$

### 2.2.4 Berechnung von Punktwolken

Soweit wurde im Kapitel 2.2.3 vorgestellt, wie der radiale Abstand Pixel-weise berechnet werden kann. Darauf aufbauend wird in diesem Kapitel die mathematische Gewinnung der Punktwolken aus dem radialen Abstand erklärt. Dazu wird das Pinhole Modell einer ToF-Kamera betrachtet (siehe auch Grafik 2.11).



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

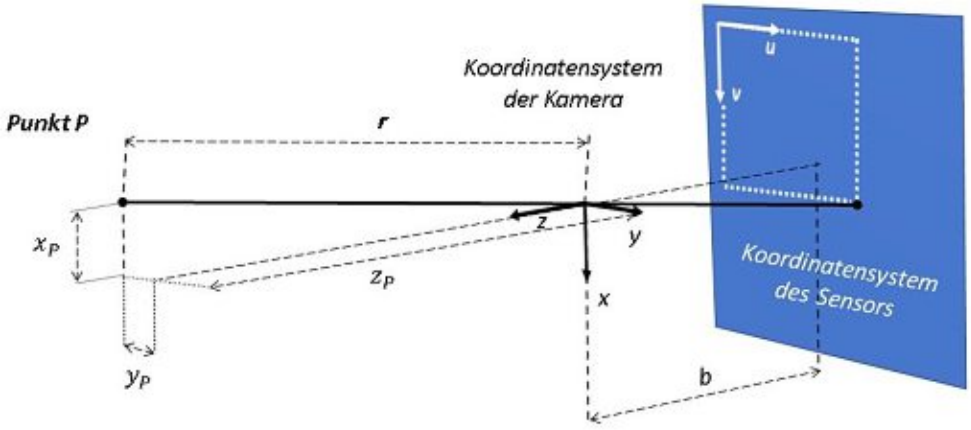



Abbildung 2.11: Koordinaten eines Punktes (in Anlehnung an [31])

Es gilt dabei für den radialen Abstand des Punktes von dem Koordinatenursprung [31]:

$$r = \|[x_P, y_P, z_P]\| = \sqrt{x_P^2 + y_P^2 + z_P^2} \tag{2.12}$$

Mithilfe projektiver Geometrie (und unter Vernachlässigung der Verzerrungen definiert in Gleichung 2.4) kann die Berechnung der $z$-Koordinate eines Objekts mit $K_T$ definiert in 2.13 und $m$, $n$ und $c$ wie in 2.6 wie folgt angenähert werden [31]:

$$K_T = \begin{bmatrix} b_x & 0 & c_x \\ 0 & b_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.13}$$

$$z_P \cdot \begin{bmatrix} u_P \\ v_P \\ 1 \end{bmatrix} = K_T \cdot \begin{bmatrix} x_P \\ y_P \\ z_P \end{bmatrix} \tag{2.14}$$

$$z_P = \frac{r_P}{\left\| K_P^{-1} \cdot [u_P, v_P, 1]^T \right\|} \tag{2.15}$$

Abschließend können durch die Invertierung der Gleichung 2.15 und Einsetzen der $z$-Koordinate die Koordinaten $x$ und $y$ wie folgt ermittelt werden [31]:

$$\begin{bmatrix} x_P \\ y_P \\ z_P \end{bmatrix} = K_P^{-1} \cdot \begin{bmatrix} u_P \\ v_P \\ 1 \end{bmatrix} \cdot z_P \tag{2.16}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

### 2.2.5 Kalibrierung von Kameras

Kameras erfordern vor ihrer Anwendung eine Kalibrierung, um eine genaue optische Abbildung durch das Gerät sicherzustellen. Als Kalibrierung wird generell die Erfassung der Abweichung der Werte zur Erhöhung der Genauigkeit sowie der Präzision bezeichnet. Die Genauigkeit bezieht sich dabei auf die Abweichung des Mittelwerts eines zu untersuchenden Parameters von seinem wahren Wert. Auf der anderen Seite bezeichnet die Präzision die Breite der Verteilung der Messwerte [31].

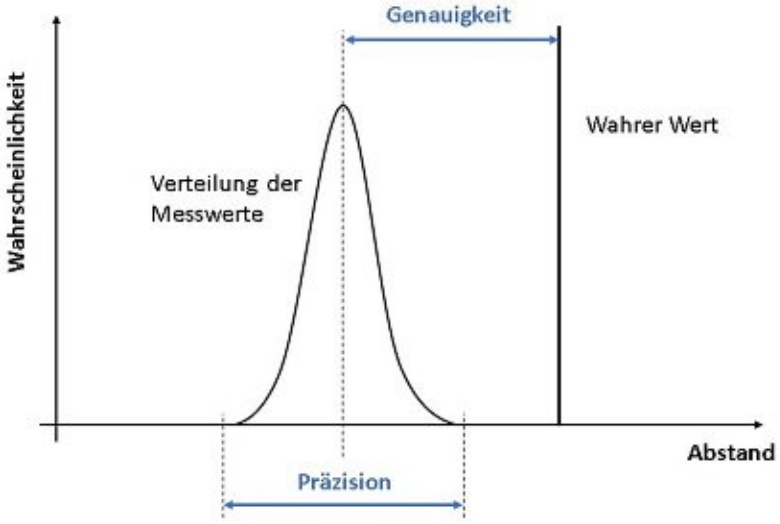



Abbildung 2.12: Veranschaulichung der statistischen Präzision und der Genauigkeit (in Anlehnung an [31])

Die Kalibrierung erfolgt entweder parametrisch (modelbasiert) oder nicht-parametrisch (ohne Model). Parametrische Kalibrierung setzt die Kenntnis bzw. die Annahme einer funktionalen Beziehung zwischen dem wahren und dem erfassten Wert der Kamera. Mit Hilfe des funktionalen Zusammenhangs, wird die Kalibrierung zu einem mathematischen Problem der Parameterermittlung. Auf der anderen Seite wird bei nicht-parametrischer Kalibrierung keine vorherige Kenntnis oder Annahme der Parameter vorausgesetzt, stattdessen wird durch einen paarweisen Vergleich von jeweils einem wahren Wert und einem erfassten Wert ein Zusammenhang im Laufe der Kalibrierung gebildet [31].

Weiterhin werden Kalibrierungsmethoden in überwachte und nicht-überwachte Methoden unterteilt. Überwachte Kalibrierung setzt die Anwendung eines Objekts mit zuvor bekannten geometrischen Eigenschaften voraus. Durch das Matching der bekannten Geometrie mit der durch das Gerät erfassten Geometrie kann anschließend ein Zusammenhang abgeleitet werden. Dies kann entweder ein 3D-Objekt, ein 2D-Objekt oder ein eindimensionales Objekt sein, häufig werden insbesondere planare Schachbretter verwendet (siehe auch Grafik 2.13). Nicht-überwachte Kalibrierung ermittelt die Werte durch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

eine Vielzahl an Aufnahmen und deren Vergleich. Dieses Verfahren erreicht allerdings in der Regel nicht dieselbe Genauigkeit, wie überwachte Methoden [29][30][31].

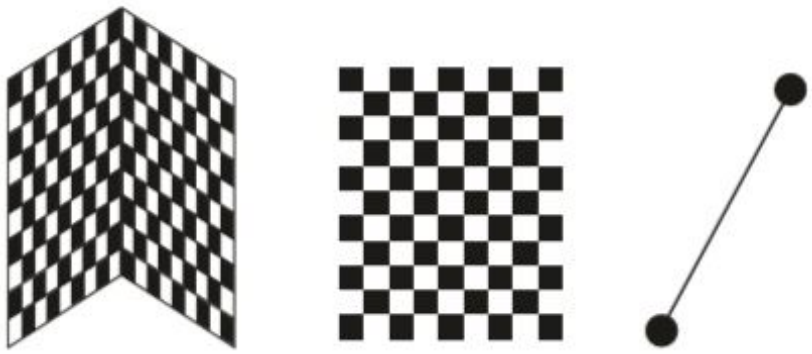


Abbildung 2.13: Möglichkeiten zur überwachten Kalibrierung von Kameras: 3D (links), 2D (Mitte) und 1D (rechts) [31]

Eine weitere Unterteilung der Kalibrierungsverfahren erfolgt in geometrische und fotometrische Kalibrierung. Als geometrische Kalibrierung wird die Erfassung der intrinsischer sowie extrinsischer Kameraparameter verstanden. Unter intrinsischen Parametern werden die Parameter der Matrix K (siehe auch Gleichung 2.13) sowie die Verzerrungsparameter (siehe Gleichung 2.4) verstanden. Die extrinsischen Parameter bezeichnen die Rotation und die Translation des Koordinatenursprungs der Kamera zu dem Weltkoordinatensystem (siehe auch Kapitel 2.3). Fotometrische Kalibrierung bezieht sich auf die Ermittlung des Zusammenhangs zwischen dem emittierten Licht der Szene und der Helligkeit des entsprechenden Pixels. Dieser Zusammenhang hängt von mehreren Faktoren ab, wie Materialeigenschaften der Szene, Lichtbedingungen etc. [29][30][31].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 2.3 Verarbeitung visueller Daten

Im Allgemeinen wird eine Punktwolke als ein Datenset von Punkten in einem mehrdimensionalen Koordinatensystem ($1, \ldots, K$ Dimensionen) definiert. Per Definition muss eine dreidimensionale Punktwolke $P$ eines Objektes mit Punkten $p_i$ folgende Kriterien erfüllen [36]:

- $\forall p_i \in P: \quad \forall p_i \in \mathbb{R}^3 \quad i = 1, 2, ...., M$
- Das zu untersuchende Objekt ist in der konvexen Hülle der Punkte

Erfüllt die Punktwolke das zweite Kriterium nicht, so wird sie oft auch als eine 2.5-dimensionale Punktwolke bezeichnet [36].

Punktwolken bestehen im Allgemeinen aus $K$-dimensionalen Tupeln, die Informationen zu jedem Punkt beinhalten. Somit weist eine Punktwolke die Struktur einer $M$ x $K$ Matrix. Weiters wird zwischen strukturierten und unstrukturierten Punktwolken unterschieden. Eine strukturierte Punktwolke besitzt die Eigenschaft der Strukturiertheit, d.h. eine formalisierte, den Koordinaten des Sensors entsprechende Indizierung. Demnach lässt sich die Struktur einer strukturierten Punktwolke mittels der in Grafik 2.8 präsentierten Koordinaten des Sensors durch $U$ x $V$ x $K$ beschreiben. Hingegen besitzt eine unstrukturierte Punktwolke keine formalisierte Struktur [36].

Die geometrische Interpretation von Punktwolken basiert auf der Verwendung dreidimensionaler Koordinaten. Zur dreidimensionalen Manipulation der Punkte und deren Koordinaten besteht eine Vielzahl von Operationen, eine Übersicht der relevanten Operationen wird in Tabelle 2.2 abgebildet [38]. Für die detaillierte Vorstellung der einzelnen Transformationen wird auf den Anhang verwiesen.

| ***Transformation*** | ***Matrix*** | ***Freiheitsgrade*** | ***Bleibt erhalten*** |
|---|---|---|---|
| Translation | $[I\|t]_{3x4}$ | 3 | Orientierung |
| Starre Transformation | $[R\|t]_{3x4}$ | 6 | Länge |
| Ähnlichkeits-transformation | $[s \cdot R\|t]_{3x4}$ | 7 | Winkel |
| Affine Transformation | $[A]_{3x4}$ | 12 | Parallelität |

Tabelle 2.2: Mathematische Operation zur Manipulation von dreidimensionalen Koordinaten [38]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Zur Ermittlung und Beschreibung geometrischer Informationen wie etwa Oberflächennormalen, Keypoints oder Deskriptoren (siehe auch Kapitel 2.3.1) werden statistische Methoden benötigt. Dazu zählen insbesondere die Berechnung der Kovarianz sowie der Eigenwerte und Eigenvektoren der Punktwolke, die zur Bestimmung räumlicher Verteilung der Punkte verwendet werden können. Die detaillierte Vorstellung dieser Methoden erfolgt im Anhang dieser Arbeit.

### 2.3.1 Merkmale der 3D-Computer Vision

Im Bereich der Computer Vision werden Daten durch Features beschrieben. Features sind interessante Merkmale der Daten, geltend sowohl für zwei- als auch dreidimensionale Daten, die einen deskriptiven bzw. informativen Charakter haben. In Hinsicht auf 3D-Daten sind geometrische Informationen von hoher Wichtigkeit. Eine Ermittlung dieser Daten basiert auf der räumlichen Anordnung der Punkte innerhalb einer bestimmten, meist vorher definierten Nachbarschaft. Als eine Nachbarschaft wird dabei in der Regel eine bestimmte Anzahl der nächsten Punkte, oder eine bestimmte geometrische Region rund um den betrachteten Punkt (z.B. Sphäre oder Zylinder), bezeichnet [39][40].

Punkte einer Punktwolke stellen stichprobenartig erfasste Punkte einer realen Oberfläche, die weiter durch Rauschen behaftet wurden, wodurch die Information über die Krümmung der Oberfläche verloren geht. Demnach zählen Oberflächennormalen (siehe auch Abbildung 2.14) zu den wichtigsten geometrischen Features einer Punktwolke und dienen zur Wiederherstellung der Information über die lokale Krümmung. Zur Berechnung der Normalen werden grundsätzlich zwei verschiedene Arten der Methoden verwendet. Die optimierungsbasierten Verfahren schätzen die Oberflächennormalen durch die Ermittlung der Minima einer mathematischen Funktion. Diese Funktion nimmt an, die Punkte der Nachbarschaft können entweder linear (d.h. planar) oder quadratisch angenähert werden. Die zweite Art zur Schätzung der Oberflächennormalen stellen die Mittelwertbildungsmethoden dar. Diese bestehen im ersten Schritt aus Bildung lokaler Dreiecke, die durch den zu untersuchenden Punkt und zwei Nachbarpunkte geformt werden. Zunächst wird für jedes Dreieck sowohl die Oberflächennormale, als auch ein Gewichtungsfaktor ermittelt. Anschließend wird die Oberflächennormale für den zu untersuchenden Punkt als ein gewichteter Mittelwert aller Oberflächennormalen der zugehörigen Dreiecke berechnet [41][42].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

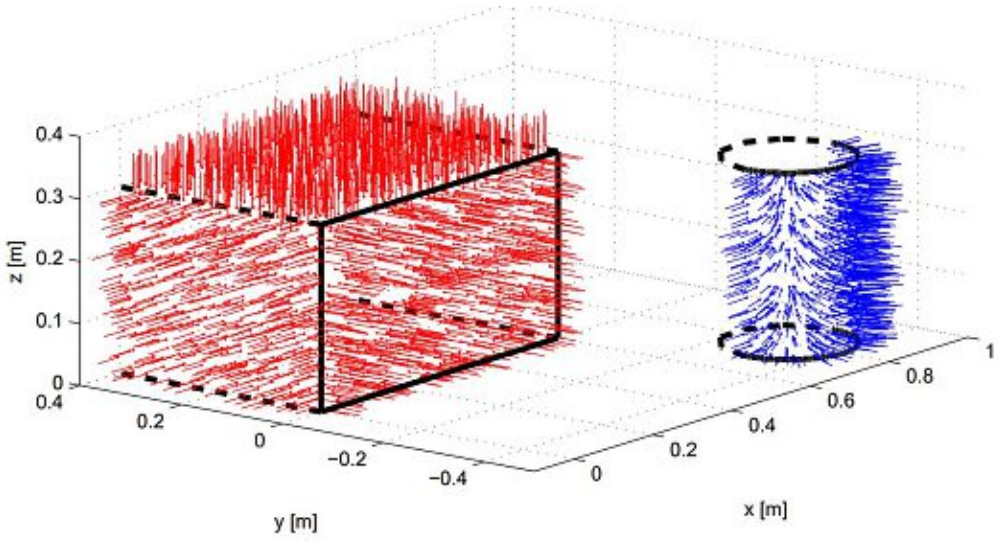



Abbildung 2.14: Darstellung der Oberflächennormalen anhand einer planaren Box und eines Zylinders [41]

3D-Deskriptoren sind Features, die geometrische Struktur einer Punktwolke invariant gegenüber einer starren Transformation (siehe auch Tabelle 2.2) beschreiben sollen. 3D-Deskriptoren werden in lokale, globale und hybride Deskriptoren unterteilt [43].

Lokale Deskriptoren beschreiben Features mit Rücksicht auf ihre lokale Nachbarschaft rund um einen Keypoint. Als Keypoints werden Punkte bezeichnet, die besondere Charakteristiken besitzen, wie z.B. Kanten oder eine Ecke. Die Ermittlung von dreidimensionalen Keypoints basiert auf den geometrischen Informationen, sowie sonstigen Informationen wie Farben- oder Intensitätskontrast. Zu den bekanntesten Algorithmen zur Ermittlung von Keypoints zählen NARF, 3D Harris und 3D SIFT [44][45][46].

Lokale Deskriptoren sind gegenüber globalen Deskriptoren in der Regel robuster hinsichtlich Okklusionen und Rauschen, jedoch sind sie sensitiv auf die Änderungen innerhalb der betrachteten Nachbarschaft um den Keypoint. Hingegen ermitteln globale Deskriptoren in der Regel einen Vektor, der die geometrischen Eigenschaften der gesamten Punktwolke samt der oben erwähnten Einschlüsse beschreibt. Daher werden oft hybride Deskriptoren verwendet, die aus Kombination aus einem lokalen und einem globalen Deskriptor bestehen. Im Folgenden werden einige ausgewählte lokale und globale Deskriptoren und deren Funktionsprinzipien kurz beschrieben [43].

**Lokale Deskriptoren**:

- **Spin Image** wandelt die 3D-Oberfläche repräsentiert durch die Punkte der Punktwolke in ein neues Set zweidimensionaler Bilder durch eine mathematische Transformation. Anhand der berechneten Abstände zwischen einem Keypoint und seinen Nachbarpunkten sowie deren Oberflächennormalen können für jeden Nachbarpunkt zweidimensionale zylindrische Koordinaten berechnet werden und anschließend gespeichert werden [47].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- **Eigenwert-basierte Deskriptoren** verwenden die Kovarianz lokaler Nachbarschaft rund um den untersuchten Punkt. Aus der Kovarianz werden die Eigenwerte ermittelt, die anschließend absteigend gereiht werden ($\lambda_1 > \lambda_2 > \lambda_3$). Diese werden dann zur Beschreibung der Point-ness („Pünktlichkeit“), die dem niedrigsten Eigenwert entspricht, der Curve-ness („Kurvigkeit“), die durch die Subtraktion von $\lambda_1$ und $\lambda_2$ ermittelt wird und Surface-ness („Flächigkeit“), die mit $\lambda_2 - \lambda_3$ berechnet wird [43].

- **Point-Feature-Histogramm (PFH)** verwendet die Beziehungen zwischen den Koordinaten aller Punkte innerhalb der zu untersuchenden Nachbarschaft, sowie ihrer Oberflächennormalen zur Konstruktion eines Darboux-Rahmens. Dieser liefert drei aufeinander rechtwinklige Koordinaten, mit deren Hilfe unter Einbeziehung der Oberflächennormalen der benachbarten Punkte drei Winkel berechnet werden, die schließlich den Deskriptor bilden [45].

- **Signature of Histogram of Orientation (SHOT)** legt ein Koordinatensystem entsprechend der Eigenwerte, die aus der Kovarianzmatrix der Nachbarschaft berechneten werden. Anschließend wird die Nachbarschaft anhand des Koordinatensystems in isotrope sphärische Zellen unterteilt. Für jede Zelle wird ein Histogramm gebildet, das aus Winkeln, die durch das Skalarprodukt der Oberflächennormale des Keypoints und der Oberflächennormale jeden Punktes der Nachbarschaft berechnet werden, besteht. Diese werden danach normalisiert, um die Dichte der Punkte berücksichtigen zu können [43][45].

**Globale Deskriptoren**:

- **Viewpoint Feature Histogramm (VFH)** basiert auf einem ähnlichen Prinzip wie PFH, erweitert um die Einbeziehung des Blickwinkels (Viewpoints). Dies erfolgt durch die Berechnung der Winkel der Oberflächennormalen der gesamten Punktwolke ausgehend von einem bestimmten Blickpunkt, die anschließend in einem Histogramm gespeichert werden. Die Einbeziehung des Blickwinkels wirkt sich dabei besonders positiv auf die Leistung und die Fähigkeit, die Geometrie zu beschreiben, aus [45].

- **Global Structure Histogramm (GSH)** berechnet für jeden Punkt einen lokalen Deskriptor. Anhand der Eigenschaften werden Punkte zu Clustern, die die Geometrie der Oberfläche charakterisieren, zugewiesen. Eine anschließende Berechnung der Beziehungen zwischen den entstandenen Clustern, basierend auf den Abständen zu dem jeweiligen Cluster, wird gefolgt durch eine Bildung eines Histogramms, was in GSH resultiert [45][48].

- **Global Ortographic Object Descriptor (GOOD)** erfordert die Auslegung eines globalen Koordinatensystems basierend auf Principal Component Analysis (PCA, siehe auch Anhang der Arbeit). Ausgehend von dem neuen Koordinatensystem werden zwischen den Achsen des Koordinatensystems drei Ebenen gelegt, auf die die Punkte der Punktwolke orthogonal projiziert werden. Danach erfolgt



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

eine Aufteilung der Ebenen in mehrere Bins, für die zwei statistische Features berechnet werden. Verschmelzung der zwei Features in einen Vektor formt letztendlich den Deskriptor [45].

### 2.3.2 Machine Learning in Computer Vision

Machine Learning wird definiert als ein Prozess der Implementierung von Lernfähigkeit in informatische Systeme. Dies impliziert, dass das Verhalten eines durch Machine Learning entstandenen Computerprogrammes nicht explizit durch den Autor des Programms definiert ist. Das Verhalten eines solchen Counterprogrammes wird bestimmt durch drei Faktoren – von den Daten, die zum Eintrainieren verwendet wurden, von der Funktion, die die Abweichung zwischen dem aktuellen und dem gewünschten Verhalten quantifiziert, sowie durch eine Feedback-Funktion, die anhand der Abweichung den Learning-Prozess steuert und die Güte verbessert [49][50].

Die Unterteilung von Machine Learning erfolgt in drei Gruppen nach der Art des Lernens: überwachtes Lernen, unüberwachtes Lernen und Reinforcement Learning. Das überwachte Lernen verwendet einen bestehenden Datensatz mit Inputs und richtig gelabelten Outputs. Dadurch kann sich das Programm prädiktiv verhalten und Outputs vorhersagen. Typische Vertreter des überwachten Lernens sind Regression und Klassifikation. Unüberwachtes Lernen verwendet hingegen keine gelabelten Daten, kann allerdings die Daten deskriptiv beschreiben, z.B. in Form von Clustering. Separat dazu wird Reinforcement Learning betrachtet. Reinforcement Learning verwendet nicht-gelabelte Datensätze, erhält trotzdem aber Feedback durch die Umgebung während des Prozesses [49][50].

Deep Learning (weiters als DL bezeichnet) zählt zu den Ansätzen von Machine Learning und basiert auf künstlichen neuronalen Netzen. Diese nähern durch ihren Aufbau, der durch eine Vielzahl von mit sich interagierenden Neuronen geprägt ist, Funktionsweise des menschlichen Gehirns. Dabei führt jedes Neuron eine Operation aus, und durch die Interaktion mehrerer Neuronen können komplexe Entscheidungen entstehen [40].

In der letzten Dekade stieg der Einsatz von Deep Learning zur Analyse und Verarbeitung visueller Daten zunehmend an [40]. Als Antriebskraft für diesen Zuwachs gelten einerseits die Verfügbarkeit der Datensätze, andererseits die Entwicklungen auf dem Gebiet des GPU-basierten parallelen Rechnens, wodurch das Eintrainieren im Vergleich zum CPU-basierten Ansatz deutlich beschleunigt wurde [51][52].

Deep Learning unterscheidet sich demnach von einem traditionellen algorithmischen Ansatz, der auf einer manueller Ermittlung von Features und Deskriptoren basiert, die oft durch eine flache Klassifizierung der Features gefolgt wird (siehe auch 2.15). Dieser Ansatz ist in der Regel anwendungsspezifisch, erfordert ein gewisses Niveau fachlicher Expertise und eine aufwendige Feinjustierung der Algorithmen. DL bietet dagegen eine End-to-End-Lösung, die aus dem vorgegebenen Datensatz selbst Features erkennen und auswählen kann, und diese in ein Muster umsetzt und anhand dessen gewünschte



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Ergebnisse liefert. Hierbei erreicht DL oft eine höhere Präzision, da durch die deskriptive Analyse des Datensatzes während des Eintrainierens auch Einflüsse bzw. Features, die durch den Menschen sonst vernachlässigt würden, mitberücksichtigt werden. Ein Nachteil des DL-Ansatzes ist allerdings, dass die Features, die durch das Eintrainieren extrahiert wurden, sehr objektspezifisch sind und nur dem Inhalt des Datensatzes entsprechen. Dies impliziert, dass das Modell bei ungeeignet gewählten Daten für den Datensatz eine mangelhafte Leistung für Objekte, die nur gering von den Daten des Datensatzes abweichen, liefert. Generalisierung ist somit nur mit einem großen und vielfältigen Datensatz möglich, wodurch allerdings die Kosten des Eintrainierens in Form von der Zeit deutlich steigen [40][52].

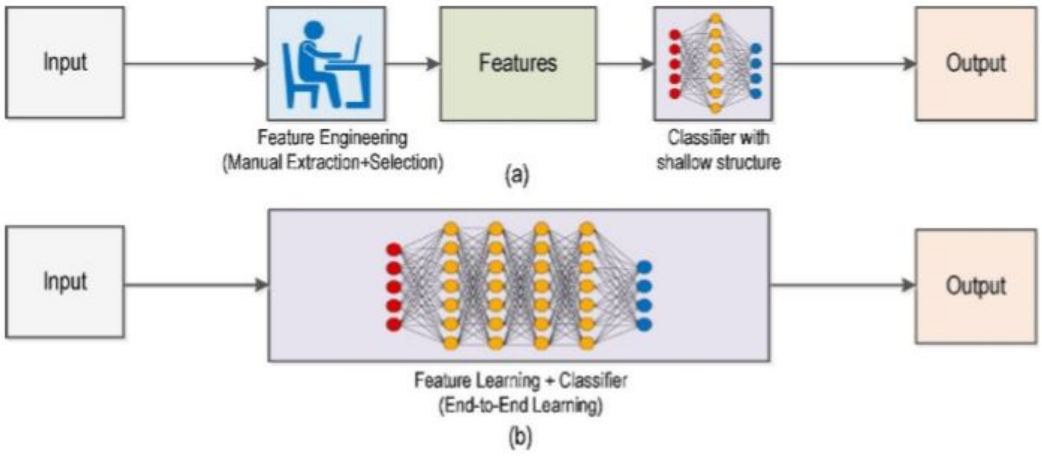



Abbildung 2.15: Veranschaulichung der unterschiedlichen Ansätze: a) traditioneller Ansatz b) Deep Leaning-Ansatz [40]

Im Hinblick auf 2D- sowie 3D-Vision werden insbesondere Convolutional Neural Networks verwendet (Deutsch: gefaltete neuronale Netzwerke). Convolutional Neural Networks (weiter als CNN bezeichnet) sind neuronale Netzwerke inspiriert durch die biologische Struktur des visuellen Prozesses und bestehen aus drei Arten der Layer – Convolution, Pooling und Fully Connected Layer [51].

Convolution Layer verwenden Kernels in Form einer Matrix zur Extraktion der Features. Für jede Position des Kernels wird elementarweise ein Produkt des Kernels und der diskreten Datenmatrix, die als Input gilt (z.B. die Pixel im Falle eines 2D-Bildes), berechnet und anschließend aufsummiert (siehe auch Abbildung 2.16). Das Ergebnis wird in eine Feature-Karte gespeichert, die für den jeweiligen Kernel gilt. Pro Modell können mehrere Kernel und somit Features-Karten angewendet werden. Der Output wird anschließend durch eine nicht-lineare Aktivierungsfunktion transformiert, die üblichste ist ReLU (Englisch: Rectified Linear Unit). Zur Anwendung kommen auch Sigmoid oder Tangens hyperbolicus [51][53].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

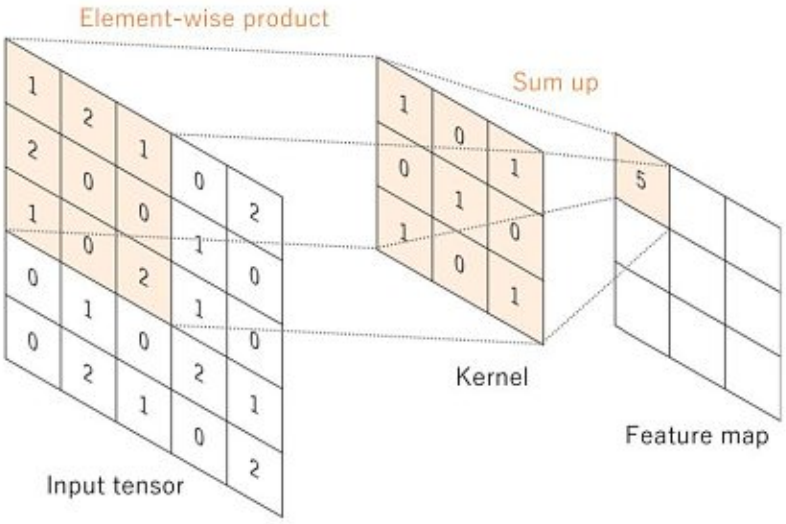



Abbildung 2.16: Veranschaulichung des Prinzips der Convolution [53]

Pooling Layer dienen zur Reduktion der Dimensionalität der Daten. Dadurch kann eine Konvergenz des Modells schneller erreicht werden. Ebenfalls werden dadurch eine bessere Invarianz gegenüber kleinen Veränderungen wie Verschiebungen und Verzerrungen der Daten erreicht. Die Reduktion der Dimensionalität erfolgt durch Downsampling, wobei zwei übliche Strategien angewendet werden: Max Pooling extrahiert für jeden Patch, d.h. einen Subset der Feature-Karte, den maximalen Wert, dagegen ermittelt Average Pooling den Mittelwert für jeden Patch [51][53].

Als letztes kommen Fully Connected Layer zur Anwendung. In diesem Layer sind alle Input-Neuronen mit allen Output-Neuronen durch Gewichte, die durch das Eintrainieren bestimmt werden, verbunden. Nach dem Durchgang durch alle Layer wird die letzte Feature-Karte üblicherweise in ein eindimensionales Array übersetzt. Der letzte Layer besitzt in der Regel eine unterschiedliche Aktivierungsfunktion, für Zweck der Klassifizierung werden die Sigmoid-Funktion oder die Softmax-Funktion eingesetzt. Die Ergebnisse enthalten dadurch Werte zwischen Null und Eins, wodurch die Wahrscheinlichkeit der Zugehörigkeit zu einer Klasse bestimmt werden kann [51][53].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# 3 Tracking und Ermittlung der Pose starrer Körper

## 3.1 Einführung

Die Bestimmung der Position und der Orientierung starrer Körper erfolgt über die Bestimmung aller ihrer sechs Freiheitsgrade. Zu diesen zählen drei translatorische sowie drei rotatorische Parameter. Daraus abgeleitet wird das Problem der Positionsermittlung und der Orientierung als 6D Pose bezeichnet. Unter der Annahme, dass ein Modell für das zu bestimmende Objekt verfügbar ist, können die Punktwolke der Szene mit der Punktwolke des Modells in eine Überlagerung gebracht werden, wodurch schließlich die sechs Freiheitsgrade in Form einer starren Transformation (siehe auch Kapitel 2.3) in einem gemeinsamen Koordinatensystem bestimmt werden. Die Ermittlung aller der Freiheitsgrade erfolgt entweder anhand zweidimensionaler Daten (2D-3D) oder anhand dreidimensionaler Daten (3D-3D) [29].

Die Methoden zur Erkennung der 6D-Pose der Objekte können grundsätzlich wie folgt aufgeteilt [54]:

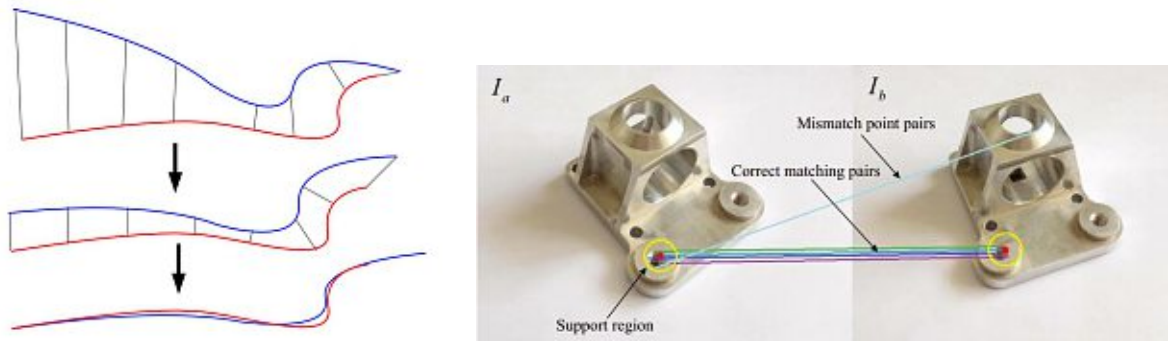
- Template-basierter Ansatz verwendet ein abgeleitetes Modell des starren Objektes, bezeichnet auch als Template, welches über die Szene gescannt wird. Üblicherweise werden die Kanten des Modells als ein Template verwendet. Die Erkennung basiert auf der Bildung einer abstandsbasierten Funktion, die beispielsweise die Distanz der Ecken des Template mit den Punkten der Szene berechnet. Durch die Minimierung der abstandsbasierten Funktion wird die beste Annäherung erreicht. Dieser Ansatz weist allerdings den Nachteil auf, stark durch Clutter (Unordnung der Objekte in der Szene) und Okklusionen beeinflusst zu sein.

- Sparse-Features-Ansatz basiert im ersten Schritt auf der Extraktion der Merkmale in Form von Keypoints, für die anschließend lokale Deskriptoren berechnet werden. Die Berechnung erfolgt sowohl für das Modell als auch für die Szene, um anschließend Matching zwischen den Deskriptoren der Szene und deren des Modells durchzuführen. Die Einschränkung dieser Methode besteht in der Erforderlichkeit gewisser diskriminierender Merkmale, z.B. in Form von Textur, um Aussagekraft der Deskriptoren sicherzustellen.

- Dense-Approach baut auf der Annahme auf, dass die Erkennung der Objekte durch eine statistische Berechnung der Zugehörigkeit zu einem Objekt über die kleinste Einheit (z.B. ein Pixel oder ein Punkt einer Punktwolke) her möglich ist. Dense-Approach verwendet daher eine Funktion, die die Wahrscheinlichkeit der globalen Zugehörigkeit jeder Einheit der Szene zu einem Modell bestimmt. Dies resultiert in einen großen mathematischen Raum der Lösungen, auf dem die Suche der Zugehörigkeit erfolgt, was schließlich den Nachteil dieser Methode bildet.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN




(a) (b)

Abbildung 3.1: a) Veranschaulichung des Prinzips von Template-Matching als Minimierung des Abstandes zwei Objekte (anhand ICP [55]), b) Veranschaulichung des Prinzips von Sparse-Features-Matching anhand 2D Daten [56]

Zur Schätzung der 3D-Pose werden traditionelle Ansätze, die durch den Vergleich der selbst-programmierten Features bzw. Deskriptoren zwischen dem Modell und der Szene die Pose ermitteln, sowie Machine Learning verwendet. Der Machine Learning-Ansatz wird zur Erkennung der Pose aus Punktwolken oft in Kombination mit zweidimensionalen Methoden verwendet. Viele Methoden bauen auf einem zweidimensionalen Machine Learning-Detektor auf (üblich sind R-CNN, Fast R-CNN, Faster R-CNN, YOLO, SSD, Decision Forrests), mit deren Hilfe sogenannte Bounding Boxes (Englisch für Begrenzungsrahmen) ermittelt werden. Eine anschließende Projizierung dieser Boxen in den dreidimensionalen Raum wird gefolgt durch eine lokale Ermittlung der Pose über den Vergleich der dreidimensionalen Daten der Szene und des Modells [57].

Zu Methoden, die die Pose mittels Machine Learning ermitteln, gehören Klassifikation, Regression und kombinierte Klassifikation und Regression. Klassifikation verwendet eine Wahrscheinlichkeitsfunktion, die die Zugehörigkeit zu einer der prätrainierten diskreten Klassen ermittelt. Die Klassen bestehen in diesem Fall aus verschiedenen Posen desselben Objektes. Die Ermittlung der sechs Freiheitsgrade basiert daher auf dem Matching der Klassen, wodurch dem untersuchten Bauteil die Freiheitsgrade der Klasse verliehen werden. Bei grober Aufteilung in Klassen, und dadurch einer groben Posenschätzung, kann die wahre Pose weiter durch Pose-Refinement verfeinert werden. Umgekehrt wird bei Regression eine mathematische Funktion durch das Einlernen mit Daten ermittelt. Durch Anwendung dieser Funktion wird die 6D-Pose aus der untersuchten Szene kontinuierlich angenähert. Viele Regressionsalgorithmen ermitteln im ersten Schritt die grobe Schätzung durch eine Bounding Box, gefolgt durch eine lokale Regression. Andere extrahieren in erster Linie Features aus der Szene, die anschließend durch Regression gematcht werden. Einige Verfahren vereinheitlichen sowohl Klassifikation als auch Regression zur Schätzung der 6D-Pose. Hierbei wird die Architektur entweder getrennt gehalten und die Methoden in zwei Schritte klar separiert, oder werden beide Methoden in eine gemeinsame Architektur integriert [57].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Punktwolken besitzen im Vergleich zu 2D-Daten eine geringere Informationsdichte und eine unregelmäßige Datendistribution. Zudem weisen sie oft keine Strukturiertheit auf, weshalb direktes Erlernen der 3D-Features mittels neuronalen Netzwerken lange eine sehr schlecht bewältigbare Herausforderung darstellte. In den letzten Jahren wurde allerdings auf diesem Gebiet ein Durchbruch geschafft. Deep Learning-Methoden für Punktwolken werden in Ansätze ohne Verwendung von Korrespondenzen und Korrespondenz-basierte Verfahren unterteilt. Ansätze, die keine Korrespondenzen bilden, setzen auf Beschreibung durch einen globalen Deskriptor und eine anschließende Regression der Paramater, über die schließlich die Pose bestimmt wird. Diese Algorithmen sind am besten geeignet für die Erkennung von Objekten einer übereinstimmenden Geometrie, die ausschließlich durch ihre Pose unterschiedlich sind. Je abweichender die Geometrie, etwa durch Ausreißer, Rauschen oder Clutter, desto abhängiger ist das Verfahren von der deskriptiven Fähigkeit des Netzes. Korrespondenz-basierte Verfahren beschreiben die Punktwolke zuerst durch lokale Features. Dies wird gefolgt durch die Suche der Korrespondenzen zwischen den Features des Modells mit den Features der Szene [58][59].

## 3.2 Vorstellung ausgewählter Methoden

Aufbauend auf der präsentierten Unterteilung der Verfahren zur Ermittlung der 6D-Pose aus Punktwolken erfolgt nun die Beschreibung ausgewählter Methoden, die für den Use-Case in Betracht kommen.

Vock et al. haben in ihrer Arbeit eine Methode zum schnellen Template Matching vorgestellt. Hierbei werden in erster Linie die Konturen des Objektes sowie der Szene ermittelt und aus den Konturen Point-Pair-Features extrahiert und in einer Hash Tabelle gespeichert (siehe auch Kapitel 5.7). Die 6D-Pose wird anschließend direkt durch die Berechnung einer passenden Transformation über die korrespondierenden Punktpaare der Szene und des Modells ermittelt. Im Anschluss daran wird die Hypothese durch einen passenden Algorithmus verifiziert. Die Methode bietet sich insbesondere für Umgebungen an, die durch Rauschen behaftet sind, sowie bei sehr niedrigen Dichte der Punktwolken, da bei 10% der ursprünglichen Dichte ungefähr eine Hälfte der erkannten Pose korrekt ermittelt werden kann [60].

Radkowski hat ebenfalls Point-Pair-Features bei dem Entwurf eines Tracking-Algorithmus zur Verwendung für ein AR-basiertes Assistenzsystem für die Montage gewählt. Die Architektur des Systems ist in zwei Schritte aufteilen. Als erstes erfolgt die Initialisierung über die Erfassung der Anfangspose mittels Point-Pair-Features und Bildung der Korrespondenzen. Anschließend erfolgt eine Schleife von ICP (Iterative Closest Point, siehe auch Kapitel 5.7). ICP wird dabei einerseits zur Verfeinerung der 6D-Pose verwendet, andererseits auch zur Ermittlung in den nächsten Aufnahmen nach der Initialisierung. Demnach lässt sich das System als eines mit Template-Matching klassifizieren. Dabei wird allerdings vorausgesetzt, dass sich die Pose zwischen den einzelnen Aufnahmen nicht viel ändert. In diesem Fall versagt der ICP-Algorithmus und es erfordert eine nochmalige Initialisierung [61].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Wang et al. haben bei der Entwicklung eines AR-Assistenzsystems ebenfalls ICP angewendet. Im Vergleich zu Radkowski setzen sie allerdings auf einen Sparse-Features Ansatz und verwenden hierfür 2D-Farbdaten zur Schätzung entsprechender Punkte der Punktwolke, die anschließend mittels ICP in eine korrespondierende Pose gebracht werden. Als erste werden 2D lokale Features ermittelt, und mit den Features der vorherigen Aufnahme verglichen (siehe auch Grafik 3.1). Die Pixel-Koordinaten der passendsten Korrespondenzen werden zunächst gespeichert, und durch die Strukturiertheit der Punktwolke können anschließend die korrespondierenden Punkte über die gespeicherten Indizes abgerufen werden. Hierdurch liegt eine initiale Schätzung vor, die im Anschluss daran über ICP verfeinert wird. Diese Methode zeichnet sich insbesondere durch ihren geringen Computing-Bedarf aus, was zu sehr schnellem Tracking führt [56].

Almhazi et al. untersuchen verschiedene hybride Deskriptoren zur Erkennung der Pose der Objekte. Sie verdeutlichen die Unterschiede der zwei Ansätze – Erkennung über lokale Deskriptoren (etwa Sparse-Features) und über globale Deskriptoren (etwa der Dense-Approach), die neben den im Kapitel 2.3.1 bereits genannten auch die Segmentierung im Falle des dichten Ansatzes beinhalten (siehe auch Grafik 3.2). Aus ihrer Untersuchung zeigt sich VFH am geeignetsten, bedingt insbesondere durch die gute Performance, wobei sich die Kombination aus Fast Point Feature Histogramme (FPFH) als den lokalen Deskriptor, sowie Viewpoint Feature Histogramme (VFH) als den globalen Deskriptor am schnellsten aufweist [62].

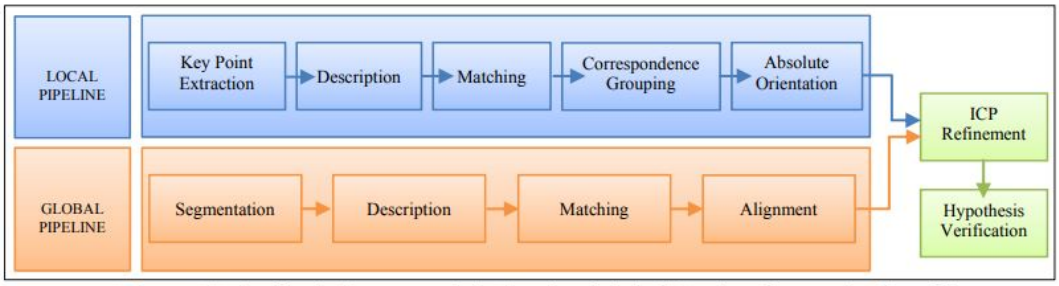



Abbildung 3.2: Unterschiede in den Pipelines bei Verwendung lokaler Deskriptoren (oberes Abbild) und globaler Deskriptoren (unteres Abbild) [62]

Li et al. verwenden einen globalen Deskriptor zur Erkennung und Registrierung der Pose der Bauteile, weshalb sich dies als Dense-Approach klassifizieren lässt. Um die Sensitivität der globalen Deskriptoren gegenüber dem Rauschen und Okklusionen zu minimieren, legen sie ein Koordinatensystem basierend auf der PCA (Principal Component Analysis) auf, und teilen anschließend die Szene in vier gleichgroße Cluster. Hierauf wird ein angepasster VFH-Deskriptor für alle der vier Cluster gebildet. Die Verfeinerung der Pose erfolgt weiters über ICP. Dieser Ansatz führt zur höheren Erkennungsrate im Vergleich zum traditionellen VFH-Deskriptor [63].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Brachmann et al. verwenden Dense-Approach und Machine Learning mittels Random Forest zur Ermittlung der Pose aus RGB-D Daten. Hierzu wurden Entscheidungsbäume gebildet, die Daten an jedem der Knoten nach verschiedenen Kriterien bzw. Features selektiv klassifizieren. Somit wird jedem Punkt der Depth-Map beim Durchgang durch den Forest (Deutsch: Wald) je ein Feature-Vektor pro Baum zugeteilt, dessen Zellen die Blätter jedes Baumes darstellen. Durch den Vergleich der Vektoren der aufgenommenen Daten und der prätrainierten Klassen kann die Wahrscheinlichkeit der Zugehörigkeit zu einer Instanz der Klassen für jeden Punkt errechnet werden. Anschließend erfolgt die Ermittlung der sechs Freiheitsgrade über eine Minimierungsfunktion, die die Koordinaten des Modells und die Koordinaten der erkannten Instanz vergleicht [54].

Deng et al. verwenden Fast-RCNN zur Schätzung einer 2D-Bounding Box aus RGB-Daten sowie einer Depth-Map, die nachfolgend in eine 3D Box transformiert wird. Die Transformation aus dem zweidimensionalen Raum in den Dreidimensionalen erfolgt dabei über eine prätrainierte Größe der 3D-Bounding Box, und die pixelweise Zuordnung durch die Strukturiertheit der Punktwolke. Die Erfassung der Pose erfolgt über die Ermittlung der Orientierung der 3D-Bounding Box, die anschließend durch Regression der Bounding Box und der prätrainierten 2D Features erfolgt [64].

Im Bereich von Deep Learning direkt aus Punktwolken stellen PointNet und PointNet++ Pionierarbeit dar, da sie als Erste das Problem der Unstrukturiertheit gelöst haben. Die zwei Methoden unterscheiden sich dabei durch die Verwendung der Korrespondenzen (PointNet++) bzw. einer globalen Beschreibung (PointNet) [59].

Su et al. (PointNet) schlagen drei mögliche Strategien zum Trainieren neuronaler Netzwerke vor, die invariant gegenüber der Sequenz der Inputdaten sind und somit zur Bekämpfung der Unstrukturiertheit geeignet sind. Diese heißen [65]:

- Punktwolke auf die kanonische Ordnung zu bringen (Strukturiertheit schaffen)
- Verwendung von Rekurrenten Neuronalen Netzwerken (Netzwerke mit einer Rückkopplungsfunktion zwischen den Layern)
- Zusammenschmelzen gesamter Informationen der Punktwolken durch Verwendung einer symmetrischen Funktion

Da die erste Lösung hinsichtlich der räumlichen Anordnung der Punkte mathematisch nicht stabil ist [65], und die zweite nur für kleine Datengrößen Sequenz-invariant ist [66], wird bei PointNet die dritte Strategie angewendet. Das Prinzip von PointNet besteht aus zwei neuronalen Netzwerken (bezeichnet als MLP – Multi-Layer Perzeptron) und einer max-Pooling Funktion (siehe auch Grafik 3.3). Dabei ist die max-Pooling Funktion für die Sequenzinvarianz verantwortlich. Die Invarianz gegenüber räumlichen Transformationen, insbesondere Rotation, wird durch zwei neuronale Netzwerke gesichert, die durch Regression eine Transformation für die Inputdaten sowie die Features abschätzen. PointNet ermittelt Features in Form eines globalen Deskriptors, der als Vektor für die gesamte Punktwolke definiert ist. Dadurch wird eine Klassifikation der Punktwolken



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

ermöglicht. Weiters kann durch die Fusion des Vektors des globalen Deskriptors und den lokalen Features Segmentation mittels eines Netzwerkes durchgeführt werden. Die Robustheit von PointNet lässt sich durch die eingelernte Fähigkeit zur Extraktion der sogenannten kritischen Punkte erklären. Zu diesen zählen diejenigen Punkte der Punktwolke, die am meisten zu der Max-Pooled-Feature beitragen und somit die Punktwolke am besten beschreiben [65].

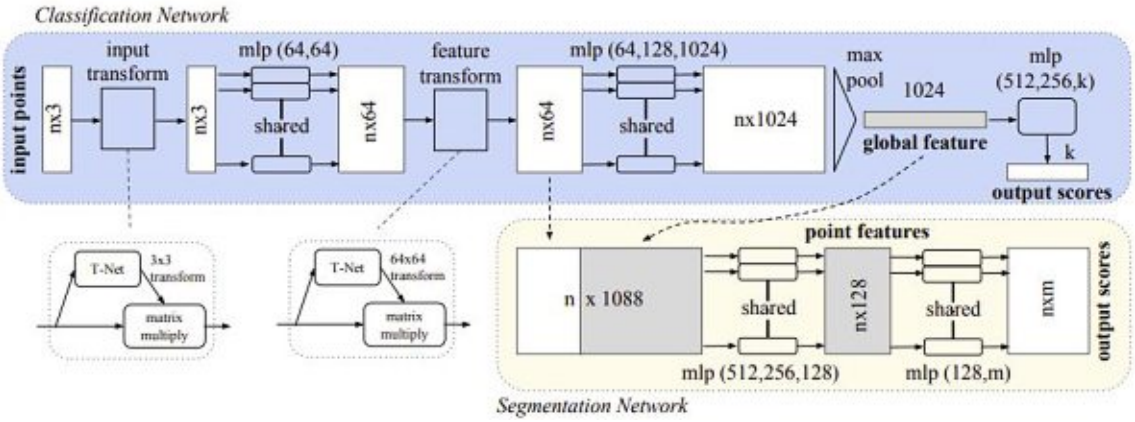



Abbildung 3.3: Architektur von PointNet [65]

PointNet++ ist eine Erweiterung von PointNet durch eine hierarchische Struktur. Hierbei wird die Architektur um Abstraktion durch Sampling und Grouping Layer erweitert (siehe auch Abbildung 3.4). Im ersten Schritt wird die Punktwolke abgetastet, wodurch Zentroide lokaler Regionen ermittelt werden (Sampling Layer). Im Anschluss daran werden Punkte rund um die Zentroide in fixe lokale Nachbarschaften gruppiert (Grouping Layer). Danach können, ähnlich wie bei einer üblichen CNN-Architektur, die lokalen Regionen durch einen PointNet-Layer in lokale Features umgewandelt werden. Dies verbessert die allgemeine Leistung des Netzwerkes [67].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

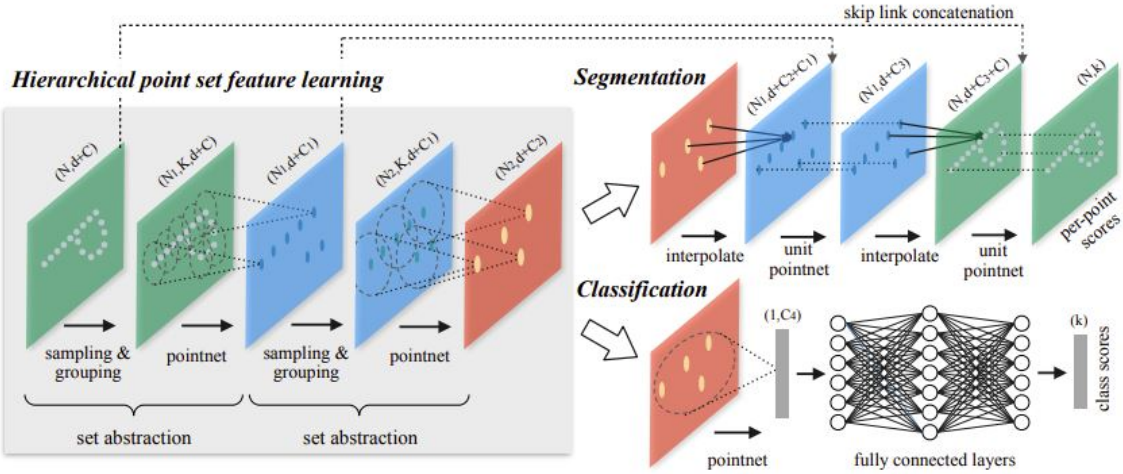



Abbildung 3.4: Veranschaulichung der Architektur von PointNet++ [67]

Obwohl PointNet und PointNet++ die 6D-Pose nicht direkt liefern, dient die Architektur als Kern vieler Verfahren zur Ermittlung der Pose starrer sowie artikulierter Objekte (siehe auch Kapitel 4.1). Zum Beispiel haben Hagelsjaer et al. PointNet zur Klassifikation der Hypothesen der Lage des Objektes in einer Szene angewendet. Sie verwenden anschließend die kritischen Punkte, geliefert durch PointNet, zur Segmentierung der benachbarten Punkte in der Umgebung des jeweiligen kritischen Punktes (siehe Grafik 3.5). Hierzu wurde eine PointNet-ähnliche Architektur gewählt. Wird die gleiche Segmentierung auch für das Modell durchgeführt, resultiert dies in 3D-3D segmentierte Korrespondenzen. Dies führt zu einer dramatischen Vereinfachung des Problems zur Ermittlung der 6D-Pose, die sich über gängige Algorithmen lösen lässt [68].

Abbildung 3.5: Veranschaulichung der segmentierter Umgebung rund um die kritischen Punkte [68]

Chen et al. haben in ihrem Werk auch PointNet zur Ermittlung der 6D-Pose aus RBG-D Daten verwendet. Hierzu wird zunächst die 2D-Lage der entsprechenden Instanz mittels 2D Detektors (YOLO V3) detektiert und die 2D-Bounding Box über die Strukturiertheit der Depth-Map in eine Punktwolke umgewandelt. Anschließend erfolgt eine Segmentierung der Instanz durch PointNet. Danach werden Punkte der Punktwolke



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

stichprobenartig ausgewählt und basierend auf den Intersektionen der Einheitsvektoren zwischen den stichprobenartig gewählten Punkten und alten Keypoints werden neue Hypothesen für Keypoints generiert (siehe auch Abbildung 3.6). Letztendlich werden Korrespondenzen der ermittelten Keypoints mit den Keypoints der Szene berechnet, und die passendste Hypothese ausgewählt [69].

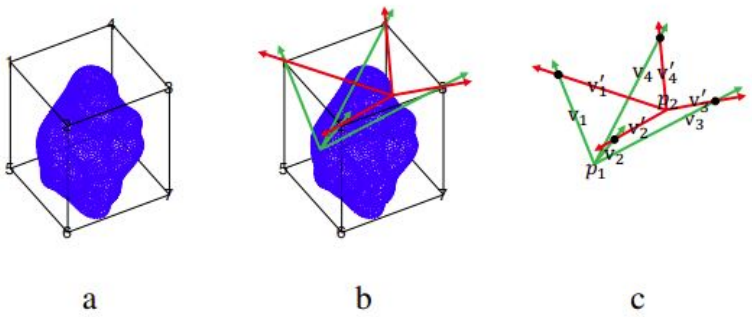



Abbildung 3.6: Veranschaulichung der Keypoints-Ermittlung (als Keypoints dienen in diesem Fall die Ecken der Bounding Box): a.) die generierte Bounding Box b.) Vektoren zwischen den Keypoints (Ecken der Bounding Box) und zwei zufälligen Punkten p1 und p2, c.) Intersektion der Vektoren als Basis für die neuen Keypoints [69]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# 4 Tracking und Ermittlung der Pose artikulierter Objekte

## 4.1 Einführung

Als ein artikuliertes Objekt wird ein Objekt bezeichnet, dessen Struktur sich aus mehreren starren Körpern zusammensetzt, die über Gelenke beweglich verbunden sind. Zu artikulierten Objekten zählen demnach Roboter sowie verschiedene Lebewesen, wenn die Annahme getroffen wird, dass Gewebe als ein starrer Körper angenommen wird (vgl. [70]). Artikulierte Objekte werden in der Regel durch eine Baumstruktur der kinematischen Kette beschrieben. Eine Darstellung hierarchischer Struktur der menschlichen Kette ist in Abbildung 4.1 ersichtlich (vgl. [71]).

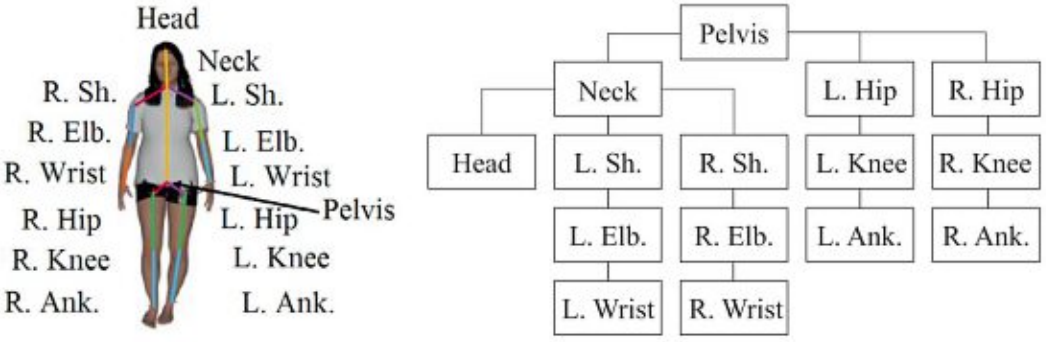



Abbildung 4.1: Veranschaulichung eines Modells der menschlichen kinematischen Kette [72]

Systeme zur Erfassung kinematischer Daten des menschlichen Körpers werden oft als Motion Capture Systeme bezeichnet. Die Taxonomie der Motion Capture Systeme wird in Tabelle 4.1 vorgestellt. Ein wesentlicher Nachteil der nicht-optischen Systeme ist, dass sie das Tragen der Sensorik voraussetzen. Zudem weisen viele der bestehenden Systeme keine akkurate Genauigkeit hinsichtlich des Trackings der absoluten Position für industrielle Anwendungen auf. Dies ist einerseits bedingt durch Einflüsse der Umgebung (z.B. Wechselwirkung zwischen den magnetischen Systemen und Metallobjekten) sowie durch akkumulierte Messfehler, die durch das passive Messen, d.h. Messen über die relative Änderung der Bewegung, besteht (dies betrifft unter anderem inertiale sowie mechanische Sensoren) (vgl. [6][73]).



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

| Nicht-optische Systeme | Optische Systeme |
|---|---|
| Inertiale Systeme | Marker-basierte Systeme |
| Magnetische Systeme | Markerlose Systeme |
| Mechanische Systeme | |

Tabelle 4.1: Unterteilung der Motion Capture Systeme (in Anlehung an [6][73])

Optische Systeme werden in Marker-basierte und markerlose Systeme unterteilt. Marker-basierte Systeme setzten das Tragen spezieller Marker voraus. Das Messprinzip besteht in der Emission infraroter Strahlung und eine anschließende Reflexion der Strahlung durch die Marker, die erfasst wird, wodurch sich die Position der Marker ermitteln lässt. Markerlose Systeme sind hingegen nicht-invasiv und verwenden Ansätze der Computer Vision zur Ermittlung der menschlichen Pose. Anforderungen an Vision-basierte Motion Capture Systeme bestehen in der Erfassung der Position und Orientation (Pose in dreidimensionalen Raum) des Menschlichen Körpers. Dies erfolgt über die Ermittlung der Freiheitsgrade aller Körperteile, wobei bei 15 Körpern der kinematischen Ketten (vergleiche mit Abbildung 4.1) besitzt der menschliche Körper 90 Freiheitsgrade (je sechs Freiheitsgrade pro einen starren Körper) [70].

Moeslund et al. unterscheiden in der Architektur verschiedener Motion Capture Systeme verallgemeinert zwischen vier Bestandteilen [74]:

- **Initialisierung** besteht unter anderem aus der Modellierung der kinematischen Kette, Bestimmung der anthropometrischen Grenzwerte für die Gelenke sowie Bestimmung der Form und Gestalt des Modells der kinematischen Kette.
- **Tracking** wird hier als Erkennung und Segmentierung des menschlichen Körpers aus den Daten definiert.
- **Ermittlung der Pose** ist der integrale Bestandteil von Motion Capture und besteht aus der Ermittlung der momentanen kinematischen Konfiguration des Menschen.
- **Erkennung** wird von Moeslung et al. als die Interpretierung der Daten zur Klassifikation von momentan durchgeführten Aktionen beschrieben (in der Regel im biometrischen Sinne, z.B. Laufen, Springen etc.) oder als die Erkennung spezifischer Parameter wie Geschlecht, Identität und andere.

Daraus abgeleitet gilt die Ermittlung der Pose als der Kern der Motion Capture. Erkennung der menschlichen Pose erfolgt entweder 2D-2D, 2D-3D oder 3D-3D (vgl. [29][72]). Im Allgemeinen können Ansätze zur Ermittlung der menschlichen Pose wie folgt klassifiziert werden [72]:



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Generativer Ansatz
- Diskriminierender Ansatz
- Hybride Ansätze

Der generative Ansatz basiert auf der Verwendung eines während der Initialisierung definierten Modells, mit dem die menschliche Pose angenähert werden kann. Dies besteht entweder aus dem gesamten Körper oder aus der Zusammensetzung einzelner Körperteile, deren Verlinkung bestimmt wird. Dementsprechend werden Modelle zur Ermittlung der Pose in Kontur-basierte (Modell des gesamten Körpers) sowie Skelettmodelle (Englisch: Skeletons) und volumetrische Modelle (Zusammensetzung mehrerer Einzelteile) unterteilt. Die Taxonomie der Modelle für Motion Capture sind in Abbildung 4.2 sichtbar [72][75][76].

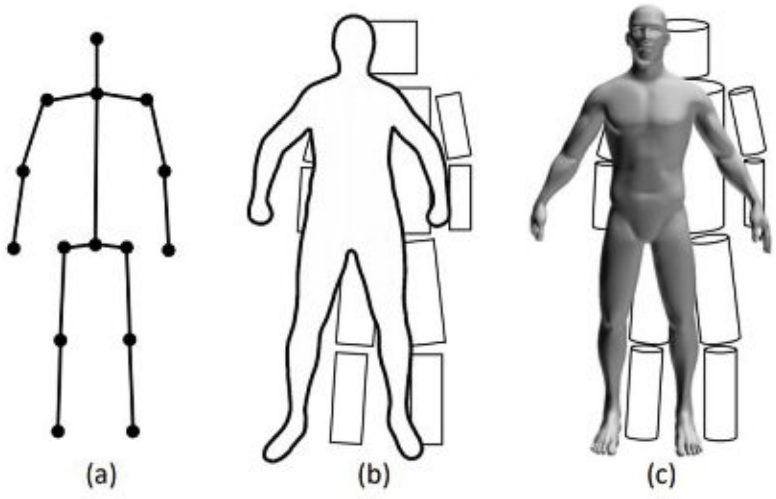



Abbildung 4.2: Modelle für generative Ansätze: a) Skeleton, b) Kontur-Modell, c) volumetrisches Modell [75]

Diskriminierende Ansätze verwenden kein Modell zur Ermittlung der Pose des menschlichen Körpers. Diese Systeme werden weiter in Learning-basierte Systeme, die ausgehend von den Daten durch eine erlernte Methode direkt die Pose bestimmen können, und Vorlagen-basierte Systeme, die zuerst die Daten mit einer gespeicherten Vorlagen matchen, um anschließend die Pose anhand der Deskriptoren der Vorlage zu interpolieren [72][75].

Hybride Ansätze bestehen aus beiden Methoden, wobei in der Regel im ersten Schritt die Pose diskriminierend ermittelt und anschließend durch den generativen Ansatz verfeinert wird [72].

Eine weitere verallgemeinerte Klassifikation der Methoden zur Ermittlung der Pose basiert auf der Unterscheidung anhand der Abstraktion. Der Top-down Ansatz führt



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

im ersten Schritt Tracking durch, wodurch die Lage der Person zuerst in Form einer groben Region der Daten angenähert wird. Im Anschluss daran wird die Pose für diese Person ermittelt. Umgekehrt werden bei dem Bottom-up Ansatz zunächst alle Teile des Körpers identifiziert, um sie danach einer Person zuzuordnen. Dies zeigt sich, besonders bei der Bestimmung der Pose mehrerer Personen aus denselben Daten, als die effizientere Variante, allerdings stellt die Zuteilung überlappender Körperteile bei dem Bottom-up Ansatz eine Herausforderung dar [75].

Ausgehend von den im Kapitel 2.3.1 präsentierten Grundlagen bieten sich sowohl der traditionelle Ansatz der Computer Vision, als auch der Deep Learning Ansatz zur Bestimmung der menschlichen Pose. Getrieben durch die Entwicklungen im Bereich der zweidimensionalen neuronalen Netzwerke, sowie durch die Verfügbarkeit geeigneter Daten, rückt in den letzten Jahren der Deep Learning Ansatz in den Vordergrund. Die Motivation dafür ist die Ausdrucksfähigkeit der erlernten Features im Vergleich zu den traditionellen Deskriptoren, wodurch die Performance der ursprünglichen Methoden überholt wurde [75][77].

Deep Learning-basierte Methoden werden grundsätzlich in regressionsbasierte und detektionsbasierte Algorithmen unterteilt. Regressionsbasierte Verfahren ermitteln die Koordinaten der Gelenke als Keypoints durch direkte kontinuierliche Regression aus den Features der Daten. Verfahren basierend auf Detektion erkennen dagegen in erster Linie die Körperteile über Patches (Sub-Sets bzw. Regionen der Daten) und produzieren diskrete Heatmaps, die die Lage der Keypoints bezeichnen [75][77].

Neben zweidimensionalen RGB-Daten können ebenfalls Tiefendaten sowie Punktwolken zur Ermittlung der Pose artikulierter Objekte mittels Deep Learning verwendet werden. Diese stoßen allerdings einerseits auf die Verfügbarkeit der Daten zum Eintrainieren der Modelle an der Seite der Depth-Maps (vgl. [78]) sowie das Problem der Unstrukturiertheit an der Seite der Punktwolken (siehe auch Kapitel 3.2) und die große Dimensionalität der Lösungen (vgl. [79]). Daher werden diese Methoden detaillierter im Kapitel 4.2 behandelt.

## 4.2 Vorstellung ausgewählter Methoden

Ausgehend von der Taxonomie der Ansätze zur Ermittlung der Pose für artikulierte Objekte erfolgt in diesem Kapitel die Beschreibung ausgewählter Methoden. Im Fokus stehen dabei Methoden, die zur Anwendung für den Use-Case geeignet sind.

Daubney et al. verwenden sparse-Features und ein Modell zur Erfassung der Pose aus RGB-Daten, wobei sich ihre Methode auf bestimmte zyklische Aktivitäten wie Joggen oder Spazieren beschränkt. Dies lässt sich somit als ein generativer Ansatz beschreiben, wobei das Prinzip ihrer Methode in Erfassung der Bewegung der ermittelten Features zwischen den einzelnen Aufnahmen eines Videos besteht. Hierzu werden im ersten Schritt zweidimensionale Features aus der Aufnahme bestimmt, um anschließend mittels eines



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Modelles die Korrespondenzen zwischen den Features und den zugehörigen Körperteilen zu ermitteln. Dafür wird ein statistisches Modell der modellierten Bewegung verwendet, das aus verketteten Gauß'schen Distributionen der erwarteten Verschiebungen der Körperteile besteht, die für die spezifische Bewegung signifikant sind (zur Veranschaulichung: eine Gauß'sche Verteilung der Bewegungsbreite der Beine wäre beim Joggen breiter als diejenige der Arme). Durch die Korrespondenzen der Features und der Körperteile kann aus der räumlichen Anordnung der Features im Bild die Lage der Gelenke mathematisch ermittelt werden [80].

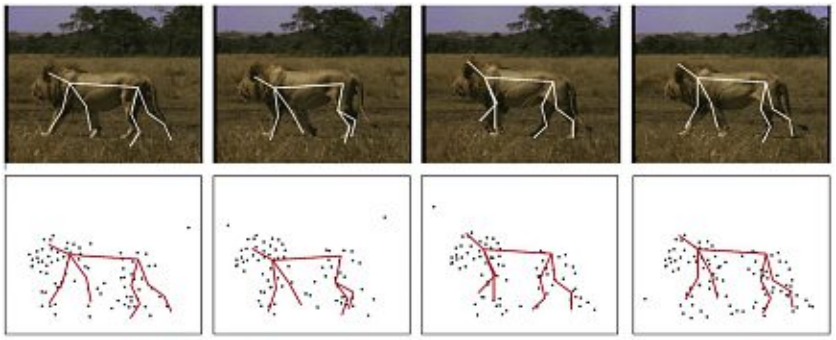


Abbildung 4.3: Ermittlung der Pose aus der Bewegung für einen Löwen: Darstellung der Features und der Keypoints [80]

Toshev und Szegedy leisteten Pionierarbeit im Bereich der Anwendung kaskadenartiger Deep Neural Networks (weiters: DNN) zur direkten Regression der Gelenke aus RGB-Bildern, was einen Durchbruch im Bereich der Learning-basierten Systemen verursachte. Hierzu wurde ein neuronales Netzwerk mit Bildern inklusive gelabelten Koordinaten der Gelenke eintrainiert. Die Architektur des Netzwerks besteht dabei aus fünf Convolution und Pooling Layern, gefolgt durch zwei Fully Connected Layer (siehe auch Grafik 4.4). Die hohe Genauigkeit der Methode verdankt der iterativen „Kaskade". Hierzu durchlaufen die Daten das Netzwerk mehrmals durch, wodurch die Regression mehrmals wiederholt wird. Gleichzeitig findet eine Verkleinerung des Bereichs des Interesses rund um die Features statt, wodurch die erste grobe Schätzung iterativ verfeinert wird [81].

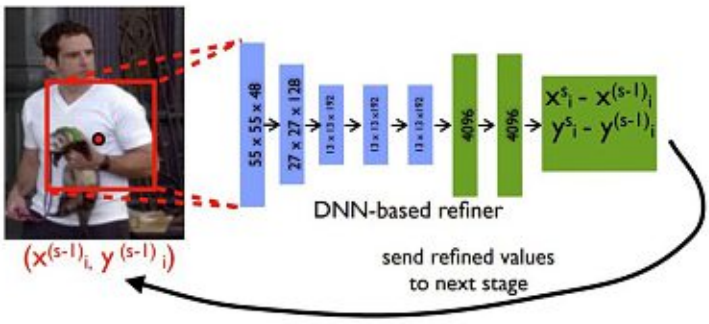



Abbildung 4.4: Architektur von DeepPose [81]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Pavlakos et al. verwenden ebenfalls Convolutional Neural Networks zur Ermittlung der 3D-Pose aus einzelnen RGB-Bildern. Zur Behebung der traditionellen Fehler der rein Learning-basierten Methoden, die sich vor allem an das Versagen bei Okklusionen richten, wird bei ihrer Methode ein Modell verwendet. Somit lässt sich das Verfahren der Klasse der hybriden Ansätze zuzuteilen. Das Verfahren verläuft in zwei Phasen ab. Als erstes wird ein CNN angewendet, das gleichzeitig eine Silhouette bzw. ein Schattenbild des Menschen generiert sowie die Gelenke als Keypoints in Form von diskreten Heatmaps ermittelt. Im Anschluss daran werden die Silhouette und die Keypoints in das 3D-Modell gefittet, wodurch die 72 Freiheitsgrade des Modells bestimmt werden. Abschließend erfolgt eine Projektion des dreidimensionalen Modells (als eine Silhouette), sowie der Keypoints in das zweidimensionale Bild [79].

Martinez-Gonzales et al. haben sich in ihrem Werk mit der Detektion anatomischer Punkte zur Erfassung der Pose aus Depth-Maps gewidmet. Verwendet wurde hierzu eine schlanke CNN-Architektur, die zuerst Features extrahiert, um anschließend die Keypoints in der Form von Gelenken über eine kaskadenartige Regression zu detektieren. Sie beheben den Mangel an Daten zum Eintrainieren durch Verwendung synthetischer Daten und Techniken zur Domänenadaptation, wodurch die eingelernten Features des Netzwerks invariant gegenüber künstlicher Repräsentation der Daten werden. Dies wird durch die Einbindung eines Netzwerks zur Domänen-Klassifizierung geschafft, das anhand der Features die Daten in reale und synthetische unterscheiden kann (siehe Grafik 4.5). Das Ziel ist hierbei, das Netzwerk zur Ermittlung der Features (bezeichnet in der Grafik als $G_f$) so einzutrainieren, damit die Domänen-Klassifizierung (bezeichnet als $G_d$) versagt bzw. getäuscht wird. Praktisch bedeutet dies eine Invarianz der Features gegenüber der Domäne, wodurch sie sich ebenfalls für reale Daten eignen [78].

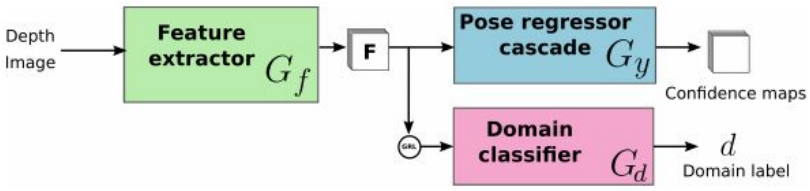



Abbildung 4.5: Veranschaulichung der Architektur zur Domänenadaptation von [78]

Chan et al. schlagen ein neues Feature zur Ermittlung der menschlichen Pose aus Punktwolken vor. Inspiriert durch die 2D-HOG Feature (Histogram of Oriented Gradients), erweitert ihre Feature VISH (Viewpoint and Shape feature Histogram) den VFH Deskriptor (siehe auch Kapitel 2.3.1) um eine räumliche Anordnung und Form-Features. Dadurch wird der Deskriptor robuster gegenüber symmetrischen Posen (zur Veranschaulichung: das Heben der linken Hand auf die Schulterhöhe kann in bestimmten Situationen in ein ähnliches VFH wie das Heben der rechten Hand auf die Schulterhöhe resultieren). Die Pose kann anschließend durch ein Matching mit den im Datenset modellierten Posen ermittelt werden. Die Methode erfordert somit keine Farbdaten und basiert einzig auf



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

der räumlichen Anordnung der Daten, die mittels ToF-Kameras aufgenommen wurden. Die Methode setzt allerdings eine einfache Segmentierung der Daten voraus [82].

Moon et al. beschäftigten sich mit der Detektion artikulierter Posen der menschlichen Hand und argumentieren, dass reine Depth-Maps aufgrund der perspektivischen Verzerrungen nicht zur zweidimensionalen Behandlung der Daten geeignet sind (siehe auch Grafik 4.6). Demnach schlagen sie einen Ansatz über eine dreidimensionale Voxelisierung vor. Hier wird die Depth-Map in eine voxelisierte Punktwolke transformiert. Der Ansatz basiert auf einer 3D-CNN Architektur, die die Wahrscheinlichkeit der Keypoints für die Hand ermittelt. Zimmermann et al. haben dies zur Ermittlung der menschlichen 3D-Pose aus RGB-D Aufnahmen genützt. Sie spalten das Problem in zwei Teile – Ermittlung der Pose aus RGB-Bildern sowie aus einer voxelisierten Punktwolke. Der Kern der 2D-Detektion der Keypoints besteht aus OpenPose, zur Ermittlung der Keypoints aus der voxeliesierten Punktwolke wurde VoxelPoseNet von Moor et al. verwendet. Die resultierenden Pose wird durch die Fusion der Koordinaten, die durch beide der Methoden extrahiert wurden, ermittelt. Dies führt zu einer guten Performance, da die Symmetrie des menschlichen Körpers einfacher durch 2D-Methoden beherrschbar ist, wobei gleichzeitig eine höhere Genauigkeit in 3D durch VoxelPoseNet erreicht werden kann [83][84].

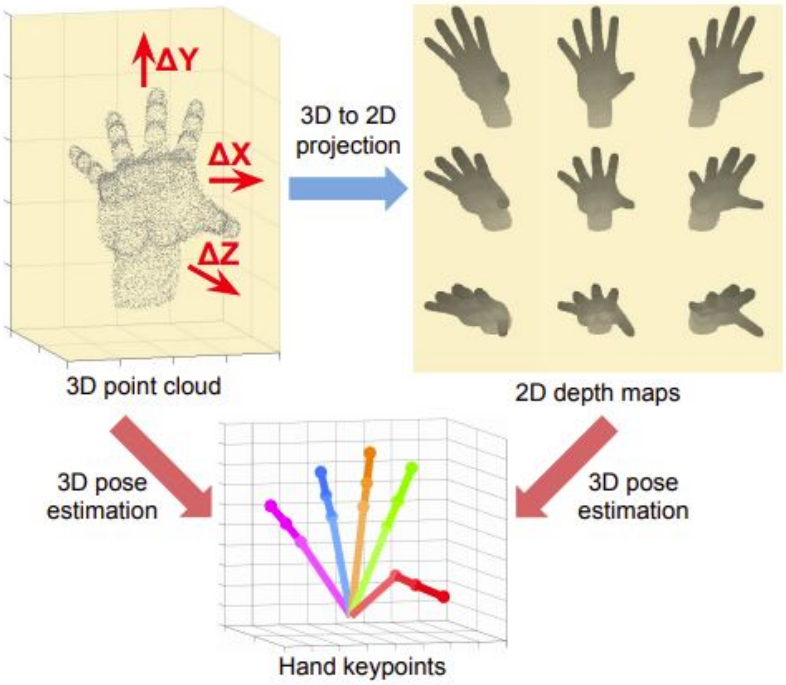



Abbildung 4.6: Bedarf der Invarianz gegenüber der perspektivischen Verzerrung bei der Rotation der Hand – Vergleich der 1 zu 1 Zuordnung von 3D-Pose zu den 3D-Keypoints (links) und N-Korrespondenzen derselben Pose in 2D zu den 3D-Keypoints durch die perspektivische Transformation [83]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Ge et al. haben die Architektur von PointNet++ (siehe auch Kapitel 3.2) durch Regression erweitert, um Gesten menschlicher Hände detektieren zu können. Um eine Invarianz gegenüber der globalen Orientierung der Hand zu schaffen, erfolgt zuerst eine Normalisierung der segmentierten Hand in ein neues Koordinatensystem definiert durch PCA. Anschließend wird die normalisierte Punktwolke in PointNet++ eingespeist, welches in eine dreifache Abstraktion der Punkte führt (siehe auch Grafik 4.7). Durch die Anwendung verschiedener Kernel-Funktionen bei jedem Niveau der Abstraktion erfolgt eine hierarchische Regression der Features, was schließlich zu den Koordinaten der Keypoints führt [85].

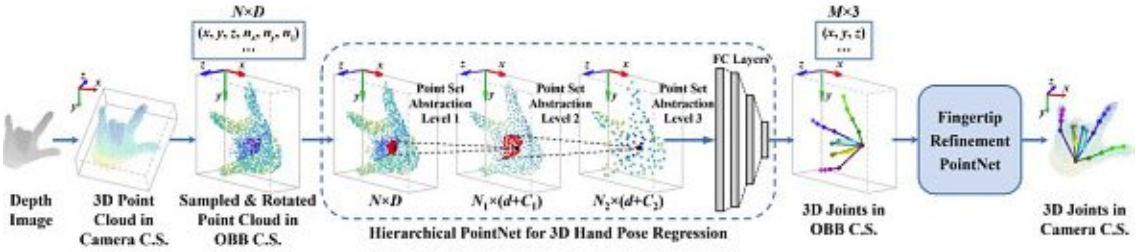



Abbildung 4.7: Architektur von Hand PointNet [85]

Zhou et al. extrahieren Keypoints des menschlichen Körpers direkt aus einer Punktwolke. Sie utilisieren ebenfalls PointNet sowie DGCNN (Dynamic Graph CNN) in ihrer Architektur (siehe auch Abbildung 4.8). Hierbei stammt das T-Net (Verweis auf Grafik 3.3), welches für eine Invarianz und die Normalisierung gegenüber räumlichen Transformationen sorgt, von PointNet und die EdgeConv Netzwerke, die für die Extraktion der Features verantwortlich sind, von DGCNN. Die Architektur richtet sich an zwei Netzwerke, die aus je zwei EdgeConvs, gefolgt durch Pooling und einem Multi-Layer-Perzeptron besteht. Im Anschluss daran folgt eine Regression der Keypoints durch drei Fully-Connected Layer [86].

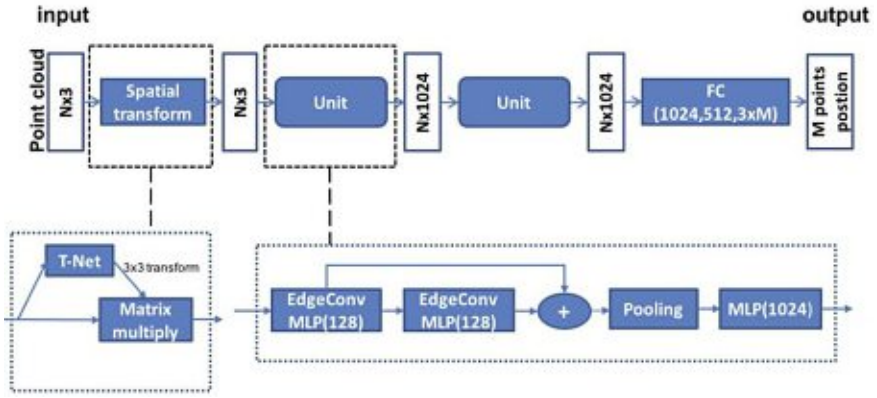



Abbildung 4.8: Architektur verwendet von Zhou et al. – räumliche Transformation durch T-Net (Bestandteil von PointNet), zwei Units bestehend aus EdgeConvs (DGCNN) gefolgt durch Fully-Connected Layer [86]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# 5 Implementierung

## 5.1 Konzept des Spatial Augmented Reality Assistenzsystems

Baustellenfertigung zeichnet sich durch große Abmessungen der Bauteile, eine hohe Varianz der Erzeugnisse und nicht-repetitive Aufgaben aus. Die Dokumentation der Arbeitsanweisungen erfolgt in der Regel in Papierform, oder gespeichert auf Rechnern außerhalb des Arbeitssystems. Eine Alternative zu der traditionellen Form der Dokumentation der Arbeitsanweisungen stellt die Verwendung der dynamischen Projektion (siehe auch Abbildung 5.1), wie von Rupprecht et al. (2020) vorgestellt [8].

Das Assistenzsystem zur dynamischen Projektion verwendet ein Lasersystem zur präzisen Darstellung von Konturen sowie die Kombination eines Projektors und eines Spiegelkopfs, wodurch die Projektion der Arbeitsanweisungen auf große Abstände möglich ist. Des Weiteren kann die Krümmung der Oberflächen durch die Einbringung einer geeigneten Verzerrung nachgebildet werden und somit die Informationen verzerrungsfrei auf die Oberfläche projiziert werden. Der Einsatz eines solchen Assistenzsystems zur Bereitstellung der Arbeitsanweisungen zeichnet sich durch eine sehr gute Usability sowie eine hohe Nutzerakzeptanz aus und führt zu einer Verringerung der Zykluszeit [8].

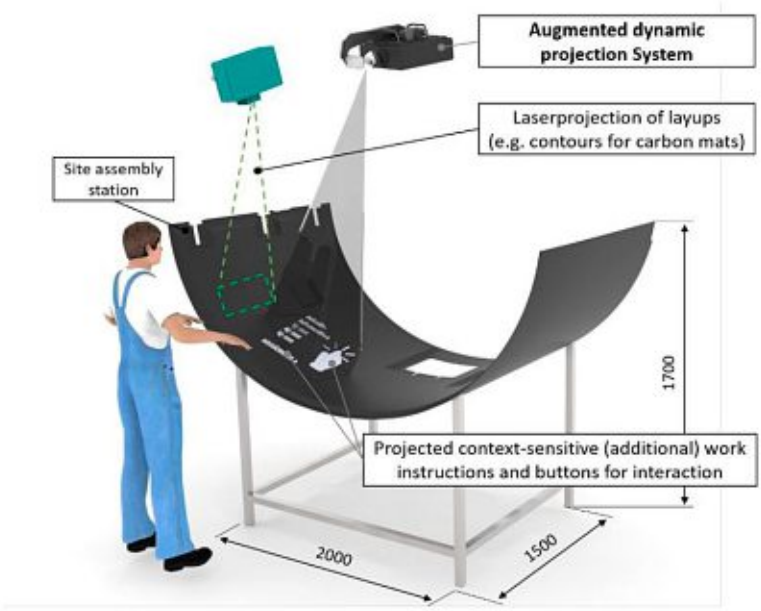



Abbildung 5.1: Veranschaulichung des Konzeptes des Spatial Augmented Reality Assistenzsystems [8]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die kontextuelle Situativität eines AR-Assistenzsystems erfordert die Wahrnehmung der Informationen aus der Umgebung. Aufbauend auf dem Konzept des Assistenzsystems vorgestellt von Rupprecht et al. wird dies durch ein System zur Informationsaufbereitung vervollständigt. Hierzu wird ein statisches Kamerasystem bestehend aus Time of Flight Kameras verwendet (siehe auch Kapitel 5.2), wodurch Mapping großräumiger Arbeitsbereiche, wie etwa im Falle der Baustellenfertigung, ermöglicht wird.

## 5.2 Beschreibung des Set-Ups

Das Set-Up besteht aus drei *Argos 3D – P330* Time of Flight Kameras der Firma *Becom*, die in der Pilotfabrik der Technischen Universität Wien aufgestellt sind. Die Kameras sind auf der Decke aufgehängt und um ungefähr 120 Grad versetzt, wodurch eine 3D-Aufnahme des dazwischenliegenden Arbeitsplatzes ermöglicht wird. Zusätzlich wurde eine weitere *Argos 3D – P330* Kamera zur Verfügung gestellt, die zum Ausprobieren der Algorithmen verwendet werden konnte.

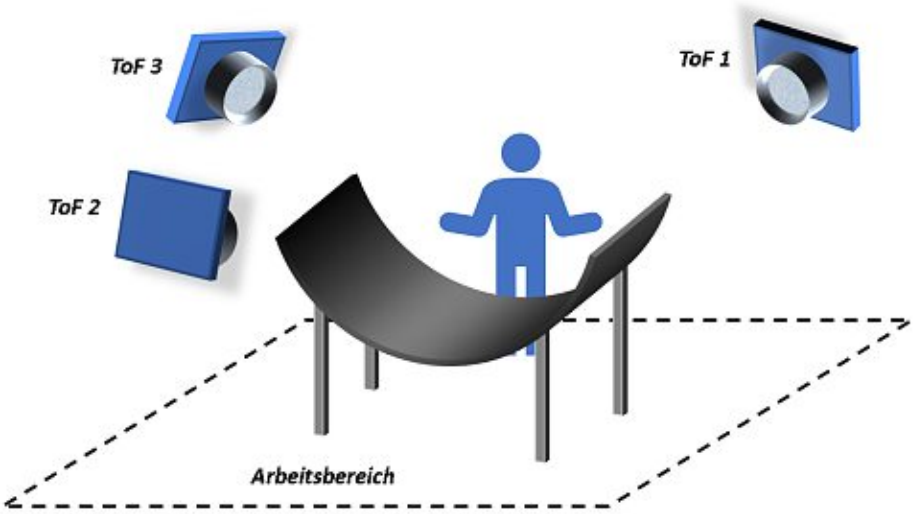



Abbildung 5.2: Veranschaulichung des Set-Ups in der Pilotfabrik der TU Wien bestehend aus drei Time of Flight Kameras und dem dazwischenliegenden Arbeitsbereich

Die Kameras verfügen über einen 352 x 287 Pixel Time-of-Flight Sensor mit einer Reichweite von zehn Metern und einem Aufnahmewinkel von 80 Grad. Die Beleuchtung der Szene durch die Kameras erfolgt im Infrarotbereich bei 850 nm. Zusätzlich besitzen die Kameras einen 2D-CMOS Farbsensor mit einer Auflösung von 1280 x 720 Pixel. Die Kommunikation zwischen den Kameras und dem Rechner erfolgt über Ethernet, wobei UDP-Protokoll zur Steuerung der Kamera und das TCP-Protokoll für den Datenstream verwendet werden. Die Daten können von der Kamera über zwei Arten abgefragt werden [87]:



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Über eine C++ Schnittstelle (als Bibliothek, bezeichnet als *BltTofApi*)
- Über eine MATLAB Schnittstelle (als Mex-Wrapper der *BltTofApi*)

Eine gleichzeitige Aufnahme aller Daten wird durch die Kamera nicht gestattet. Die Aufnahme richtet sich nach 11 Aufnahmemoden, die eine Kombination von bis zu drei Arten der Daten ermöglichen. Somit ist zum Beispiel ein zeitgleiches Abfragen der Farbdaten und der Punktwolke nicht gestattet, sondern zuerst muss der jeweilige Modus geändert werden. Im Rahmen dieser Arbeit wurden folgende Modi am häufigsten verwendet:

- Punktwolke
- Punktwolke, Amplituden und Depth-Map
- 2D Farbaufnahme, Amplituden und Depth-Map

Die Kameras wurden überwacht kalibriert und beziehen sich auf ein gemeinsames Koordinatensystem. Somit wird im Rahmen dieser Arbeit das System als vorkonfiguriert angenommen, die Aufgabe bezieht sich lediglich auf die Erstellung einer Schnittstelle zur Erfassung und Weitergabe der ausgewählten Informationskanäle zum Processing. Die Schnittstelle wurde mittels *BltTofApi* erstellt und setzt sich aus folgenden Komponenten zusammen:

- Die Initialisierung besteht aus dem Einstellen der Networking-Parameter zur Verbindung mit den Kameras
- Als Frame-Capture wird hier die Erfassung der Daten durch die Kamera bezeichnet
- Die Verarbeitung der Daten besteht im Wesentlichen aus der Transformation auf das Format der entsprechenden Bibliothek
- Abschließend erfolgt eine Speicherung der verarbeiteten Daten im geeigneten Format zum weiteren Processing

## 5.3 Anforderungen an das System

Aufbauend auf der Architektur der Augmented Reality-basierten Assistenzsysteme (präsentiert in Abbildung 2.4) erfolgt nun die Abgrenzung des Vision-basierten Systems als eines Teilsystems des AR-basierten Assistenzsystems. Diese ist ersichtlich in der Abbildung 5.3. Des Weiteren wird das System in zwei funktionale Kernkomponenten nach dem Objekt des Interesses in das Teilsystem **„Bauteil“** (Erfassung der Pose des Bauteils) sowie das Teilsystem **„Mensch“** (Erfassung der Pose des Mitarbeiters und Mensch-Maschine-Interaktion) aufgeteilt.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

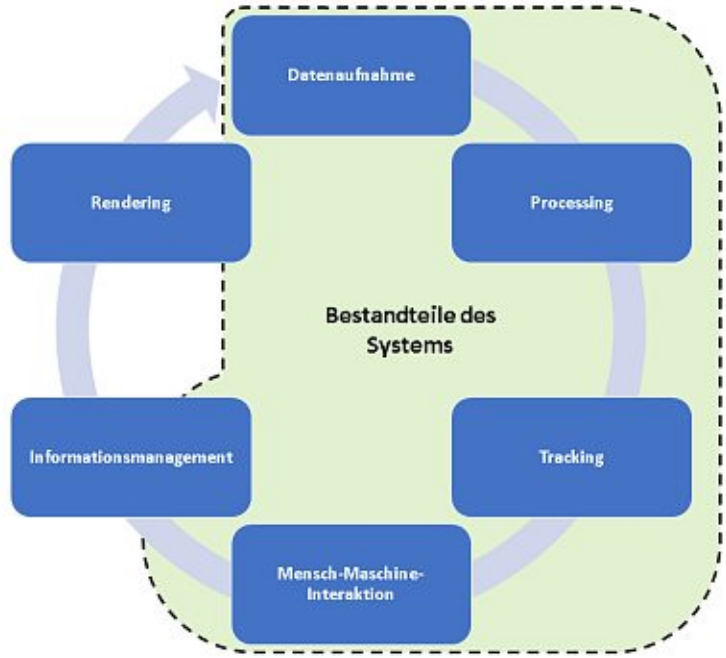



Abbildung 5.3: Darstellung der Komponenten eines markerlosen Vision-basierten Trackingsystems im Rahmen der AR-Architektur

In Anlehnung an den Leitfaden zur Qualitätsbewertung von Softwareanwendungen der ISO/IEC 25000 [88] werden grobe Anforderungen an das Konzept definiert. Die Anforderungen werden im Rahmen dieser Arbeit in funktionale und nicht-Funktionale Anforderungen aufgeteilt. Abgeleitet von der Abgrenzung des Systems im Kontext der AR-basierten Architektur richten sich die funktionalen Anforderungen an die Datenaufnahme und die darauffolgende Vorbereitung der Daten (Processing) zur Ermöglichung des Trackings. Dies bedarf zumindest im Falle des Bauteils der Einbeziehung eines CAD-Modelles (Informationsmanagement). Im Falle des Mitarbeiters soll weiters eine Mensch-Maschine-Schnittstelle im Sinne von NUI (Natural User Interface, siehe auch Kapitel 2.1.1) aufgebaut werden.

Weiters werden nicht-funktionale Anforderungen an das System definiert, die sich insbesondere an die Änderbarkeit des Systems sowie die Effizienz richten. Dabei bedeutet die Änderbarkeit die Möglichkeit zur effektiven Systemänderung, ohne die Funktion des Systems zu beeinträchtigen. Die Effizienz eines Systems richtet sich an folgende Kriterien (vgl. [89]):

- Zeitverhalten
- Ressourcen-Auslastung
- Kapazität



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Um das Potential des System zur Ansteuerung eines Spatial Augmented Reality-Assistenzsystems abschätzen zu können, erfolgt nach der Implementierung die Evaluierung auf die ausgewählten Kriterien. Diese wird im Kapitel 5.8 behandelt.

## 5.4 Konzeptionierung und Beschreibung des Vorgehens

Ausgehend von der Abgrenzung der Systemkomponenten in Grafik 5.3, wird in diesem Kapitel die Erstellung eines Konzeptes zum markerlosen 3D-Tracking von Bauteil und Mitarbeiter beschrieben. Dies wird auf folgende Art strukturiert: als erstes erfolgt eine generelle Beschreibung der Ausgangssituation und der Herausforderungen, die im Rahmen des Use-Cases bewältigt werden müssen. Zunächst werden die Herausforderungen und die Begründung des gewählten Ansatzes für das Teilsystem zum Tracking des Bauteils beschrieben. Im Anschluss daran erfolgt eine analoge Beschreibung für das Teilsystem „Mensch“. Abschließend erfolgt eine Zusammenfassung und die Vorstellung des Konzeptes für ein Spatial-AR-basiertes Assistenzsystem für den Use-Case mit Time of Flight Kameras.

### 5.4.1 Ausgangssituation

Die allgemeinen Herausforderungen an das Vision-System werden hier in Technik-bedingte und Arbeitswissenschaftlich-bedingte Herausforderungen strukturiert. Die Aufteilung ist in Tabelle 5.1 sichtbar, im darauffolgenden Abschnitt erfolgt eine kurze Beschreibung der Herausforderungen.

| **Arbeitswissenschaftlich-bedingt** | **Technik-bedingt** |
|---|---|
| Clutter und Okklusionen verursacht durch die Interaktion zwischen dem Menschen und dem Arbeitssystem | Auflösung und Distanz der Kameras zu der Szene |
| Bewegungen ausgehend aus der Interaktion zwischen dem Menschen und dem Arbeitssystem | Synchronisierung des Kamerasystems |

Tabelle 5.1: Aufteilung der allgemeinen Herausforderungen von markerlosem Tracking mit einem ToF-Multi-Kamerasystem

Das zu erfassende Arbeitssystem wird, im Kontext der Arbeitswissenschaft, als eine Baustellenfertigung klassifiziert. Dementsprechend besteht die Interaktion zwischen dem Mitarbeiter und dem Bauteil aus Arbeit um ein statisches Bauteil (in Anlehung an [7]). Dies führt zu wechselwirkenden Okklusionen und Clutter zwischen dem Mitarbeiter sowie dem Bauteil, wodurch eine vollständige Aufnahme aller Daten von Interesse nie garantiert wird. Praktisch heißt dies insbesondere, dass in bestimmten Posen der Mensch nicht vollständig im Sichtfeld der Kamera ist. Umgekehrt bedeutet dies ebenfalls eine teilweise Okklusion des Bauteils durch den Menschen.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Des Weiteren stellen die schnellen Bewegungen des Mitarbeiters eine Herausforderung dar. Dies ist einerseits bedingt durch die Wechselwirkung zwischen der Geschwindigkeit der Bewegung und der Erfassungsrate der Kamera (siehe auch [90]), sowie durch die Synchronisierung des Kamerasystems, auf die im nächsten Abschnitt näher eingegangen wird.

Zu den Technik-bedingten Herausforderungen, hervorgerufen durch die Wahl des Set-Ups, gehören die Auflösung der Punktwolke sowie die Synchronisierung der Kameras. Für eine simultane Operation mehrerer Time of Flight Kameras besteht das Problem der Interferenz der Strahlung emittiert durch die einzelnen Kameras, die hardware- bzw. softwaretechnisch gelöst werden muss (in Anlehnung an [91]). Dies impliziert, dass eine perfekte Synchronisierung in der Regel nicht erreicht werden kann, da durch die Interferenz die Aufnahme der Daten zur Zeit der Abfragung stattfinden kann. Praktisch bedeutet dies bei sehr schnellen Bewegungen eine Verschiebung der Punktwolke des Menschen in Richtung der Bewegung (siehe auch Grafik 5.4), wodurch das Tracking des Mitarbeiters beeinflusst werden könnte. Eine softwaretechnische Abhilfe über Algorithmen zur allgemeinen Registrierung der drei Punktwolken ist aufgrund der statischen Umgebung, die wegen der Position des Bauteils ebenfalls von Interesse ist, ohne Segmentierung der Punktwolken der Person nicht gestattet.

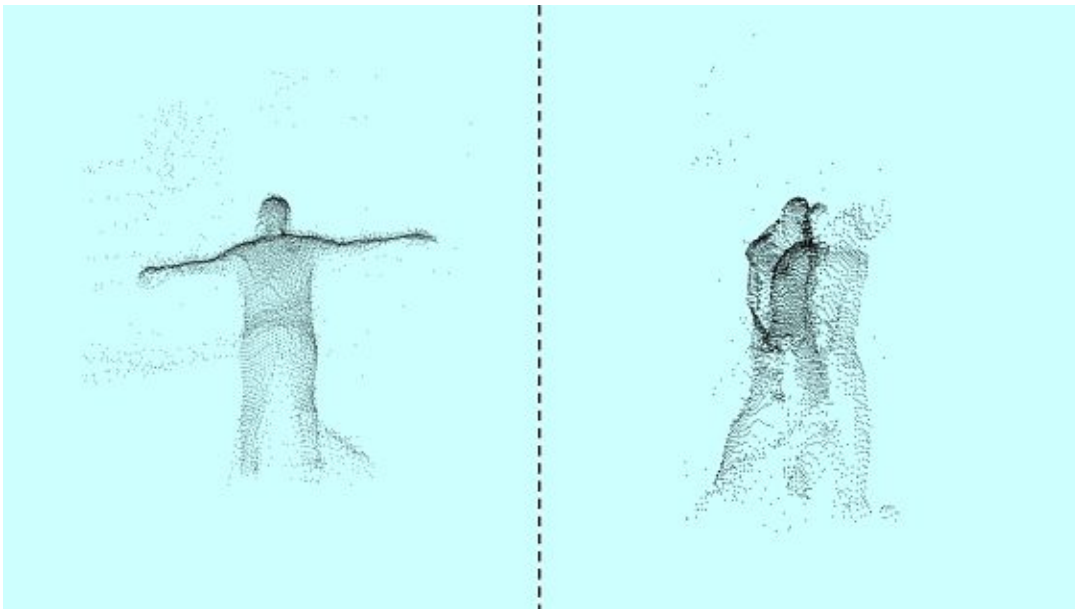


Abbildung 5.4: Synchronisierungsprobleme der Punktwolke des Menschen bei schnellen Bewegungen: links eine statische Pose und rechts Bewegung (Spazieren)

Weiters lässt sich das Set-Up als ein statisches Multi-Kamerasystem beschreiben, demnach gilt insbesondere die Distanz der Kamera zu der Szene und die damit verbundene Auflösung der Punktwolke als eine wesentliche Herausforderung (vgl. [25]). In Kombination mit der generell geringeren Auflösung der Time of Flight Kameras im Vergleich zu anderen Methoden (vgl. [92]) ergibt sich daher eine „Sparse-Punktwolke“ (Englisch:



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

dünnbesetzt bzw. geringe Dichte) der Szene, und insbesondere der in der Mitte der Szene liegenden Objekte von Interesse. Dies mindert die Qualität der Punkwolke und kann dementsprechend die Ableitung der Features beeinflussen.

Des Weiteren wird bei dieser Arbeit auf die Farbdaten verzichtet, wodurch ein wichtiger Informationskanal verloren geht. Dies wird einerseits durch die Generalisierung des Konzeptes zur Erkennung der Objekte mittels Time of Flight Kameras begründet, die oft über keinen Farbsensor verfügen, als auch durch die Notwendigkeit der mathematischen projektiven Transformation der Farbdaten, aufgrund der unterschiedlichen intrinsischen Parameter der Sensoren begründet (vgl. Kapitel 2.2.1). Daher wird in dieser Arbeit die Annahme getroffen, dass Amplitudenbilder die Farbdaten vollständig ersetzten können, was im Kapitel 5.6 ebenfalls gezeigt wird.

### 5.4.2 Teilsystem Bauteil

Spezifisch für das Bauteil zählt die Reflexion der Strahlung durch die glatte Oberfläche des Bauteils zu den wesentlichen Herausforderungen. Daher wird die Mitte des Bauteils nicht ausreichend aufgenommen, was zu einer Abweichung von der realen Geometrie und des CAD-Modells führt (siehe auch Grafik 5.5). Des Weiteren können durch die Sparse-Auflösung der Punktwolke nicht alle geometrischen Merkmale des Bauteils dargestellt werden. Das Bauteil weist ebenfalls keine Textur auf und verfügt aufgrund der Geometrie nur über wenig relevante lokale Merkmale.

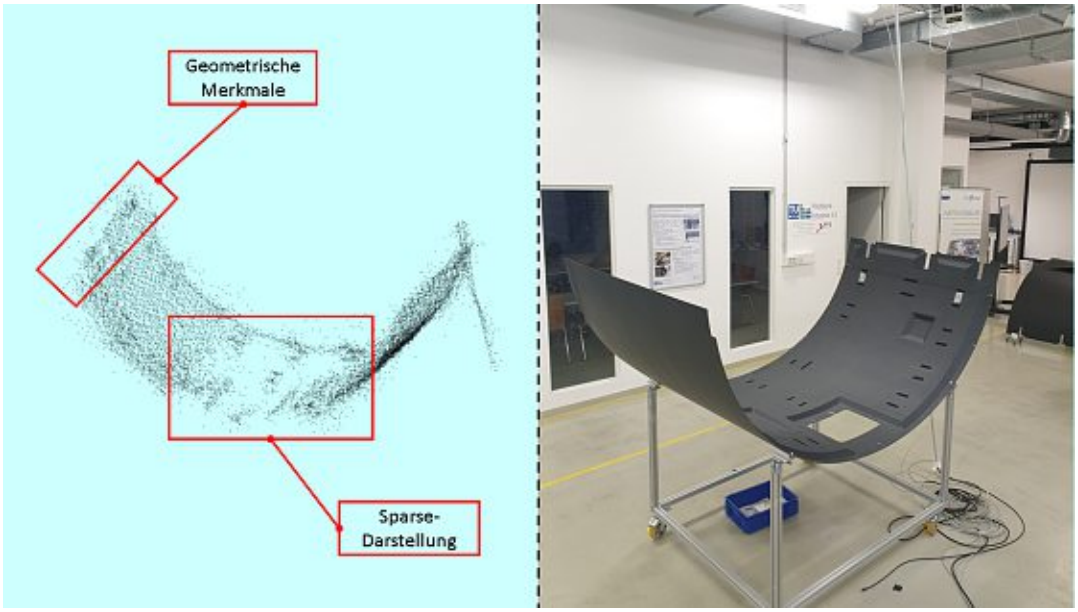



Abbildung 5.5: Darstellung der Herausforderung bei der Erkennung des Bauteils: Punktwolke des Bauteils (links) und ein Foto des realen Objektes (rechts)

Abgeleitet aus den im Kapitel 3.1 präsentierten Grundlagen widmet sich dieser Teil der Arbeit der Evaluierung der dort gelisteten Ansätze, die hier kurz zusammenge-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

fasst werden. Die Ansätze werden generell in Template-Matching, Sparse-Features und Dense-Approach unterteilt, des Weiteren erfolgt die Erkennung der Pose direkt aus einer Punktwolke entweder mittels im Voraus gewählten und konstruierten Features und Deskriptoren (traditioneller Ansatz), oder über Deep Learning.

In Hinsicht auf die Unterteilung in den traditionellen sowie den Deep-Learning Ansatz wird bei dieser Arbeit wie folgend argumentiert. Obwohl DL-basierte Methoden vielversprechende Ergebnisse hinsichtlich der Robustheit aufweisen (siehe auch Kapitel 3.2), gilt, dass sie generell mit der Aufbereitung eines großvolumigen Datensatzes zum Eintrainieren eines Netzwerks verbunden sind (siehe auch Kapitel 2.3.1). Des Weiteren muss die Qualität der Daten berücksichtigt werden. Einerseits gilt, dass neuronale Netzwerke mit einem großen Volumen an CAD-Daten eintrainiert werden können (vgl. [93]). Auf der anderen Seite besteht allerdings die Gefahr der Domänenabhängigkeit der dadurch eingelernten Features (vgl. [78]). Dies kann durch die Abbildungsfehler der Punktwolke (etwa Sparse-Abbildung und dadurch das Fehlen der Merkmale) noch verstärkt werden. Eine mögliche Abschaffung der Domänenabhängigkeit stellt die Einbeziehung realer Daten in den Datensatz (siehe auch [94]). Die Voraussetzung hierfür ist eine hohe Anzahl an Aufnahmen der Szene, aus denen zusätzlich das Bauteil geeignet segmentiert werden muss (vgl. [94]). Eine geeignete Segmentierung für den Use-Case erfordert allerdings umfangreiches Preprocessing, um das Bauteil aus der Szene zu segmentieren. Als nächstes wird ebenfalls auf die Anforderungen eingegangen. Hier stellt insbesondere die Änderbarkeit ein wichtiges Kriterium im Rahmen des Use-Cases. Eine effiziente Übertragbarkeit des Systems auf andere Bauteile ist bei Deep Learning nur bedingt möglich, da sie mit einem erneuten Eintrainieren verbunden ist. Aus diesen Gründen wird in dieser Arbeit der traditionelle Ansatz zur Erfassung der Pose des Bauteils vorgeschlagen.

Template-Matching weist eine sehr gute Robustheit gegenüber Sparse-Darstellung (vgl. [60]), gleichzeitig eignet sich das Verfahren sehr gut für texturlose Objekte, da bei der Erstellung des Template bzw. Verwendung eines CAD-Modelles in der Regel auf die Textur verzichtet wird (vgl. [54][60]). In Hinsicht auf die beschriebenen Herausforderungen stellen sich allerdings Okklusionen und Clutter durch den Mitarbeiter als potenzielle Probleme dieses Ansatzes, da dies die Darstellung des Bauteils in der Szene wesentlich beeinflusst (vgl. [54]).

Sparse-Features, welche um bestimmte Keypoints konstruiert werden, sind in der Regel sehr robust hinsichtlich Okklusionen und Rauschen, allerdingts werden sie durch lokale Änderungen der Nachbarschaft beeinflusst. Zeitgleich ist in der Regel eine bestimmte Dichte notwendig, um geeignete Features konstruieren zu können (vgl. [43]). Weiters kann in Hinsicht auf die bereits erwähnten Herausforderungen das Fehlen von Textur sowie fehlerhafte Darstellung jeglicher geometrischen Merkmale außer der globalen Form des Bauteils zu Problemen in Bestimmung der Keypoints führen.

Aus diesen Gründen wird der in dieser Arbeit der Dense-Approach mittels robuster Deskriptoren zur Erkennung der Pose des Bauteils aus der Punktwolke vorgeschlagen.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Obwohl Dense-Approach durch Clutter und Okklusionen behaftet ist und oft eine Segmentierung voraussetzt (vgl. [43] [62]), wird in dieser Arbeit angenommen, dass die globale Struktur der Punktwolke des Bauteils das signifikanteste Merkmal zur Erkennung darstellt. Das Vorgehen wurde inspiriert durch [61] und die Methode baut auf Point Pair Features [95] auf, mit denen die Erkennung des Bauteils aus einer unstrukturierter Punktwolke möglich ist. Das Vorgehen wird näher im Kapitel 5.4.4 vorgestellt, die Implementierung wird ausführlich im Kapitel 5.7 beschrieben.

### 5.4.3 Teilsystem Mensch

Die allgemeinen Herausforderungen in der Ermittlung der Pose des Mitarbeiters richten sich im Wesentlichen an die Herausforderungen, die im Kapitel 5.4.1 präsentiert wurden. Das folgende Kapitel widmet sich dementsprechend der Begründung der Wahl des Vorgehens unter Berücksichtigung der genannten Herausforderungen.

Generative Ansätze verwenden ein Modell eines menschlichen Körpers zur Bestimmung der Pose. Dies führt zu einer sehr guten Generalisierung, und somit zu einer hohen Präzision, da sie in der Regel robuster gegenüber nicht-eingelernten Posen sind. Hingegen verwenden diskriminierende Methoden kein Modell des menschlichen Körpers, stattdessen erfolgt die Bestimmung der Pose entweder durch das Einlernen von geeigneten Methoden zur Bestimmung der Pose direkt aus den Daten (Learning-basiert), oder durch einen Vergleich der Aufnahme mit den zuvor gespeicherten Vorlagen (Vorlagenbasiert). Der wesentliche Vorteil des diskriminierenden Ansatzes besteht in einer höheren Geschwindigkeit dieser Methoden. Weiters werden die Methoden in Top-down sowie Bottom-up Verfahren gegliedert. Top-down Methoden segmentieren in erster Linie die Personen aus dem Bild, um anschließend Keypoints aus dem extrahierten Segment zu ermitteln. Bottom-up Algorithmen ermitteln dagegen die Keypoints direkt aus dem Abbild ohne jegliche Segmentierung. Dies führt auf einer Seite insbesondere bei Multi-Personen-Erkennung zu einem geringeren Rechenaufwand, auf der anderen Seite erschweren Überlappungen der Körperteile bei Bottom-up Verfahren die Bildung der Korrespondenzen zwischen den Körperteilen und den zugehörigen Keypoints (vgl. [72][75]).

Im Rahmen dieser Arbeit wurde weiter für die Implementierung der diskriminierende, Learning-basierte Ansatz gewählt. Begründet wird dies durch die hohe Verfügbarkeit an Open-Source Methoden, sowie durch die Fähigkeit dieser Methoden, die Pose ganzheitlich zu beschreiben, die durch die Verwendung von Modellen und Deskriptoren nicht gegeben ist (vgl. [81]). Zunächst gilt insbesondere für das zu entwickelnde System das Problem der Bestimmung der Inputdaten zur Erkennung der Pose. Anhand der im Kapitel 4.2 präsentierten Recherche lassen sich die potenziellen Datenformate in zweidimensionale (Bilder oder Depth-Maps) sowie dreidimensionale (strukturierte oder unstrukturierte Punktwolke) aufteilen.

Demnach soll das Format der Inputdaten auf die Eignung zur Erkennung evaluiert werden. Der Vorteil der dreidimensionalen Methoden besteht bei dem Use-Case insbesondere durch die Einfachheit der Datenaufnahme und Datenverarbeitung, da lediglich



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

eine Punktwolke verwendet werden müsste. Diese darf dabei entweder als eine voxelisierte Depth-Map und somit strukturiert (vgl. [83][78]) oder unstrukturiert (vgl. [86]) vorkommen. Abgeleitet aus der Literaturrecherche gelten insbesondere die Invarianz der Perspektive sowie die dreidimensionale Genauigkeit als die Vorteile dieser Methoden. Zu den Nachteilen dieser Methoden zählen in Hinsicht auf die allgemeinen Herausforderungen die Sparse-Darstellung bzw. die Qualität der Abbildung als eine Punktwolke. Im Falle der strukturierten Methoden könnten lediglich separate Einzelaufnahmen der drei Kameras verwendet werden, da durch die Verschmelzung der Einzelaufnahmen zu einer Punktwolke der gesamten Szene die Strukturiertheit verloren geht. Hierbei besteht die Gefahr, dass die Dichte der Punktwolke des Menschen zu gering wäre, um Features zu ermitteln. Des Weiteren stellt die Synchronisierung der Kameras eine potenzielle Gefahrquelle dar, da hierdurch bei schnellen Bewegungen oft keine ganzheitliche und stetige Abbildung des Mitarbeiters als einer Punktwolke möglich ist (siehe auch Abbildung 5.4).

Daher werden zur Lösung des Problems der Posenerkennung des Mitarbeiters in dieser Arbeit zweidimensionale Methoden vorgeschlagen. Hierzu sollen weiters die Amplituden, die sich in Bilder transformieren lassen, verwendet werden. Die Implementierung basiert auf prätrainierten Netzwerken von Mask R-CNN [96] als auch OpenPose [97], wobei zur Projizierung der zweidimensionalen Pose in die Punktwolke die anfängliche Strukturiertheit der Kameraaufnahmen und ein vereinfachtes Modell des menschlichen Körpers verwendet wird. Des Weiteren wird Mask R-CNN zur Segmentierung der Punktwolke des Menschen aus der Punktwolke der Szene verwendet, wodurch der wechselwirkende Clutter zwischen dem Mitarbeiter und dem Bauteil teilweise behoben werden kann. Die Implementierung wird weiter ausführlich im Kapitel 5.6 vorgestellt.

### 5.4.4 Ableitung der Architektur des Assistenzsystems

Entsprechend der Abgrenzung der Systemkomponenten, sowie ausgehend von den genannten Herausforderungen und der Wahl des Vorgehens erfolgt die Ableitung der Architektur für das Spatial Augmented Reality Assistenzsystem mit Time of Flight Kameras. Die vorgeschlagene Architektur ist in der Abbildung 5.6 sichtbar, wobei der Datenflow zwischen den Systemkomponenten durch einen blauen Pfeil modelliert wird.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

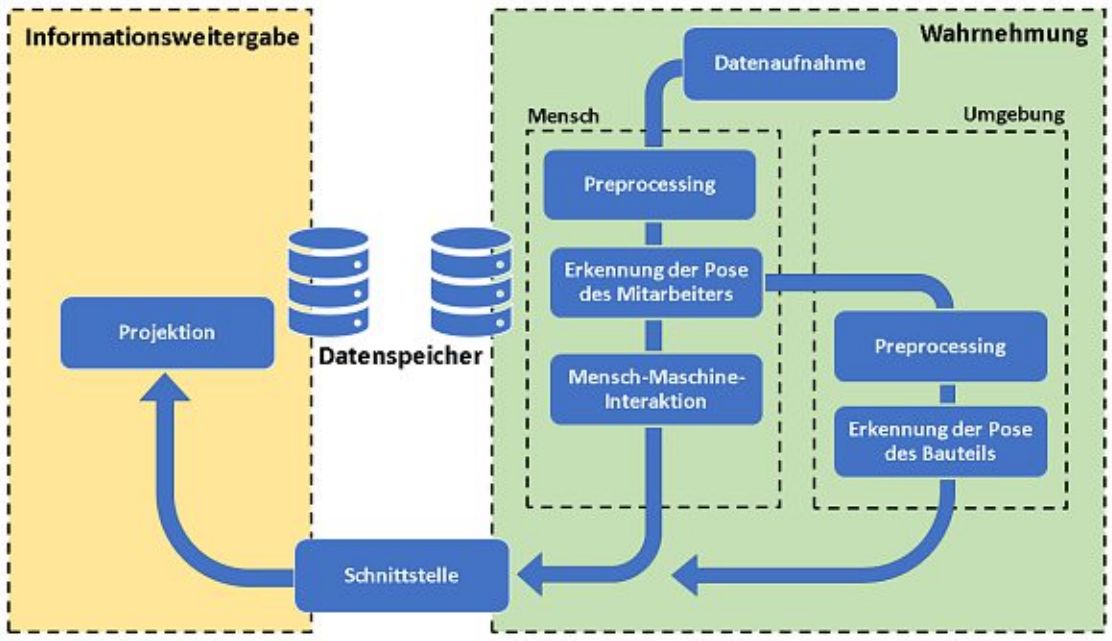



Abbildung 5.6: Vorgeschlagene Architektur für das Spatial Augmented Reality Assistenzsystem in der Pilotfabrik der TU Wien

In Anlehnung an Kapitel 2.1 wird das Assistenzsystem in ein wahrnehmendes System, das den Menschen und die Umgebung (das Bauteil) erfasst, sowie ein System zur Informationsweitergabe aufgeteilt. Die zwei Systeme sind über eine Schnittstelle verbunden, die räumliche Informationen aus dem wahrnehmenden System bezieht und diese, entsprechend den Anforderungen des Systems, zur Informationsweitergabe verarbeitet und weiter transferiert. Des Weiteren wird das Informationsmanagement in einen Datenspeicher zur Speicherung der Informationen zur Projektion sowie einen Datenspeicher für das wahrnehmende System aufgeteilt.

Die Beschreibung des Datenflows der wahrnehmenden Komponente des Assistenzsystems erfolgt in mehreren Schritten:

1. **Datenaufnahme** - die Datenaufnahme erfolgt über die programmierte Schnittstelle für alle Kameras direkt nacheinander, anschließend erfolgt die Verarbeitung und Speicherung im geeigneten Format. Zu den extrahierten Daten zählen die Punktwolke und die Amplituden je Kamera.

2. **Mensch** - zunächst erfolgt das Preprocessing der Amplitudenbilder und die Erkennung der Pose des Menschen. Das Ergebnis dieser sind die räumlichen Koordinaten der Keypoints, welche die Pose des Menschen beschreiben, die aussegmentierte Punktwolke des Menschen und die Punktwolke der Szene ohne den Menschen. Anhand der Koordinaten der Keypoints kann im Anschluss daran die Mensch-Maschine-Interaktion gestaltet werden.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. **Bauteil** - die Punktwolke der Szene ohne den Menschen wird in die Datenpipeline zur Erkennung der Pose des Bauteils eingespeist. Hierzu wird die Punktwolke zuerst verarbeitet (Preprocessing), abschließend wird die Pose des Bauteils ermittelt.

Die getrennte Struktur des Datenflows wird einerseits durch die Notwendigkeit der Strukturiertheit zur Extraktion der räumlichen Koordinaten der Keypoints aus der Punktwolke begründet, andererseits führt die zuerst stattfindende Segmentierung der Punktwolke des Menschen aus der Punktwolke der Szene zu einer Reduktion von Clutter, was sich positiv auf die Beschreibungsfähigkeit des globalen Deskriptors zur Erkennung der Pose des Bauteils einwirkt.

Der Output des wahrnehmenden Systems setzt sich somit aus drei Komponenten zusammen. Einerseits werden die dreidimensionalen Koordinaten der Keypoints des menschlichen Körpers geliefert. Diese bilden die Basis zur Bestimmung der räumlichen Pose und Orientierung des Mitarbeiters relativ zu dem Bauteil. Des Weiteren können die mittels Koordinaten berechneten Winkel, unter Einbeziehung zeitlicher Angaben, zur Erkennung der derzeitigen Körpergeste verwendet werden. Abschließend wird die Pose des Bauteils erkannt, und als eine Transformationsmatrix wiedergegeben, die die Informationen über die Translation und die Rotation des Bauteils beinhaltet.

## 5.5 Übersicht verfügbarer Bibliotheken

Dieses Kapitel stellt eine Übersicht ausgewählter Programmierbibliotheken und Tools dar, die zum Processing von Punktwolken, sowie den zugehörigen zweidimensionalen Amplitudenbildern, verwendet werden können. Die Kriterien zur Auswahl einer geeigneten Bibliothek richten sich dabei insbesondere an die Verfügbarkeit aller Methoden, die zur Implementierung eines ToF-basierten wahrnehmenden Assistenzsystems notwendig sind. Die Bibliotheken werden in Tabelle 5.2 gelistet.

| **Name** | **Programmier-sprache** | **Fokus auf** | **Quelle** |
|---|---|---|---|
| Scikit-Image | Python | Bilder | [98] |
| BoofCV | Python, Kotlin | Bilder | [99] |
| OpenCV | Python, C++, Java | Bilder (Punktwolken) | [100] |
| MATLAB Computer Vision Toolbox | MATLAB | Bilder und Punktwolken | [101] |
| Cilantro | C++ | Punktwolken | [102] |
| Open3D | Python, C++ | Punktwolken | [103] |
| Point Cloud Library (PCL) | C++ | Punktwolken | [104] |

Tabelle 5.2: Übersicht der Bibliotheken



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Im Rahmen dieser Arbeit wurde das System mittels OpenCV und Point Cloud Library entwickelt. OpenCV (Open Source Computer Vision Library) ist eine Programmierbibliothek basierend auf C++. Zur Interaktion mit der Bibliothek kann sowohl C++ als auch Python und Java verwendet werden. Die Bibliothek bietet eine Vielzahl von Modulen zum Processing und Analyse von Bildern, Extraktion von Features und Erkennung von Objekten. Des Weiteren verfügt die Bibliothek über ein Modul zur einfachen Integrierung prätrainierter DL-Modelle [100]. PCL (Point Cloud Library) ist eine holistische C++-Bibliothek zur Entwicklung Punktwolkenbasierter Anwendungen. Die Interaktion mit der Bibliothek erfolgt ausschließlich über C++. Die Funktionalität deckt eine Vielzahl von Methoden zu Preprocessing, der Bestimmung und dem Matching von Features, der Rekonstruierung von Oberflächen und der Erkennung von Objekten [104].

Die Wahl der Bibliotheken erfolgte anhand der großen Verfügbarkeit der Algorithmen, die zur Implementierung benötigt werden. Auch bieten die ausgewählten Bibliotheken durch ihren Umfang ein großes Potential für zukünftige Erweiterungen. Bedingt durch die Programmiersprache der Schnittstelle der Kamera, sowie durch die Programmiersprachen zur Interaktion mit den Bibliotheken, wurde C++ als Implementierungssprache gewählt.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 5.6 Implementierung des Teilsystems Mensch

Das Teilsystems „Mensch“ wird in fünf funktionale Blöcke aufgeteilt (siehe Abbildung 5.7). Die globale Pipeline des Teilsystems „Mensch“ fängt mit der Datenaufbereitung der extrahierten Daten an. Im Anschluss daran erfolgt das Preprocessing zweidimensionaler Daten. Zunächst erfolgt die Erkennung des Menschen aus dem 2D-Amplitudenbild und eine Segmentierung. Hierzu wird ein prätrainiertes Deep Learning Netzwerk (in Abbildung 5.7 als DL-Modell bezeichnet) verwendet. Anhand der Erkennung wird weiter eine Region extrahiert und diese zur Ermittlung der menschlichen Pose in 2D angewendet, wozu ebenfalls ein prätrainiertes Netzwerk hereingezogen wird. Abschließend erfolgt die Projektion der zweidimensionalen Pose in die Punktwolke, wozu ein vereinfachtes Modell des menschlichen Körpers verwendet wird, um die Güte der Projektion zu verbessern (in Abbildung 5.7 als Mensch-Modell bezeichnet). Bei der Implementierung erfolgt die Erkennung der Pose nur durch eine der drei Kameras. Hierzu hat sich die Kamera 3 bewährt (siehe Abbildung 5.2), da diese im Vergleich zu den anderen über die beste, nicht-geclutterte Perspektive verfügt.

Abbildung 5.7: Globale Pipeline des Teilsystems „Mensch“

Aufbauend auf der groben Beschreibung der Schnittstelle zu den Kameras erfolgt als erstes die Aufnahme und die Datenaufbereitung der Daten. Dies bedeutet das Umwandeln der Daten in ein passendes Format entsprechend der verwendeten Bibliotheken. Hierzu wird zuerst die Punktwolke auf die Einheit Meter normalisiert. Zunächst werden die Amplitudenbild in ein Grayscale-Bild umgewandelt. Dies setzt ein Thresholding der Amplituden voraus. Das Ziel ist es hierbei, Werte, die durch sehr hohe Reflektivität der am Bauteil sowie in der Pilotfabrik angebrachten Marker entstehen, zu mindern. In Rahmen dieser Arbeit hat sich ein Thresholding der Intensität gleich 550 bewährt. Die Visualisierung verschiedener Thresholding-Werte wird in Grafik 5.8 vorgestellt.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Abbildung 5.8: Vergleich verschiedener Thresholding-Werte: links – 1800, mittig – 550, rechts – 200

Im Rahmen der Implementierung wurde festgesetzt, dass eine Detektion aus Amplitudenbildern ohne Preprocessing eine sehr geringe Güte besitzt. Um weiter die globale Pipeline zur Erkennung der Pose erklären und begründen zu können, wird daher die Architektur des verwendeten OpenPose-Netzwerkes im Detail beschrieben.

OpenPose ist ein Vertreter des Learning-basierten, Bottom-up-gerichteten Ansatzes zur Ermittlung der 2D-Pose, wobei das Netzwerk die Keypoints des menschlichen Körpers in Form von Heatmaps ermittelt. Herzu wird eine multi-Stage Architektur angewendet, die aus zwei Sets besteht. Das erste Set ermittelt Konfidenz-Maps (de facto Heatmaps), die die Wahrscheinlichkeit der Lage der Keypoints kennzeichnen. Das zweite Set ermittelt sogenannte Affinitätsfelder der Körperteile (Englisch: Part Affinity Fields – PAF), die die Assoziationen der einzelnen Keypoints zueinander beschreiben [97].

Die Funktionsweise von OpenPose wird in Abbildung 5.9 visualisiert. Die Architektur besteht im ersten Schritt aus der Ermittlung der Anfangs-Feature-Map (bezeichnet als F). Hierzu werden die ersten 10 Layer von [105] angewendet. Diese bestehen aus verschiedenen Filtern mit je einem 3 x 3 Kernel. Anschließend wird die anfängliche Feature-Map den beiden Sets von OpenPose zugeführt. Das Set zur Ermittlung der Affinitätsfelder wird in Abbildung 5.9 blau dargestellt, das Set zu Ermittlung der Heatmaps wird dagegen rot dargestellt. Die beiden Sets bestehend aus Blöcken von je drei Convolutions (Deutsch: Faltung) mit jeweils einem 3 x 3 Kernel, deren Output schließlich zusammenführt wird. Nach dem Durchgang durch die zwei Sets werden die neu detektierten Features der anfänglichen Feature-Map beigeführt. Der Vorgang wird anschließend mit der neuen Feature-Map wiederholt. Die Architektur führt demnach zu einer iterativen Verfeinerung der Detektion über die Anzahl der Iterationen [97].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

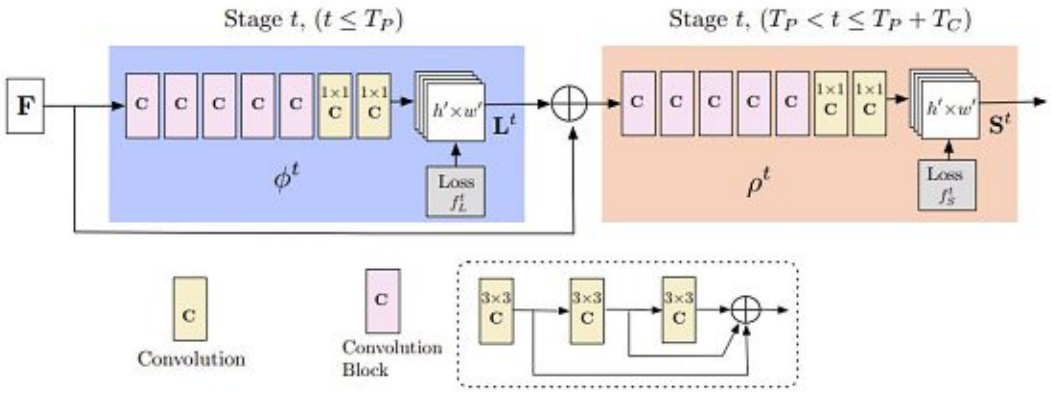



Abbildung 5.9: Multi-stage Architektur von OpenPose [97]

Die Notwendigkeit der Verarbeitung lässt sich wie folgt erklären. Stützend auf der Darstellung des Menschen in Grafik 5.8 sowie der Sensordimension der verwendeten Time of Flight Kameras, die 352 x 287 Pixel beträgt, beträgt die Region in der sich der Mensch befindet, aufgrund der hohen Distanz vom Mensch und der Kamera, nur ungefähr 90 x 75 Pixel. Dies stellt demnach lediglich 7% des Abbildes dar. Dazu findet durch die hohe Anzahl der Convolutions von OpenPose eine laufende Verkleinerung der Features (siehe auch Kapitel 2.3.2) statt. Hierdurch gehen viele der Features im Laufe der Rechnung verloren, was schließlich zu einer schlechten Güte der Erkennung führt. Deshalb wird dieses Problem in dieser Arbeit wie folgt behandelt:

1. Als erstes wird das Amplitudenbild verarbeitet und einem Segmentierungsnetzwerk zugeführt.

2. Die durch das erste Netzwerk generierte Bounding Box wird anschließend zur Extraktion einer Region des Interesses (weiters: ROI – Region Of Interest), in der sich der Mensch befindet, verwendet. Hierdurch wird das Verhältnis der Größe des Menschen zum Input in das Netz erhöht. Dies verbessert die Qualität und der gesamte Zeitbedarf der globalen Pipeline wird dadurch verringert.

Das Preprocessing (die Verarbeitung) des Amplitudenbildes besteht im ersten Schritt aus einer Interpolation auf die dreifache Größe des ursprünglichen Bildes. Dies gewährleistet eine passende Größe der Features, und ermöglicht somit die Detektion mit den prätrainierten Netzwerken. Um eventuelle Fehler der Interpolation zu beheben, insbesondere weiche Kanten, die durch mangelhaften Kontrast bei der Interpolation entstehen können, wird das Bild anschließend mit einem Schärfungskernel (dargestellt in 5.1) verarbeitet.

$$K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{5.1}$$

Die Extraktion von ROI (d.h. den Menschen) wurde in dieser Arbeit durch eine Bounding Box generiert mittels Mask R-CNN implementiert. Mask R-CNN ist eine Erweiterung von Faster R-CNN zur semantischen Segmentierung 2D-Daten [96]. Der Output



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

des Netzwerkes ist eine pixelweise Klassifizierung von Objekten, wodurch eine Maske entsteht (siehe auch Grafik 5.10). Die Maske wird weiter verwendet, um die Punktwolke der Szene um die Figur des Menschen zu bereinigen. Dies reduziert den Clutter sowie die Anzahl der Punkte und somit auch den Rechenaufwand des Teilsystems „Bauteil“.

Mask R-CNN baut auf dem Region-basierten Ansatz von Faster R-CNN, der in zwei Schritten erfolgt. Als erstes wird ein Region Proposal Network (weiters: RPN) angewendet. RPN nimmt ein Bild als Input und erkennt Features, anhand deren Regionen (de facto Bounding Boxes) für Objekte vorgeschlagen werden [106]. Der zweite Schritt besteht aus der Extraktion der Features aus der vorgeschlagenen Region. Darauf basierend erfolgt eine Regression der Bounding Box sowie eine Klassifikation des Objektes und die Ermittlung der Maske. Die durch Mask R-CNN ermittelte Bounding Box wird weiterhin vor dem Input in das Netzwerk zur Ermittlung der Pose gestreckt, um zu sichern, dass die gesamte Kontur und die unmittelbarste Umgebung des Menschen sich im Abbild befinden. Die Outputs von Mask R-CNN sind in der Abbildung 5.10 dargestellt. Die dadurch Segmentierte Punktwolke wird zur Visualisierung der Pose in 3D verwendet und ist dementsprechend in Grafik 5.13 visualisiert [96].

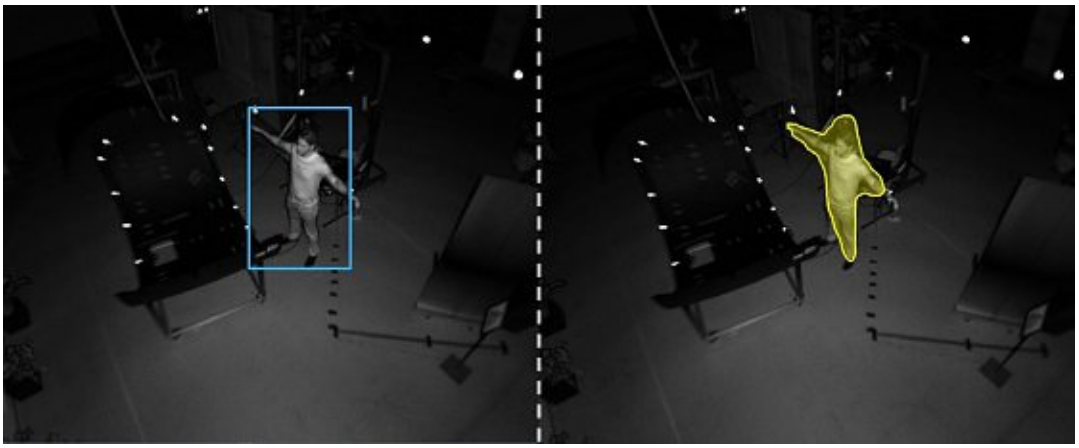


Abbildung 5.10: Ergebnisse von Mask R-CNN: links die ermittelte Bounding Box, rechts die Maske

Nach der Streckung und Extraktion der ROI wird dieses Abbild dem OpenPose-Netzwerk zugeführt. Zur Erkennung der Pose wird ein prätrainiertes Modell, basierend auf dem „COCO“ Datensatz, verwendet [107]. Dieses Modell besteht aus 18 Keypoints, die links in Abbildung 5.11 dargestellt sind. Zwecks einer besseren Veranschaulichung wird in dieser Arbeit die rechte Hand grün visualisiert, um die Asymmetrie des menschlichen Körpers in 3D besser darstellen zu können.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

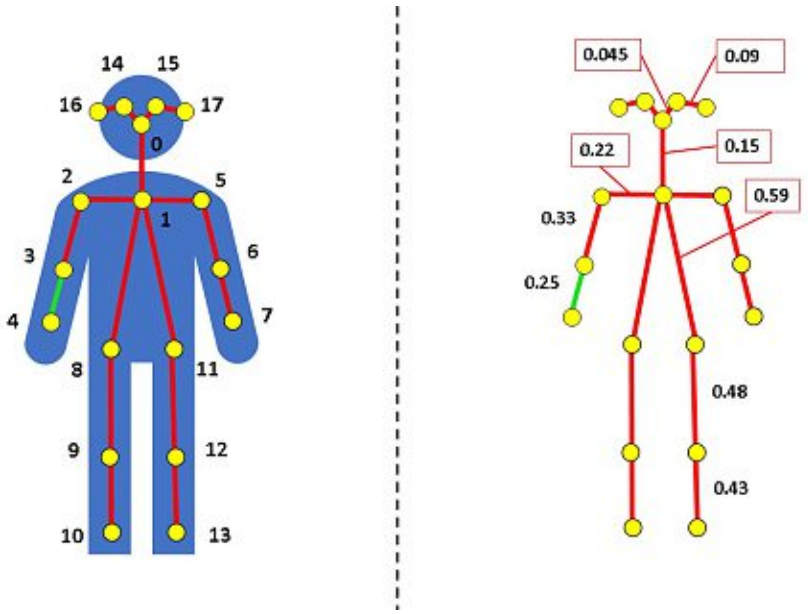



Abbildung 5.11: COCO-Modell: Übersicht der Keypoints (links), abgeleitetes räumliches Modell der Maßlängen des menschlichen Körpers (rechts)

Der Output des Netzwerkes erfolgt in Form von Heatmaps. Zur weiteren Berechnung wird hierbei der Pixel mit der höchsten Wahrscheinlichkeit der Heatmap verwendet. Das Resultat der zweidimensionalen Erkennung wird in Grafik 5.12 rechts veranschaulicht. Links ist als Vergleichsbasis das Ergebnis ohne die Extraktion der ROI durch das erste Netzwerk (Mask R-CNN) dargestellt.

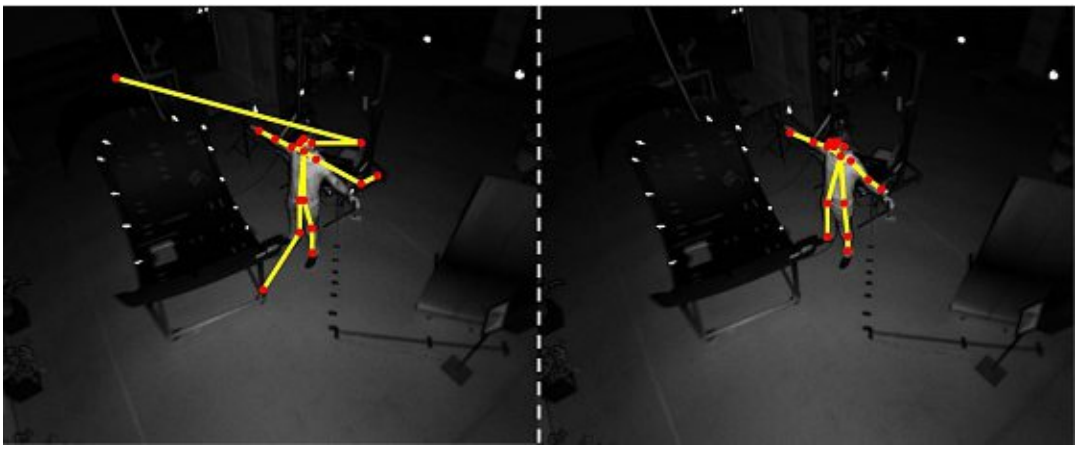


Abbildung 5.12: Vergleich der zweidimensionalen Pose: links ohne das Segmentierungsnetzwerk, rechts mit dem Segmentierungsnetzwerk

Aufbauend auf der zweidimensionalen Pose erfolgt die Projektion in die noch strukturierte Punktwolke. Hierzu werden die ermittelten Koordinaten der Pixel der Keypoints zunächst auf die ursprüngliche Größe des Abbildes (d.h. 352 x 287) normalisiert. An-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

schließend werden entsprechende Keypoints hierarchisch gepaart, wodurch eine Baumstruktur der Gliedmaße bzw. Körperteile gebildet werden. Ausgehend von dem Keypoint 1 (siehe Abbildung 5.11), der den Anfangsknoten der Hierarchie darstellt, erfolgt anschließend eine Validierung der Erkennung in 3D. Diese wird unter Einbeziehung eines aus DIN 33402-2 [108] abgeleiteten Modells des menschlichen Körpers, das die Maße zwischen den einzelnen Keypoints enthält (siehe auch Abbildung 5.11 rechts), durchgeführt.

Zur Validierung und Anpassung werden zuerst die Koordinaten des Anfangspunktes nach der Hierarchie entnommen. Für jeden Endpunkt des Punktpaares wird ein Patch von 16 am nächsten benachbarten Pixeln zu der normalisierten Lage des Keypoints zur Berechnung der Distanz der zwei Keypoints hereingezogen. Wenn die berechnete Distanz in den Bereich des menschlichen Modelles fällt, wird der Punkt angenommen und die 3D-Koordinaten des Punktes in den hierarchischen Baum gespeichert. Wird der Punkt abgelehnt, wird der ganze Patch durchgescannt, bis ein Punkt gefunden wird, der dem Modell entspricht. Wird trotzdem kein passender Punkt gefunden, so erlischt der Knoten und der zugehörige Ast der Hierarchie. Das Vorgehen wird detaillierter in dem unteren Abschnitt des Programmes beschrieben. Praktisch bedeutet die Validierung, dass Keypoints der Körperteile mit einer schlechten Güte der Erkennung womöglich angepasst oder nicht angenommen werden. Dies betrifft insbesondere die Beine (siehe auch Abbildungen 5.13 und 5.15), beispielsweise wenn die Hüfte nicht erkannt wird. Der Grund hierfür sind vor allem Okklusionen und perspektivische Varianz der Posen.

**Algorithmus 1** Validierung und Anpassung der 3D-Projektion der Keypoints

```
1:  for Punktpaar = 1, 2, ... 18 do
2:      Pixel-Koordinaten des Anfangspunkts (StartP) aus Hierarchiebaum laden
3:      Pixel-Koordinaten des Endpunkts (EndP) als Hypothese laden
4:      for Nächster Nachbar von EndP = 1, 2, ... 16 do
5:          3D-Koordinaten über StartP und EndP abrufen
6:          Berechne die 3D-Distanz zwischen StartP und EndP
7:          if Validierung(Distanz, Modell) = true then
8:              Keypoint angenommen
9:              Keypoint im Hierarchiebaum speichern
10:         else
11:             continue
12:         end if
13:     end for
14:     if Keypoint nicht angenommen then
15:         Keypoint erlischt
16:     end if
17: end for
```



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Das Ergebnis der Erkennung wird zunächst zur besseren Visualisierung anhand der segmentierten Punktwolke des Menschen, die durch Mask R-CNN ermittelt wurde, veranschaulicht (Abbildung 5.13). Das Endresultat ohne die Punktwolke des Menschen und inklusive der Erkennung der Pose des Bauteils wird am Ende von Kapitel 5.6.1 vorgestellt.

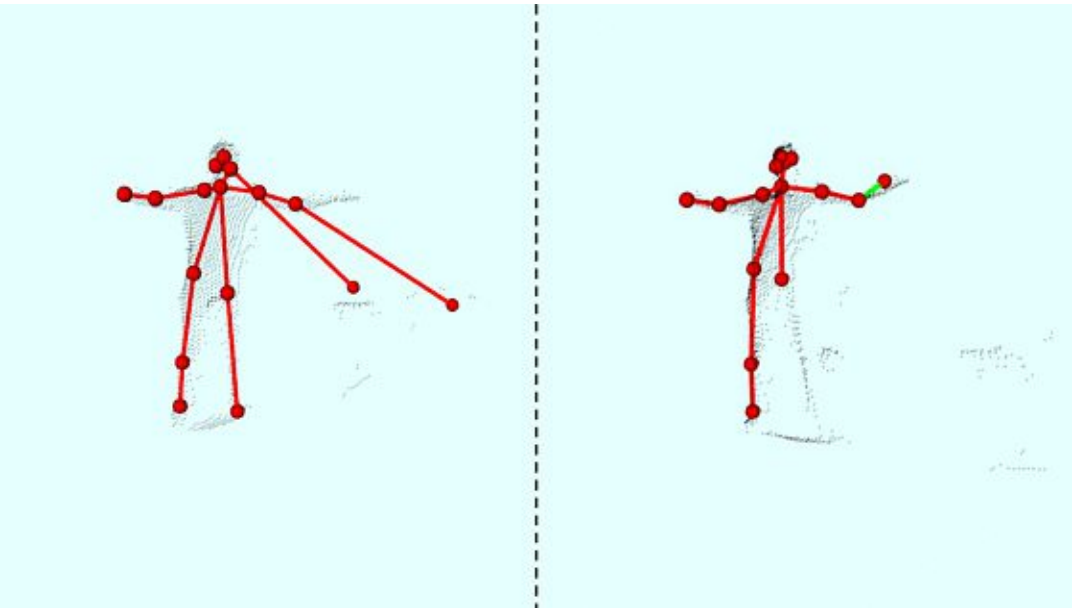


Abbildung 5.13: Veranschaulichung der Pose anhand der segmentierten Punktwolke: 3D-Projektion ohne Anpassung und Verifizierung (links) Projektion der 2D-Pose mit Anpassung und Verifizierung

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

### 5.6.1 Konzept zur Mensch-Maschine Interaktion

Aufbauend auf der Einführung in Mensch-Maschine Interaktion (Kapitel 2.1.1) wird nun die Problematik der Vision-basierten Interaktion über Körpergesten näher beschrieben. Die Vision-basierte Mensch-Maschine Interaktion richtet sich an die Erkennung der menschlichen Aktivität (weiters: HAR, Englisch für Human Action Recognition), dementsprechend wird die Vision-basierte Mensch-Maschine Interaktion, ähnlich wie HAR, als eine Klassifikation einer Sequenz von N Aufnahmen zu einer Klasse der Aktivität beschrieben. HAR wird verallgemeinert durch die räumliche sowie die zeitliche Dimension gekennzeichnet. Dies impliziert eine Extraktion räumlicher Features aus den Aufnahmen und eine anschließende Integration in ein zeitliches Modell. Zur räumlichen Darstellung können dabei sowohl RGB-Daten, Depth-maps als auch Skeletons herangezogen werden. HAR wird, ähnlich wie die Erkennung der Pose, unterteilt in (in Anlehnung an [109][110]):

- Template-basierte Ansätze bilden im ersten Schritt räumlich-zeitliche Features, die anschließend mit Vorlagen verglichen werden.
- Generative Ansätze verwenden ein statistisches Modell zur Klassifikation der Aktivität.
- Diskriminierende Ansätze stützen in der Regel auf eingelernten ML-Modellen zur direkten Klassifikation der Aktivität.

Für die Implementierung der Mensch-Maschine Interaktion wurde in dieser Arbeit der Ansatz über Skeletons zur Erkennung von Gesten der Arme adoptiert. Begründet wird dies durch die perspektivische Invarianz des 3D-Skeletons und der Einfachheit der Implementierung basierend auf der bereits ermittelten Pose. Inspiriert durch [111] wurden Winkel zwischen den einzelnen Körperteilen zur Klassifizierung der Geste hereingezogen. Im Vergleich zu [111] wird das Problem der Mensch-Maschine Interaktion allerdings als Erkennung einer statischen Geste betrachtet. Dies reduziert das Problem der zeitlichen Klassifikation der visuellen Daten signifikant, da eine statische Geste die zeitliche Aufrechthaltung bestimmter Winkel zwischen den Körperteilen voraussetzt. Dementsprechend wird in dieser Arbeit das Konzept zur Vision-basierten Mensch-Maschine Interaktion anhand einer Geste (dargestellt in Grafik 5.14) vorgestellt.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

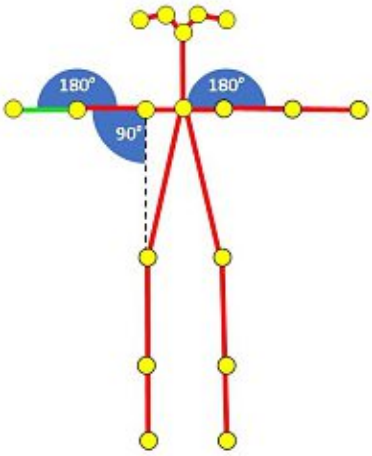



Abbildung 5.14: Veranschaulichung der Geste und der zugehörigen Winkel

Zur Definition einer statischen und symmetrischen Geste über Arme müssen lediglich drei Winkel herangezogen werden. Hierzu werden in einem ersten Schritt Vektoren zwischen den Keypoints der Schulter und Hüften, Schulter und Ellbogen sowie Ellbogen und Handgelenken berechnet. Im Anschluss daran werden für die Vektoren Winkel ermittelt, die in der Abbildung 5.14 visualisiert wurden. Schließlich werden die Winkel mit den zuvor gespeicherten Werten verglichen. Die Berechnung erfolgt für den linken als auch den rechten Arm. Fallen die berechneten Winkel in den Bereich der Winkel der Geste, und wurde die statische Geste über eine voreingestellte Zeit gehalten, so wird die Geste als erkannt bezeichnet. Dies wird anhand zweier Posen in der Abbildungen 5.15 und 5.16 dargestellt, wobei im ersten Fall die Geste richtig klassifiziert wird. Die zweite Abbildung stellt eine Pose beim Arbeiten am Bauteil dar, diese Pose wurde nicht eintrainiert und wird daher richtig abgelehnt.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

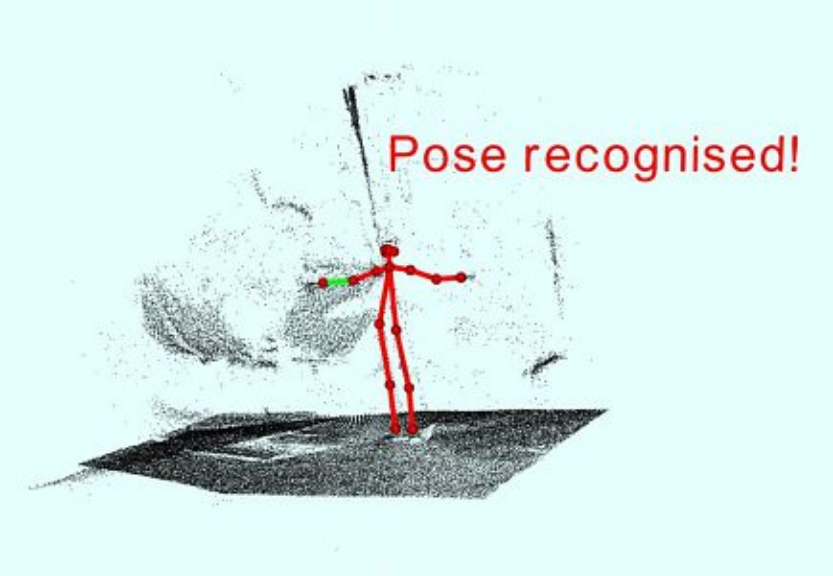



Abbildung 5.15: Nicht-okkludierte, vollständige Darstellung des Skelets und Erkennung der Geste

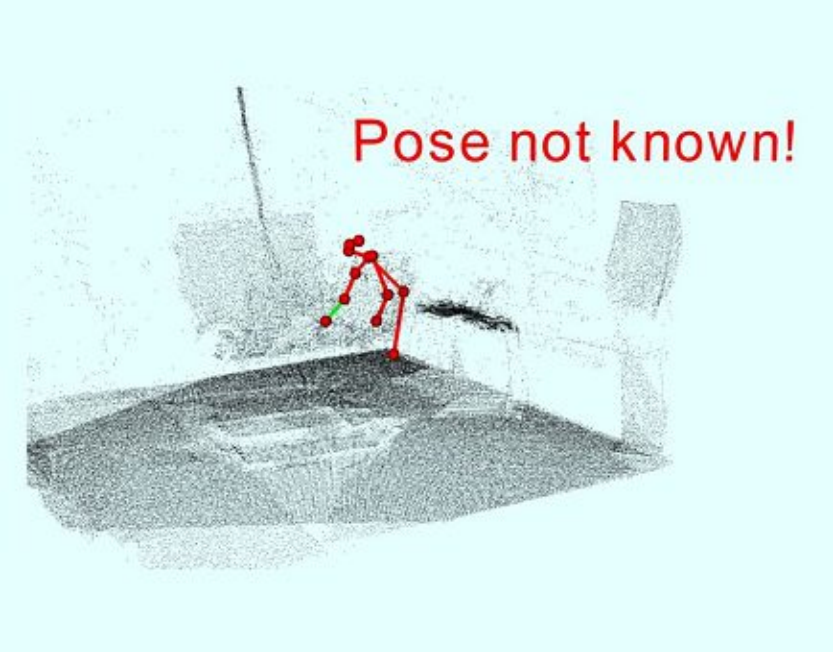



Abbildung 5.16: Abgelehnte Geste („Arbeiten am Werkstück“). Okklusion durch das Werkstück führt zur Nicht-Erkennung der Beine



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## 5.7 Implementierung des Teilsystems Bauteil

Aufbauend auf der Beschreibung der Anforderungen und der Konzeption der Architektur des Systems erfolgt in diesem Kapitel die Beschreibung der Implementierung des Teilsystems „Bauteil“. Zunächst wird die grobe Datenpipeline in Grafik 5.17 vorgestellt. Das Teilsystem besteht aus vier Kernkomponenten und basiert auf der Punktwolke der Szene, die durch die Verschmelzung der drei Aufnahmen des Teilsystem „Mensch“ ermittelt wird. Dies bedeutet, in dieser Punktwolke ist der Mensch nicht mehr explizit dargestellt, da die Silhouette des Menschen bereits aussegmentiert wurde. Die Pipeline beginnt mit dem Preprocessing der Punktwolke. Das Ziel ist die Qualität zu verbessern und die Datengröße zu verringern. Weiters erfolgt die Erkennung des Bauteils durch die Einbeziehung des CAD Modells. Abschließend wird die Pose verfeinert und verifiziert.

Abbildung 5.17: Übersicht der globalen Datenpipeline des Teilsystem „Bauteil“

### 5.7.1 Preprocessing

Preprocessing der Daten zur Ermittlung der Pose des Bauteils wird in mehrere Funktionale Blöcke aufgeteilt, die in Grafik 5.18 ersichtlich sind und in Rahmen dieses Kapitels ausführlich vorgestellt werden.

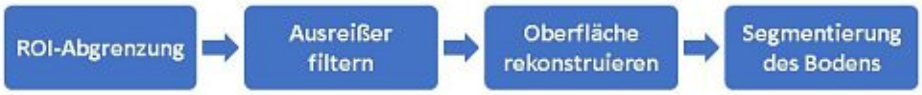



Abbildung 5.18: Datenpipeline von Preprocessing

Im ersten Schritt werden im Rahmen des Preprocessing die drei separaten Punktwolken, ermittelt durch das Teilsystem Bauteil, zu einer Punktwolke der Szene verschmolzen. Im Anschluss daran erfolgt die Abgrenzung der ROI (Region Of Interest). Dies hat das Ziel die Dimensionalität der Daten bei gleichzeitiger Beibehaltung aller Objekte zu reduzieren, was einerseits die Rechenzeiten verkürzt sowie den potenziellen Lösungsraum verkleinert. Die ROI richtet sich nach Grafik 5.2 und verfügt über folgende Abmessungen:

- $X = [-2.2m, 2.2m]$
- $Y = [-2.2m, 2.2m]$
- $Z = [1m, 5m]$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Diese Abmessungen richten sich an den voreingestellten Koordinatenursprung, der etwa in der Mitte des Bauteils liegt. Durch die Abgrenzung der ROI wird die Größe der Punktwolke auf etwa 17% der Originalgröße reduziert (siehe auch Abbildung 5.19).

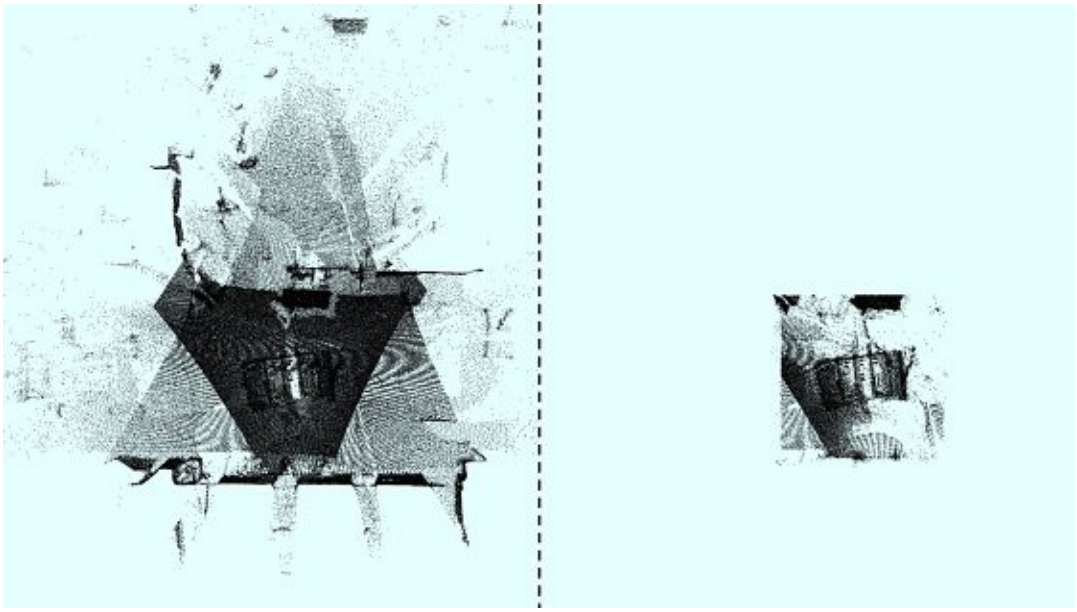


Abbildung 5.19: Darstellung der Verschmelzung der rohen Punktwolken der drei Kameras (links) und die abgegrenzte ROI (rechts)

Punktwolken sind durch die begrenzte Fähigkeit der Sensoren Realität abzubilden mit Fehlern behaftet. Diese Fehler tauchen bei der Analyse der Daten als undefinierte Dateneinträge sowie Rauschen und Ausreißer auf und können die Ergebnisse beeinflussen, weshalb ein Filtern dieser Artefakte im Rahmen des Preprocessing durchgeführt wird. Ein guter Filteralgorithmus soll dabei einen Kompromiss zwischen Filterung und der Beibehaltung der Aussagekraft der Merkmale der Punktwolke darstellen. Gängige Methoden zum Filtern der Ausreißer basieren auf statistischen Verfahren, Analyse der Nachbarschaft, Projektion auf ein Modell etc. (vgl. [112].

Im Rahmen dieser Arbeit wird zur Entfernung der Ausreißer (Englisch: Denoising bzw. „Entrauschen“) der Punktwolke die statistische Methode stammend aus [113] angewendet. Hierzu wurde für eine Nachbarschaft, die voreingestellter Anzahl der Punkte besteht, die mittlere Distanz zwischen den einzelnen Punkten sowie deren Standardabweichung berechnet (siehe 5.2). Anschließend lassen sich alle Punkte, die das Kriterium 5.2 nicht erfüllen, aus der Punktwolke extrahieren. Das Ergebnis nach Denoising ist in Grafik 5.20 sichtbar.

$$\|p_i, p_j\| = |\mu \pm \alpha \cdot \sigma| \tag{5.2}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

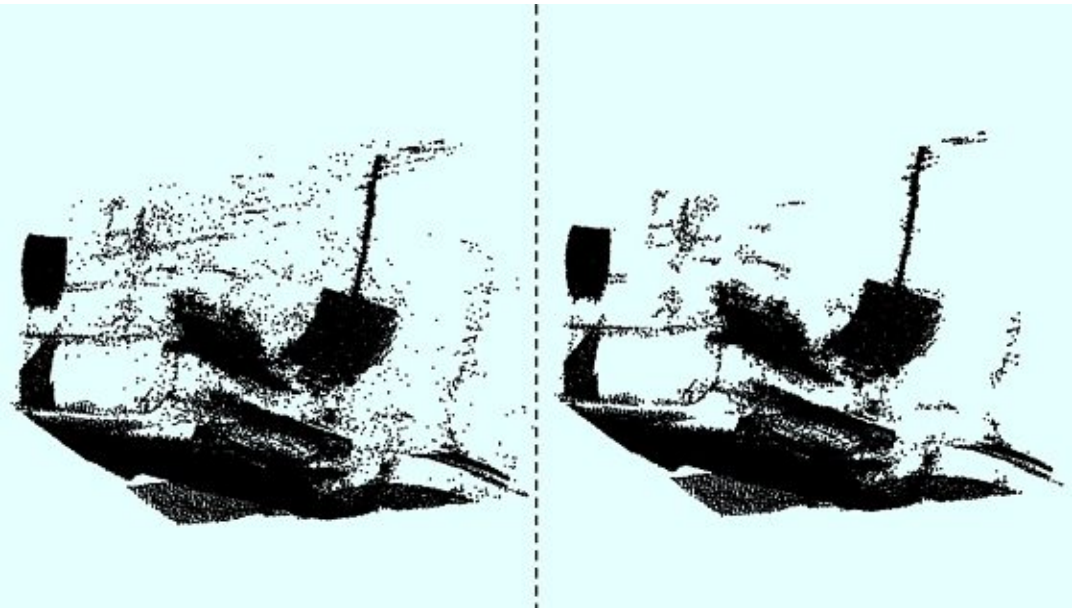


Abbildung 5.20: Veranschaulichung von Denoising (Darstellung bei dreifacher Punktgröße): Ausreißer (links) und gefilterte Punktwolke (rechts)

Trotz der Filterung der Artefakte gilt, dass die aufgenommene Oberfläche der Punktwolke durch Fehler behaftet ist. Zu diesen Fehlern zählen insbesondere Ausreißer, die durch Denoising nicht extrahiert wurden, Abweichungen durch die Verschmelzung der Aufnahmen der drei Kameras sowie Fehler durch abweichende Messungen der Kameras. Dementsprechend richtet sich die Aufgabe der Oberflächenrekonstruktion an die Wiederherstellung der Geometrie, um eine möglichst gute Grundlage für die Erkennung zu schaffen. Dies wird insbesondere in Hinsicht auf die Berechnung der Oberflächennormalen berücksichtigt (vgl. Kapitel 2.3.1), deren Berechnung im Kapitel 5.7.2 näher beschrieben wird (vgl. [114]).

Zur Oberflächenrekonstruktion wurden in dieser Arbeit Moving-Least Squares (weiters: MLS) verwendet. Die Kernannahme von MLS basiert auf der Annäherung der Oberfläche durch mathematische Funktionen. Die mathematische Definition der Oberfläche erfolgt für jeden Punkt der Punktwolke separat und verläuft als Regression in zwei Schritten. Im ersten Schritt wird für jeden Punkt eine lokale Ebene und eine Normale, die den Punkt schneidet, definiert. Dieser Schritt erfolgt iterativ, bis sich eine zufriedenstellende Konvergenz der Regression der Ebene und der Punkte der Nachbarschaft aufstellt. Der zweite Schritt besteht anschließend aus der lokalen Approximation der Nachbarschaft durch ein Polynom. Dies wird ebenfalls als eine Regression modelliert und der Vorgang wird wiederholt, bis die Konvergenz der Koeffizienten des Polynoms das voreingestellte Abbruchkriterium nicht erreicht. Nach der Berechnung der Koeffizienten kann der Punkt, unter Einbeziehung des Polynoms. entlang der Normale auf die Oberfläche, die zuvor durch das Polynom geschätzt wurde, projiziert werden [115].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die Rekonstruktion der Oberfläche führt zu einer gleichmäßigeren, stetigeren Verteilung der Punkte und somit auch der Oberflächennormalen, was schließlich zur besseren Erkennungsrate führt. Die Ergebnisse sind in Grafik 5.21 sichtbar.

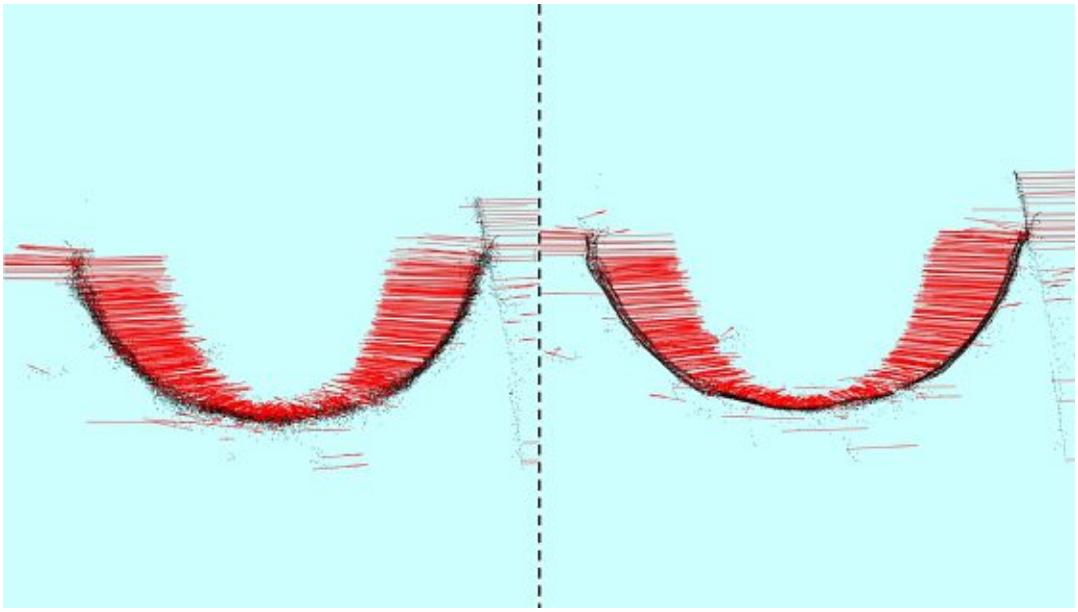


Abbildung 5.21: Darstellung von MLS anhand des segmentierten Bauteils (Anzahl der Punkte auf 20% reduziert): rohe Oberfläche (schwarz) samt Oberflächennormalen dargestellt in rot (links) und rekonstruierte Oberfläche mit gleichmäßigen Normalen (rechts)

Zwecks einer weiteren Dimensionsreduktion erfolgt eine Abstraktion und Extraktion von Daten, die nicht direkt von Interesse sind. Hierzu zählt ebenfalls die Bodenoberfläche, die dementsprechend aus der Szene aussegmentiert werden kann. Die Implementierung der Segmentierung erfolgt mittels RANSAC (Random Sample Consensus), die auf der mathematischen Beschreibung bestimmter Formen stützt und diese zur Extraktion verwendet.

RANSAC ist eine effiziente, generalisierte und gegenüber Rauschen robuste Methode zur Extraktion dreidimensionaler geometrischer Formen wie Sphären, Zylinder und Ebenen. Das Verfahren basiert auf einer iterativen Schleife der Generierung von Hypothesen und einer anschließenden Verifizierung dieser. Zur Generierung der Hypothese werden geometrische Primitiven aus den Daten ermittelt. Hierzu wird dabei die minimale Anzahl der Punkte, die zur Definition benötigt werden, herangezogen (zur Veranschaulichung: eine Ebene ist durch mindestens drei Punkte definiert). Die ermittelte Primitive stellt weiterhin eine Hypothese für den nächsten Schritt dar, der die Güte der Beschreibung gegenüber allen Punkten der Punktwolke verifiziert. Nach einer bestimmten Anzahl der Iterationen wird diejenige Approximation (bzw. diejenige Primitive), die für die meisten Punkte der Szene signifikant ist, als verifiziert angenommen. Die Methode liefert



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

dadurch alle Indizes der Punktwolke, die für die Hypothese tauglich sind. Im Anschluss daran lassen sich diese Indizes aus der Szene extrahieren (vgl. [116][117]).

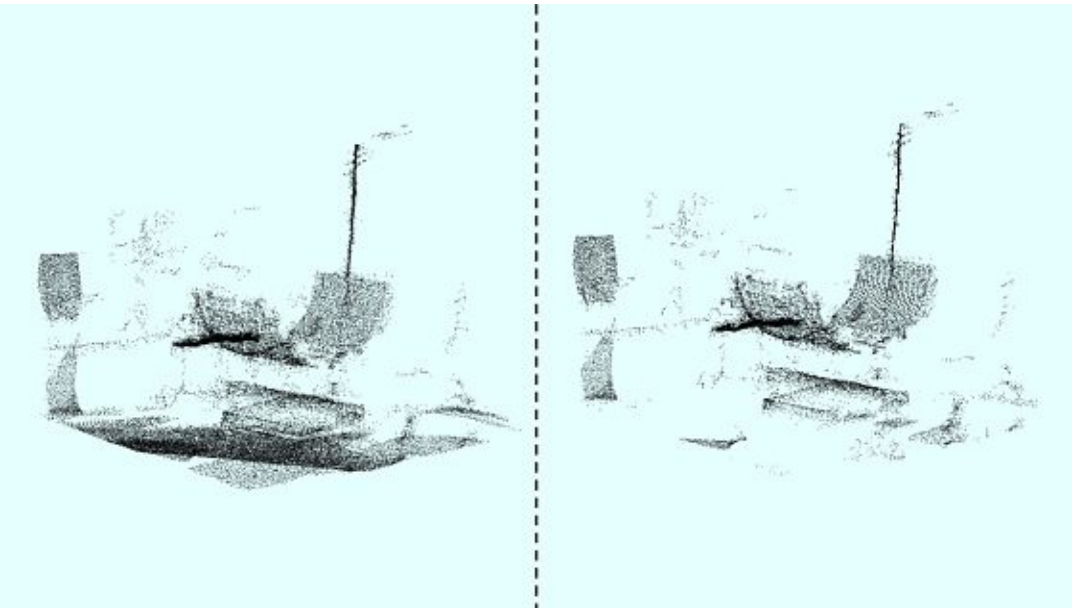


Abbildung 5.22: Veranschaulichung der Extraktion der Punktwolke des Bodens: Punktwolke mit Boden (links) und Ergebnis der Szene nach Preprocessing (rechts)

### 5.7.2 Erkennung, Verfeinerung und Verifizierung der Pose

Basierend auf den durch Preprocessing verarbeiteten Daten erfolgt die Erkennung der Pose des Bauteils. Die Datenpipeline zur Erkennung der Pose lässt sich in vier Blöcke aufteilen. Die Erkennung der Pose erfolgt unter Einbeziehung eines CAD-Modells, dessen Oberfläche direkt in der CAD-Software uniform in Punkte abgetastet wurde. Somit wurde das Modell in eine Punktwolke umgewandelt, die die Basis für die Erkennung darstellt. Die funktionalen Blöcke werden in der Abbildung 5.23 visualisiert, dabei stellt der blaue Pfeil den Datenflow der Punktwolke der Szene dar, der Datenflow des CAD-Modells wird grün modelliert.

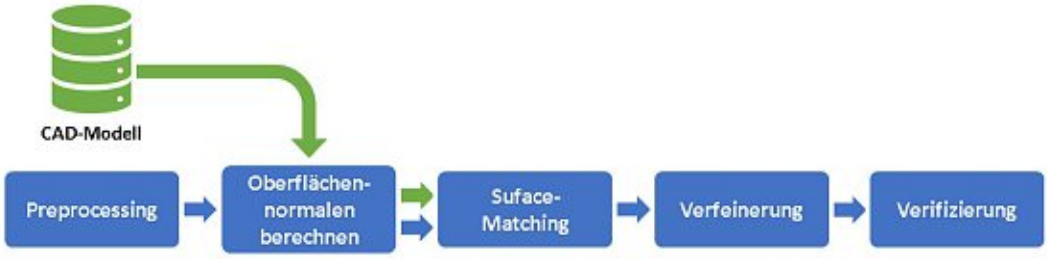



Abbildung 5.23: Datenpipeline zur Erkennung der Pose des Bauteils



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Zunächst erfolgt sowohl für die Punktwolke der Szene als auch für das CAD-Modell die Ermittlung der Oberflächennormalen. Die verwendete Methode richtet sich nach [113] und berechnet die Oberflächennormalen basierend auf einem planaren Modell der Nachbarschaft rund um den betrachteten Punkt. Hierzu wird zuerst eine PCA (Principal Component Analysis) für die Nachbarschaft berechnet, die planare Oberfläche wird anschließend an den kleinsten Eigenvektor gelegt, der ebenfalls die Richtung der Oberflächennormale festlegt. Der wesentliche Nachteil dieser Methode besteht allerdings in der mangelhaften Konsistenz der Ausrichtung der Normalen (vgl. [113]). Dies wird in Grafik 5.24 veranschaulicht. Eine mögliche Behebung der Nicht-Konsistenz kann über die Kenntnis der Perspektive erfolgen, in dem alle Normalen hin zu dem Aussichtspunkt (weiters: Viewpoint) gedreht werden (d.h. zu der Kamera). Dies ist bei dem Use-Case allerdings nicht möglich, da die Kenntnis des Viewpoints durch die Verschmelzung der drei Aufnahmen verloren geht. Deshalb wird zur Behebung der Orientierung die Z-Komponente der Oberflächennormalen des Bauteils verwendet, die Beschreibung erfolgt im unteren Code-Ausschnitt. Das Ergebnis der Normalenberechnung wurde in Grafik 5.21 veranschaulicht.

**Algorithmus 2** Orientierungskonsistenz der Oberflächennormalen

```
for Surface Normals = 1, 2, ... do
    if Z − Richtung < 0 then
        drehe alle Richtungen der Oberflächennormale
    end if
end for
```

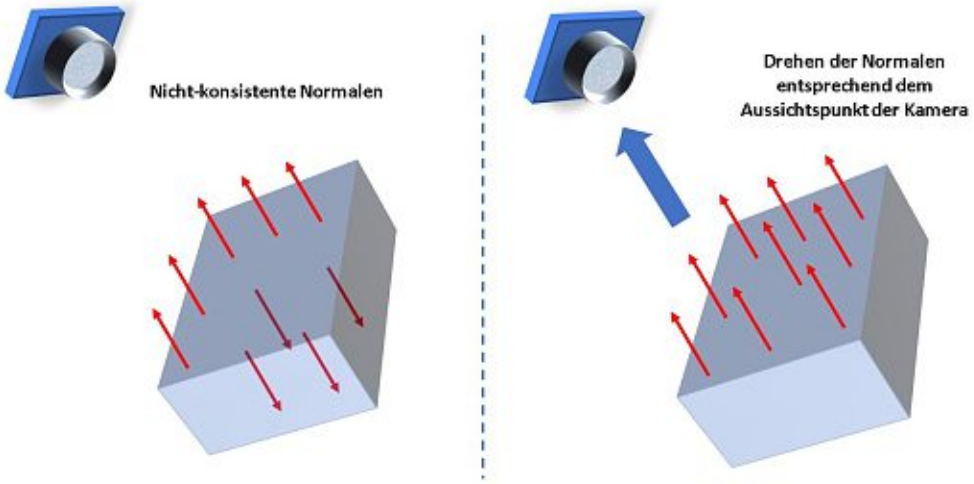



Abbildung 5.24: Veranschaulichung der Normalenberechnung: nicht-konsistente Ausrichtung der Oberflächennormalen (links) und übliche Behebung durch die Kenntnis des Aussichtspunkts (rechts)

Liegen die Oberflächennormalen der Szene und des Modells fest, so können Deskriptoren für den Vergleich der Szene und des Modells berechnet werden. In dieser Arbeit



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

werden zur Erkennung der Pose aus der Szene Point Pair Features (weiters: PPF) verwendet. PPF bilden einen globalen Deskriptor und basieren auf der relativen Position und Orientierung zwei beliebiger Punkte. Der wesentliche Vorteil der Methode besteht in der großen Robustheit gegenüber Sparse-Punktwolken. Der Nachteil dieser Methode besteht in der direkten Abhängigkeit von der Qualität der Oberflächennormalen (vgl. [95]). Die Implementierung von PPF in dieser Arbeit basiert direkt auf der Methode von Birdal [118]. Diese Methode macht die Robustheit von PPF gegenüber Sparse-Punktwolken zunutze und verwendet zusätzlich dichteabhängiges Downsampling (Deutsch: Abtasten), um die Anzahl der zu untersuchenden Datenpunkte zu verringern, und dadurch den Zeitaufwand zu reduzieren. Hierdurch können geringe Erkennungszeiten erreicht werden. Der Outline der Methode ist in Grafik 5.25 sichtbar.

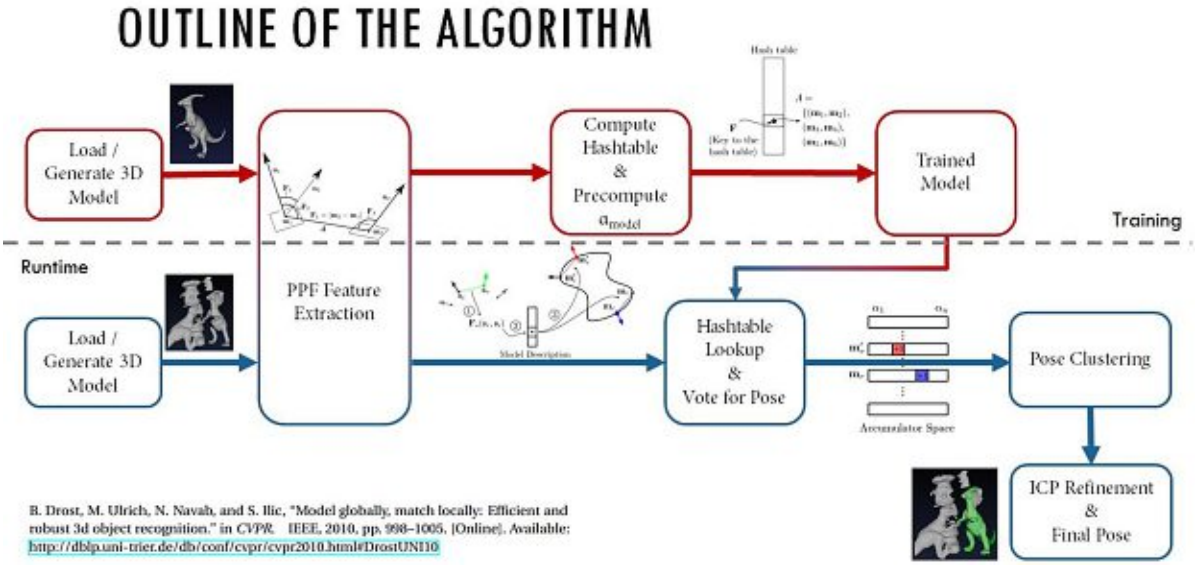



Abbildung 5.25: Outline der Methode von [118] zur Erkennung der Pose mittels PPF und anschließender Verfeinerung

Point Pair Features sind asymmetrische Features, welche aus zwei Punkten $p_1$ und $p_2$ und deren korrespondierenden Oberflächennormalen $n_1$ und $n_2$ berechnet werden. Dementsprechend wird das Feature für diese zwei Punkte definiert als [95]:

$$F(p_1, p_2) = (\|d\|, < (n_1, d), < (n_2, d), < (n_1, n_2)) \tag{5.3}$$

mit $d$ gleich dem Vektor von $p_1$ nach $p_2$ und $< (n_1, n_2)$ als der Winkel zwischen zwei Vektoren. Zuerst werden die Features $F(m_i, m_j)$ für alle Punktpaare des Modells berechnet und gespeichert. Anhand der Werte aller Features $F(m_i, m_j)$ erfolgt anschließend eine Gruppierung und Indexierung aller ähnlichen Punktpaare in einer Hashtabelle. Danach werden die Features $F(s_i, s_j)$ der Punktwolke der Szene berechnet. Im Anschluss daran werden gemäß dem Wert der Feature $F(s_i, s_j)$ nach ähnlichen Punktpaaren mit $F(m_i, m_j)$ des Modells, die zuvor in der Hashtabelle gespeichert wurden, gesucht [95].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Um die Pose anhand eines gefundenen Punktes $s_r$ der Szene korrespondierend zu dem Punkt $m_r$ des Modells zu bestimmen, werden die Oberflächennormalen der Punkte $m_r$ und $s_r$ ausgerichtet. Dies erfolgt über eine Transformationsmatrix, die die Normale des Punktes auf die $X$-Achse ausrichtet. Da das Punktepaar $s_r, s_i$ und das Punktepaar $m_r, m_i$ aufgrund der ähnlichen Werte ihrer Features ähnliche Vektoren besitzen, muss nach der Ausrichtung entlang der $X$-Achse lediglich ein Winkel bestimmt werden, der die Rotation zwischen dem Punktpaar des Modells und dem Punktpaar der Szene bestimmt. Die Berechnung wird in Abbildung 5.26 veranschaulicht, die Gleichung der Korrespondenz zwischen den Punkten $m_i$ und $s_i$ lautet demnach:

$$s_i = T_s^{-1} \cdot R_x(\alpha) \cdot T_m \cdot m_i \tag{5.4}$$

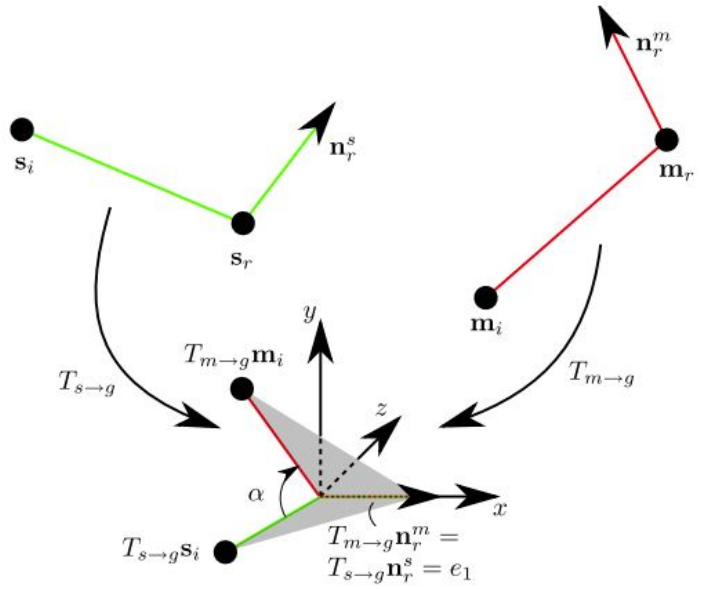



Abbildung 5.26: Veranschaulichung der Suche nach korrespondierenden Punktpaaren [95]

Eine vernünftige Pose erfordert eine Maximierung der Anzahl der korrespondierenden Punkte zwischen dem Modell und der Szene. Deshalb wird die zuvor beschriebene Berechnung von $\alpha$ für alle Punkte $s_i$ rund um den Punkt $s_r$ wiederholt. Dieselbe Berechnung erfolgt für alle Punktpaare $m_r, m_i$. Ob eine Korrespondenz zwischen dem Modell und dem Teil der Szene beschrieben durch $s_r$ besteht, ergibt sich durch die Akkumulation der berechneten Werte von $\alpha$. Hier gilt, dass die Posen, die die Lage des Modells am besten beschreiben, die höchsten akkumulierten Werte von $\alpha$ aufweisen. Anschließend kann mit $\alpha$ die Transformation für alle Punkte des Modells mittels Gleichung 5.4 berechnet werden. Dies stellt eine Hypothese der Pose dar [95].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die Verfeinerung der Pose erfolgt anschließend über Iterative Closest Point (weiters: ICP). ICP ist ein iteratives Verfahren zur Registrierung der Transformationsmatrix zwischen zwei Punktwolken über eine globale Distanzfunktion. Hierzu wird zuerst für jeden Punkt einer Punktwolke der am nächsten liegende Punkt der anderen Punktwolke bestimmt. Anschließend wird die globale Distanzfunktion, die die mathematische Übereinstimmung der Punktwolken beschreibt, für alle Punkte der Punktwolke minimiert. Diese zwei Schritte werden wiederholt, bis die Übereinstimmung konvergiert (vgl. [119]).

Bei der Methode von Birdal [118] wurde der klassische Algorithmus von ICP angepasst, um eine bessere Performance zu schaffen. Die Unterschiede werden hier kurz beschrieben. Als erstes erfolgt ein Downsampling basierend auf [120]. Dies verwendet die Eigenvektoren zur Bestimmung der Verbreitung der Punkte, wodurch schließlich nur Punkte verwendet werden, die sowohl zur Berechnung der Translation als auch der Rotation beitragen können. Dadurch kann die Instabilität der Konvergenz vermieden werden. Des Weiteren wird Picky-ICP verwendet [119], um das Konvergenzverfahren hierarchisch einzustufen. Hierzu werden Punkte der Punktwolke in eine Hierarchie angeordnet, wobei ICP für jede Stufe der Hierarchie durchgeführt wird, um anschließend zu der nächsten Stufe überzugehen. Dies reduziert die Anzahl der Punkte und führt bei großen Punktwolken zur Reduktion des Zeitbedarfs.

Die verwendete Methode liefert demnach Hypothesen zu den Posen, aus denen abschließend eine Pose, die die Szene am besten beschreibt, angenommen werden soll. Hierzu werden im ersten Schritt die drei besten Posen selektiert. Dies führt zu einer Zeitersparung, da die Verfeinerung der restlichen Hypothesen nicht stattfindet. Aus diesen drei Hypothesen wird anschließend durch die Verifizierung eine Pose angenommen. Die Verifizierung richtet sich an die Methode von Papazov et al. [121]. Die verwendete Methode basiert auf einer Akzeptanzfunktion, die die Überlappung der Hypothese mit der Punktwolke der Szene bewertet. Anschließend werden alle Hypothesen, die eine bestimmte Güte der Überlappung besitzen, miteinander verglichen, um die beste Hypothese zu bestimmen. Hierzu wird ein Konflikt-Graph erstellt, dessen Knoten die Hypothesen darstellen. Bei einer Intersektion zweier Knoten werden die Knoten durch eine Linie verbunden. Um die beste Hypothese zu finden, werden alle Knoten entlang der Linien durchgegangen. Wenn die Nachbarn von den betrachteten Knoten die Pose besser beschreiben können, wird der betrachtete Knoten gelöscht, bis nur noch die beste Hypothese verbleibt. Dementsprechend bildet die verifizierte Pose das Resultat. Die Darstellung des Ergebnisses erfolgt in Abbildung 5.27.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Abbildung 5.27: Erkannte Pose des Bauteils

## 5.8 Evaluierung und Testing

### 5.8.1 Bekannte funktionale Einschränkungen

Dieses Kapitel widmet sich der Beschreibung der funktionalen Grenzen der beiden Teilsysteme, die im Rahmen der Evaluierung entdeckt wurden. Im Falle des Systems zur Erkennung der Pose des Menschen, zählen zu den bekannten Problemen die Erfassung bestimmter Posen sowie starke Okklusion durch äußere Objekte. Beispiele für diese zwei Probleme werden in Abbildung 5.28 veranschaulicht. Der Grund für das Versagen bei bestimmten Posen, wie z.B. Vorbeugen, sind insbesondere die Selbst-Okklusionen der einzelnen Körperteile. Nach der Beschreibung des OpenPose-Netzwerkes können daher bei Okklusionen, gleichgültig ob äußere oder selbst-verursachte, die Konfidenz-Maps sowie die Affinitätsfelder nicht richtig ermittelt werden. Hierbei sei angemerkt, dass dies das Segmentierungsnetzwerk in der Regel nicht beeinflusst. Somit wird die Präsenz des Menschen in der Szene erkannt, allerdings ist in diesen Fällen das Tracking nicht möglich.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

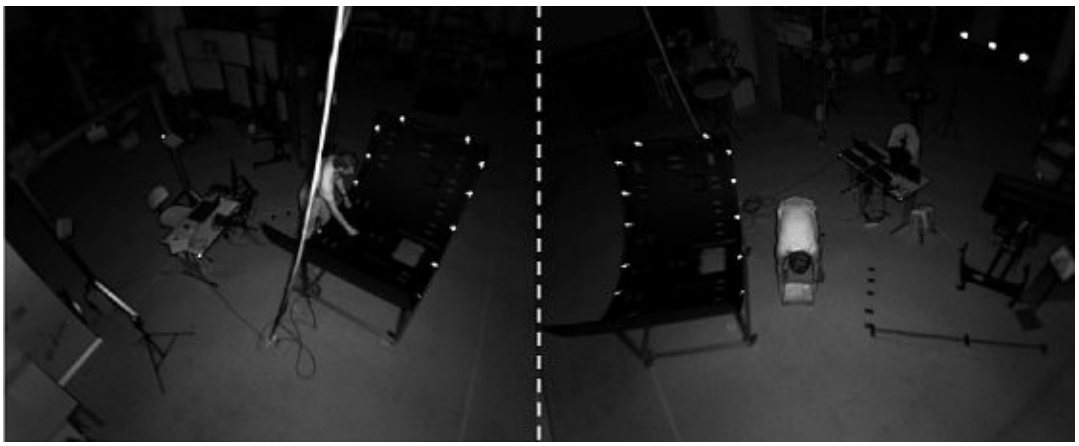


Abbildung 5.28: Schwer erkennbare Posen: starke Okklusion durch andere Objekte betreffend insbesondere Kamera 1 und Kamera 2 (links) und Selbstokklusion beim Heben einer Kiste (rechts)

In Rahmen dieser Arbeit wurde die Erkennung der Pose des Bauteils als erneute Erkennung der Pose mittels PPF aus einzelnen Szenen betrachtet. Dementsprechend fließt die vorherige Pose nicht in die Erkennung ein. Dies bedeutet praktisch, es wurde eine Schleife gebildet, die die Methoden beschrieben im Kapitel 5.7.2 erneut einsetzt. Im Laufe des Testings wurden allerdings Speicherprobleme der verwendeten Methode zur Berechnung von Point Pair Features entdeckt. Hierdurch ist eine Serialisierung de facto instabil. Eine mögliche Behebung der Serialisierungsprobleme der Methode kann über die Implementierung der Funktionen zum Speichern und anschließendem Laden der prätrainierten Hashtabelle, die oft zu Speicherproblemen führt, erfolgen. Zur näheren Beschreibung dieses Ansatzes wird hier auf die offizielle Dokumentation verwiesen [118].

Die Initialerkennung bei dem ersten Durchlauf der Schleife wird allerdings durch die Speicherprobleme nicht betroffen. Dementsprechend wird nun eine Behebung in Form eines Prototyps vorgestellt. Der Ablauf des Prototyps lässt sich in zwei Schritte aufteilen. Als erstes erfolgt die Initialerkennung mittels Point Pair Features. Für die weiteren Durchläufe der Schleife wurde eine robuste Einstellung von ICP implementiert. Demnach ist der Ablauf der Erkennung der Pose des Bauteils identisch zu der Methode vorgestellt von Radkowski [61]. Der Nachteil dieser Implementierung liegt insbesondere in dem hohen Rechenaufwand. Stützend auf der Beschreibung von ICP im Kapitel 5.7.2 wurde die maximale Anzahl der Iterationen auf 20 eingestellt, um den Rechenaufwand und somit die maximale Erkennungszeit zu beschränken. Dies impliziert, dass nach 20 Iterationen das Verfahren abgebrochen wird, gleichgültig ob eine Konvergenz erreicht wurde oder nicht, wodurch die Stabilität der Lösung nicht vollständig garantiert werden kann. Ein Versagen der Erkennung der Pose des Bauteils setzt dementsprechend einen Neustart des Systems voraus, wodurch die Pose mittels Point Pair Features wieder erkannt wird.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

### 5.8.2 Evaluierung auf die nicht-funktionalen Anforderungen

Aufbauend auf den Anforderungen an das wahrnehmende Assistenzsystem erfolgt in diesem Kapitel 5.3 die Evaluierung und die Diskussion der Performance des Systems, wodurch das Potential für eine praktische Implementierung des Konzeptes abgeschätzt werden kann.

Das System wurde in Hinsicht auf das Zeitverhalten getestet, die Resultate werden für die beiden Teilsysteme in Tabellen 5.3 und 5.4 vorgestellt. Dabei wurde zwischen den Bezugsgrößen Einzelaufnahme und Szene (Verschmelzung von drei Einzelaufnahmen zu einer Punktwolke) unterschieden. Als Vergleich werden die Hardware-Parameter gelistet. Die Implementierung sowie das Testing erfolgte auf einem Rechner mit einer *Intel i5-4210M* CPU mit 8 GB RAM. Der Einsatz einer GPU zur Beschleunigung sowie die erwartete Performance auf geeigneter Hardware wird weiters diskutiert.

| ***Block*** | ***Bezug*** | ***Stichprobe*** | ***Mittelwert (Sekunden)*** |
|---|---|---|---|
| Datenaufnahme | Einzelaufnahme | 150 | 0.0047 |
| Preprocessing und Segmentierung | Einzelaufnahme | 20 | 4.6211 |
| 2D Erkennung der Pose | Einzelaufnahme | 20 | 4.3541 |
| 3D Projektion der Pose | Einzelaufnahme | 20 | 0.4623 |
| Erkennung der Geste | Einzelaufnahme | 20 | 0.0037 |

Tabelle 5.3: Übersicht der Zeiten für das Teilsystem **Mensch**

| ***Block*** | ***Bezug*** | ***Stichprobe*** | ***Mittelwert (Sekunden)*** |
|---|---|---|---|
| Preprocessing | Szene | 50 | 1.0961 |
| Normalen-berechnung | Szene | 50 | 0.1210 |
| Erkennung über PPF | Szene | 20 | 0.0647 |
| Verfeinerung und Verifizierung | Szene | 20 | 0.9647 |

Tabelle 5.4: Übersicht der Zeiten für das Teilsystem **Bauteil**

Viele der implementierten Algorithmen bieten das Potential zur Implementierung auf der GPU. Dies würde, in Anlehnung an die Dokumentation der Algorithmen, zu ei-



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

ner signifikanten Beschleunigung führen. Als Richtwerte für gängige Hardware dienen ungefähr 0.2 Sekunden für Mask R-CNN [96] und 0.1 Sekunden für OpenPose [122]. Die Implementierung dieser ist ausgehend von dem bestehenden Code sehr einfach und erfordert lediglich eine geeignete Hardware und Freistellung einer Zeile im Code.

In Hinsicht auf die Ressourcen benötigt die Datenaufnahme und das Teilsystem „Bauteil“ jeweils knapp 200 MB an Arbeitsspeicher, das Teilsystem „Mensch“ dagegen etwa 1 GB. Diese Anforderungen stellen keinesfalls ein Hindernis für die modernen Rechner. Dementsprechend wäre eine Implementierung des Assistenzsystem mittels gängiger Hardware möglich. Mit einem Blick auf die Speicherkapazität bedarf das wahrnehmende System lediglich die Speicherung der Deep Learning-Modelle und des CAD-Modells, deren Datengröße gering ist.

Eine Änderung des Teilsystems „Bauteil“ durch die Erweiterung anderer CAD-Modelle ist ebenfalls möglich. Durch den geringen Zeitbedarf zur Erstellung einer neuen Hashtabelle, was etwa 0.7 Sekunden beträgt, lassen sich neue Modelle schnell einspielen. Dementsprechend ist ebenfalls eine Rekonfiguration des Systems sehr effizient. Simultane Erkennung mehrerer CAD-Modelle ist allerdings nicht möglich.

Zusätzlich bietet auch das Teilsystems „Mensch“ Möglichkeiten zur Erweiterung. Die Auswahl der Methoden für das Teilsystem „Mensch“ erfolgte in Hinsicht auf die Fähigkeit zur Erkennung mehrerer Personen. Die Lösung zur Multi-Personen Erkennung lässt sich relativ einfach umsetzten. Der Grund dafür ist die Integration des Segmentierungsnetzwerks in die Pipeline, wodurch die klassische Bottom-Up Architektur von OpenPose in eine Top-Down Architektur umgewandelt wird. Die einzige Voraussetzung zur Multi-Personen Erkennung besteht dementsprechend in der Zuordnung einer Identität zu der Bounding Box, da die Erkennung der Pose von der Bounding Box ausgeht. Dadurch wird garantiert, dass die Korrespondenzen zwischen den einzelnen Keypoints mehrerer Personen richtig zugeteilt werden. Demnach wäre ein Monitoring mehrerer Arbeitsstationen, solange sie im Blickwinkel der Kameras sind, gleichzeitig möglich. Dies impliziert allerdings eine Steigerung des Rechenaufwandes.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# 6 Zusammenfassung und Diskussion

Diese Arbeit hat sich mit der Umsetzung eines Konzeptes zur Erfassung der Pose des Mitarbeiters, und des Bauteils, zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems befasst. Als Motivation dient hierzu der Bedarf nach einem Monitoring großräumiger Arbeitsbereiche, wie etwa im Falle einer Baustellenmontage. Dies wurde durch die Extraktion relevanter Informationen mittels mehrerer Time of Flight Kameras erreicht, die die dreidimensionale Darstellung des dazwischenliegenden Arbeitsbereiches liefern. Der Fokus dieser Arbeit richtete sich an die Beantwortung folgender Forschungsfragen:

**(F1)** Welche Herausforderungen sind zu bewältigen, um markerloses 3D-Tracking von Mitarbeiter und Bauteil mittels Time of Flight Kameras implementieren zu können?

**(F2)** Wie kann die Erkennung und das Tracking mittels Time of Flight Kameras generierten Daten gestaltet werden?

**(F3)** Welche Kriterien bestehen auf das Tracking und wie leitet sich daraus das Potential von 3D-Tracking mit Punktwolken zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems ab?

Zunächst werden die Forschungsfragen **(F1)** und **(F2)** beantwortet, anschließend erfolgt im Kapitel 6.1 eine Diskussion der Ergebnisse und des Potentials, in Rahmen deren die Frage **(F3)** beantwortet wird.

Die Herausforderungen für markerloses Tracking können in arbeitswissenschaftlichbedingte und technologische Herausforderungen aufgeteilt werden. Zu den arbeitswissenschaftlich bedingten zählen die Wechselwirkung zwischen dem Mitarbeiter und dem Bauteil. Dadurch entstehen Clutter sowie Okklusionen zwischen dem Mitarbeiter und dem Bauteil, weshalb die Segmentierung der relevanten Daten von Interesse zu einer Herausforderung wird. Dies wird durch die Nicht-Verwendung der Farbdaten noch erschwert, da hierdurch ein wichtiger Informationskanal nicht verfügbar ist. Des Weiteren stellt die Erfassung von schnellen Bewegungen des Menschen eine schwierige Aufgabe dar. Die schnellen Bewegungen, zusammen mit dem Problem der Synchronisierung, das bei Time of Flight Kameras durch die Multi-Kamera Interferenz verursacht wird, führt dazu, dass keine stetige dreidimensionale Darstellung des Menschen möglich ist. Als letzte Herausforderung, die durch die Technologie bedingt ist, gilt die Auflösung der Punktwolke. Der Einsatz von Time of Flight Kameras kennzeichnet sich durch eine geringe Dichte der Punktwolke, wodurch die mathematische Beschreibung der Objekte zu einer Herausforderung wird. Dies wird durch die relativ hohe Distanz der Kameras zu der Szene noch verstärkt.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Entsprechend den Herausforderungen bedarf es an robusten Algorithmen zur Erfassung der Pose in 3D, die sich unter gegebenen Umständen durch eine stabile Performance aufzeichnen. Hierzu wurde zunächst eine Literaturrecherche durchgeführt, die basierend auf der generellen Taxonomie der entsprechenden Problematik auf die Findung geeigneter Ansätze zielte. Die Ansätze richten sich verallgemeinert an Algorithmen, deren Verhalten durch das manuelle Feature-Engineering (traditioneller Ansatz) bestimmt wird, sowie Algorithmen, die selbst eine geeignete Methodik zur Beschreibung erlernen können (Machine Learning). Des Weiteren gliedert sich der traditionelle Ansatz zur Erkennung der Pose starrer Körper, wie etwa das Bauteil, in Template-basierten Ansatz, sowie Sparse-Features und Dense-Approach. Die Erkennung der Pose des Menschen kann entweder mithilfe eines Modells (generativer Ansatz) oder ohne eines Modells (diskriminativer Ansatz) erfolgen.

Zur Umsetzung wurde das System in zwei Teilsysteme aufgeteilt. Im Falle des Teilsystems „Mensch“ wird als erstes der diskriminative Ansatz mittels neuronaler Netzwerken zur Erkennung der Pose in 2D verwendet. Als Input gelten hier die Bilder, die aus den Amplituden generiert werden können. Für die Robustheit der anschließenden dreidimensionalen Transformation der Pose gegenüber den Abbildungsfehlern sorgt die Verwendung eines menschlichen Modells. Somit lässt sich dies als ein hybrider Ansatz klassifizieren, da die Erkennung durch den diskriminativen Ansatz im zweiten Schritt durch ein Modells verfeinert wird.

Abgeleitet aus der geringen Dichte der Punktwolke wurde zur Erkennung der 6D Pose des Bauteils der Ansatz mittels globaler Deskriptoren gewählt. Als Begründung hierfür gilt neben dem Fehlen lokaler Merkmale ebenfalls die Robustheit gegenüber lokaler Änderungen, die etwa durch Clutter durch den Menschen entstehen können. In Hinsicht auf die Funktionalität zeigt das System vielversprechende Resultate, die im nächsten Abschnitt diskutiert werden.

## 6.1 Diskussion der Ergebnisse und des Potentials des Systems

Aufbauend auf der Beschreibung der Implementierung sowie der Evaluierung, werden zunächst die Kriterien zusammengefasst. Anschließend werden die Resultate diskutiert.

Die Kriterien zur Anwendung eines Systems richten sich verallgemeinert an funktionale, sowie nicht-funktionale Kriterien. Abgeleitet aus der Architektur eines AR-Assistenzsystems zählt zu den funktionalen Kriterien die Fähigkeit des Systems, Tracking des Menschen und des Bauteils, sowie Mensch-Maschine-Interaktion, zu gewährleisten. Des Weiteren bestehen auf das System nicht-funktionale Kriterien, die sich aus ISO/IEC 25000 [88] ableiten lassen. Dazu gehören insbesondere die Anforderungen bzgl. Zeitverhalten, Ressourcen-Auslastung und Kapazität, sowie die Änderbarkeit des Systems. In folgenden Abschnitten wird zuerst auf die funktionalen Anforderungen eingegangen, anschließend wird das System in Hinsicht auf die nicht-funktionellen Kriterien evaluiert.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Um Tracking des Menschen gewährleisten zu können, besteht der Output des wahrnehmenden Systems einerseits aus den dreidimensionalen Koordinaten der Keypoints des menschlichen Körpers, mittels denen die Pose des Menschen, die räumliche Lage und die Gesten beschrieben werden können. Des Weiteren wird die Pose des Bauteils als eine starren Transformation wiedergegeben, wodurch das Tracking des Bauteils ermöglicht wird. In Hinsicht auf die Funktionalität weist das Systems stabiles Verhalten auf, solange sich alle relevanten Informationen im Sichtfeld der Kameras, d.h. im Arbeitsbereich, befinden. Ausnahmen stellen insbesondere spezielle Posen des Menschen, in denen zu starken Selbst-Okklusionen kommt. Somit gelten die Anforderungen an die Funktionalität des Systems als erfüllt.

Unter der Annahme, dass das System über die gewünschte Funktionalität verfügt, kann ein potenzieller industrieller Einsatz eines Systems weiters anhand der nichtfunktionellen Kriterien evaluiert werden. Dies bestimmt einerseits die Realisierbarkeit der Implementierung, die sich aus den Anforderungen an die Hardware ergibt, sowie das Zeitverhalten, das die Usability eines AR-Assistenzsystems stark beeinflusst.

Der Einsatz des 3D-Trackings mittels Time of Flight Kameras ist, abgesehen von der Notwendigkeit des Kamerasystems, mittels gängiger Rechner realisierbar. Dies ergibt sich durch den Bedarf an Arbeitsspeicher, der sich in Grenzen hält. In dieser Hinsicht stellt das wahrnehmende System keinerlei Hindernis dar, denn es bedarf lediglich eine geringe Speicherkapazität zur Speicherung der Modelle zur Erkennung. Diese bestehen einerseits aus dem CAD Modells des Bauteil, sowie den DL-Modellen zur Erkennung des Menschen. Hinsichtlich der Variantenvielfalt in der Baustellenfertigung wird die Änderbarkeit zu einem wichtigen Kriterium. Das System bietet die Möglichkeit an, ein neues CAD Modell schnell und effizient einzuspielen, wodurch dies ebenfalls gewährleistet wird. Des Weiteren wäre es möglich, das System auf mehrere Arbeitsstationen zu erweitern. Eine Voraussetzung hierfür ist die Fähigkeit zum Tracking mehrerer Personen. Dies ist grundsätzlich gegeben und lässt sich durch eine einfache Anpassung des Systems implementieren.

Die Interaktionsfähigkeit in Echtzeit stellt eine der Kernkomponenten der AR-Systeme dar. Im Rahmen der Evaluierung auf die Anforderung bezüglich der Weitergabe der Informationen zur richtigen Zeit hat sich das System als mangelhaft aufgewiesen. Als eine minimale Rate der Erkennung der Pose des Menschen wird hierbei mindestens zwei ausgewertete Aufnahmen pro Sekunde, d.h. eine Erkennungsdauer von 0.5 Sekunden, vorgeschlagen. Für das Teilsystem Bauteil wird hier die minimale Erkennungsdauer gleich 1 Sekunde vorgeschlagen. Dies wird durch die großen Abmessungen des Bauteils begründet, wodurch schnelle Verschiebungen des Bauteils nicht möglich sind. Die derzeitige Implementierung erfüllt diese Echtzeitanforderungen nicht. Für das Teilsystem „Bauteil“ liegt die Erkennenungsdauer oberhalb zwei Sekunden, für das Teilsystem „Mensch“ liegt dies im Bereich von 9 Sekunden pro Aufnahme. Dementsprechend weist das System Optimierungsbedarf auf. Der zeitliche Bedarf ist hierbei allerdings abhängig von der verwendeten Hardware. Der Autor dieser Arbeit erhofft, dass dies durch den Durchbruch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

im Bereich des parallelen Rechnens mittels GPU, der in den letzten Jahren stattgefunden hat, lösbar ist. Heutzutage ist die Integration leistungsstärker GPUs, die durch die sinkenden Preise und hohe Rechenfähigkeit getrieben ist, in die gängige Hardware omnipräsent. Zunutze wird dies gerade im Bereich von Computer Vision gemacht (vgl. [123]). Ein erneutes Testing und eine Evaluierung der Zeiten mittels geeigneter Hardware werden empfohlen.

Zusammenfassend lässt sich feststellen, dass Time of Flight Kameras zum 3D-Tracking geeignet sind. Dies ergibt sich aus der Funktionalität des Systems. Um das Potential von 3D-Tracking zur Ansteuerung eines Spatial Augmented Reality Assistenzsystems ableiten zu können, bedarf es die Einbeziehung mehrerer Kriterien. Hierzu zählen insbesondere das Zeitverhalten, die Änderbarkeit und die Ressourcen-Auslastung. In dieser Perspektive ergibt sich Optimierungsbedarf hinsichtlich des Zeitverhaltens des Systems. Geringe Erkennungsdauer ist anzustreben, denn diese dürfte die Usability des Systems stark beschränken. Die Lösung hierfür stellt allerdings der Einsatz von GPU zur Beschleunigung der Rechenleistung dar, wodurch die Tauglichkeit der Time of Flight Kameras zur Ansteuerung eines in der Industrie anwendbaren Spatial Augmented Reality Assistenzsystems gewährleistet werden kann.

## 6.2 Weitere Entwicklung und Optimierung des wahrnehmenden Systems

Die entscheidenden Faktoren zur Projektion der richtigen Informationen und zur richtigen Zeit sind die Geschwindigkeit und eine stabile Güte der Erkennung. Aufbauend auf der Beschreibung der Implementierung sowie der zuvor angemerkten Vorschlägen zur Implementierung auf einer GPU, werden nun Vorschläge zur weiteren Optimierung gelistet.

Mit einem Blick auf die umgesetzte globale Pipeline zur Erkennung der Pose des Menschen, lässt sich diese durch die Vereinigung der Segmentierung und Erkennung der Pose in eine Methode vereinfachen. Die Implementierung dieser kann, beispielsweise, durch Mask R-CNN erfolgen, das durch eine Erweiterung des Datensatzes um die Keypoints und ein erneutes Eintrainieren ebenfalls die Pose ermitteln kann [96]. Dies führt allerdings zu einer Verschlechterung der Güte der Ergebnisse, weshalb auf die Arbeit von Zhang et al. verwiesen wird [124]. Ihre Methode zielt auf die detektionsfreie Segmentierung des Menschen durch den Einsatz Skeleton-basierter Features. Richtwerte für die zeitliche Performance dieser Methode, die demnach gleichzeitig die Pose ermittelt sowie die Segmentierung durchführt, betragen ungefähr 0.05 Sekunden pro Aufnahme (GPU-basiert) [124].

Des Weiteren weist die Projektion der Pose in die 3D-Punktwolke Optimierungspotential auf. Hierbei soll sowohl nach der Verbesserung der Güte der Erkennung, sowie nach einer Beschleunigung gestrebt werden. Eine Möglichkeit zur Verbesserung der Güte bietet die Einbeziehung eines Netzwerkes zur Ermittlung der Koordinaten in 3D an.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Als eine Vorlage dient hierbei der Ansatz von [78]. Dieser Ansatz muss allerdings auf die Tauglichkeit genauer untersucht werden, da aufgrund der Okklusionen, und der geringen Dichte der Punktwolke, die Erkennung in 3D erschwert wird.

In Rahmen dieser Arbeit wurde ein einfaches Konzept zur Gestaltung der Mensch-Maschine Interaktion basierend auf 3D-Skeletons vorgestellt. Dies weist allerdings große Einschränkungen hinsichtlich der Klassifikation mehrerer Gesten auf. Hierzu wird der Einsatz von Machine Learning zur Klassifikation der Gesten vorgeschlagen. Eine einfache Möglichkeit zur Erweiterung besteht in Anwendung einer flachen Klassifizierung, oder beispielsweise mittels SVM (Englisch: Support Vector Machine), zur einfachen Klassifikation der Gesten. Zur besseren Abbildung der zeitlichen Komplexität der menschlichen Tätigkeiten werden Temporal-Evolution Netzwerke vorgeschlagen. Diese ermöglichen eine Klassifikation basierend auf dem lokalen zeitlichen Verlauf der Tätigkeit, wodurch die Ausdrucksfähigkeit im Vergleich zur statischen Erkennung und unter Einbeziehung der Zeit, wie etwa in dieser Arbeit vorgeschlagen, deutlich verbessert wird (in Anlehnung an [110]).

Abschließend wird, insbesondere in Hinsicht auf die Erkennungszeit, der Bedarf nach einer Analyse Deep Learning-basierter Ansätze zur Erkennung der 6D Pose des Bauteils geäußert. Die Motivation hierfür ist der Abbau der Pipeline zum Preprocessing, wodurch der Zeitbedarf verringert werden könnte. Dies setzt allerdings eine sehr große Robustheit der Methode gegenüber Rauschen und einer mangelhafter Qualität der Punktwolke voraus. Dementsprechend können Algorithmen, wie etwa PointNet [65], auf die Performance und die Güte der Erkennung ohne bzw. mit minimalen Preprocessing untersucht werden.

## 6.3 Entwicklung einer Schnittstelle zur Projektion

An der vorgeschlagenen Architektur des Spatial Augmented Reality Assistenzsystems aufbauend, widmet sich dieser Abschnitt der groben Definition der Schnittstelle zwischen der Wahrnehmung der Informationen und der Projektion. Die zur Projektion notwendigen Informationen richten sich an die vorgestellten Ergebnisse dieser Arbeit, nämlich die Pose des Bauteils, sowie des Mitarbeiters im dreidimensionalen Raum und die Gestaltung der Kommunikation zwischen dem Menschen und dem System.

Zur Extraktion der richtigen Informationen zu der richtigen Zeit und somit zur Optimierung des Zeitverhaltens des Systems wird eine klare Klassifikation der zu projizierenden Kontexte vorgeschlagen. Begründet wird dies durch die verschiedenen Stufen der Interaktion zwischen dem Mitarbeiter und dem Bauteil. Hier gilt beispielsweise, dass für bestimmte Tätigkeiten, wie das Holen von Material, eine minimale Interaktion besteht. Andererseits besteht beim Bewegen des Bauteils oder bei der Aufbringung von Material auf das Bauteil eine sehr hohe Interaktion. Dementsprechend lässt sich anhand des Kontextes der projizierten Arbeitsanweisungen die Interaktion abschätzen und der Umfang der abzufragenden Daten steuern. Dies impliziert, dass bei einer Arbeitsanweisung, die



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

keine Interaktion voraussetzt, die Erfassung der Pose des Bauteils überflüssig wäre und dementsprechend zu keinem Informationsgewinn führen würde, und lediglich die Performance beeinträchtigen würde. Der Kontext der projizierten Arbeitsanweisungen bildet ebenfalls die Vorgaben zur Extraktion weiterer Informationen über den Menschen. Somit reicht beispielsweise für bestimmte Tätigkeiten die Erfassung der Lage des geometrischen Mittelpunkts des menschlichen Körpers. Umgekehrt erfordern bestimmte Arbeitstätigkeiten eine durchgehende Projizierung neuer Arbeitsanweisungen, wodurch die Pose und insbesondere die räumliche Ausrichtung des Menschen relevant ist.

Abschließend wird auf die koordinatentechnische Abhängigkeit der zwei Systeme kurz eingegangen. Die Erfassung der 3D-Pose des Mitarbeiters und des Bauteils führen schließlich zu keinem Vorteil, es sei denn, die Koordinatensysteme des wahrnehmenden Systems und des Systems zur Projektion sind zueinander mathematisch referenziert. Basierend auf einer Referenz der zwei Koordinatensysteme kann die Geometrie des Bauteils im Projektionssystem nachgebildet werden. Hierzu werden weiters die aktuelle Position des Bauteils und ein CAD-Modell bzw. ein Mesh benötigt. Darauf aufbauend kann die perspektivische Varianz der Projektion, die sich durch die konkave Geometrie des Bauteils ergibt, durch geeignete projektive Transformation behoben werden.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Abbildungsverzeichnis


1.1 Veranschaulichung der Struktur der Arbeit . . . . . 6

2.1 Umfrage zur Bedeutung verschiedener Kernkompetenzen und das Potential der digitalen Unterstützung [17] . . . . . 9
2.2 Veranschaulichung der Komponenten der Mensch-Maschine Interaktion (in Anlehung an [18]) . . . . . 10
2.3 Taxonomie der informationstechnischen Schnittstellen (in Anlehnung an [14]) . . . . . 12
2.4 Bestandteile der Architektur von Augmented Reality Assistenzsystemen (in Anlehung an [24]) . . . . . 13
2.5 Darstellung des Pinhole-Modells (vgl. [29][31]) . . . . . 14
2.6 Veranschaulichung des Prinzips der passiven Triangulation (in Anlehnung an [31]) . . . . . 16
2.7 Veranschaulichung des Prinzips der aktiven Triangulation (in Anlehnung an [33]) . . . . . 17
2.8 Veranschaulichung des Prinzips von Time of Flight Kameras (in Anlehung an [31]) . . . . . 18
2.9 Darstellung der Amplituden(links), Intensitäten(Mitte) und der Depth-Map(rechts) [31] . . . . . 18
2.10 Emittiertes und reflektiertes Signal bei gepulster Modulation(links) und bei CW-Modulation(rechts) [31] . . . . . 19
2.11 Koordinaten eines Punktes (in Anlehnung an [31]) . . . . . 21
2.12 Veranschaulichung der statistischen Präzision und der Genauigkeit (in Anlehnung an [31]) . . . . . 22
2.13 Möglichkeiten zur überwachten Kalibrierung von Kameras: 3D (links), 2D (Mitte) und 1D (rechts) [31] . . . . . 23
2.14 Darstellung der Oberflächennormalen anhand einer planaren Box und eines Zylinders [41] . . . . . 26
2.15 Veranschaulichung der unterschiedlichen Ansätze: a) traditioneller Ansatz b) Deep Leaning-Ansatz [40] . . . . . 29
2.16 Veranschaulichung des Prinzips der Convolution [53] . . . . . 30



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.1 a) Veranschaulichung des Prinzips von Template-Matching als Minimierung des Abstandes zwei Objekte (anhand ICP [55]), b) Veranschaulichung des Prinzips von Sparse-Features-Matching anhand 2D Daten [56] 32
3.2 Unterschiede in den Pipelines bei Verwendung lokaler Deskriptoren (oberes Abbild) und globaler Deskriptoren (unteres Abbild) [62] . . . . . 34
3.3 Architektur von PointNet [65] . . . . . 36
3.4 Veranschaulichung der Architektur von PointNet++ [67] . . . . . 37
3.5 Veranschaulichung der segmentierter Umgebung rund um die kritischen Punkte [68] . . . . . 37
3.6 Veranschaulichung der Keypoints-Ermittlung (als Keypoints dienen in diesem Fall die Ecken der Bounding Box): a.) die generierte Bounding Box b.) Vektoren zwischen den Keypoints (Ecken der Bounding Box) und zwei zufälligen Punkten p1 und p2, c.) Intersektion der Vektoren als Basis für die neuen Keypoints [69] . . . . . 38

4.1 Veranschaulichung eines Modells der menschlichen kinematischen Kette [72] . . . . . 39
4.2 Modelle für generative Ansätze: a) Skeleton, b) Kontur-Modell, c) volumetrisches Modell [75] . . . . . 41
4.3 Ermittlung der Pose aus der Bewegung für einen Löwen: Darstellung der Features und der Keypoints [80] . . . . . 43
4.4 Architektur von DeepPose [81] . . . . . 43
4.5 Veranschaulichung der Architektur zur Domänenadaptation von [78] . . . 44
4.6 Bedarf der Invarianz gegenüber der perspektivischen Verzerrung bei der Rotation der Hand – Vergleich der 1 zu 1 Zuordnung von 3D-Pose zu den 3D-Keypoints (links) und N-Korrespondenzen derselben Pose in 2D zu den 3D-Keypoints durch die perspektivische Transformation [83] . . . . . 45
4.7 Architektur von Hand PointNet [85] . . . . . 46
4.8 Architektur verwendet von Zhou et al. – räumliche Transformation durch T-Net (Bestandteil von PointNet), zwei Units bestehend aus EdgeConvs (DGCNN) gefolgt durch Fully-Connected Layer [86] . . . . . 46

5.1 Veranschaulichung des Konzeptes des Spatial Augmented Reality Assistenzsystems [8] . . . . . 47
5.2 Veranschaulichung des Set-Ups in der Pilotfabrik der TU Wien bestehend aus drei Time of Flight Kameras und dem dazwischenliegenden Arbeitsbereich . . . . . 48
5.3 Darstellung der Komponenten eines markerlosen Vision-basierten Trackingsystems im Rahmen der AR-Architektur . . . . . 50
5.4 Synchronisierungsprobleme der Punktwolke des Menschen bei schnellen Bewegungen: links eine statische Pose und rechts Bewegung (Spazieren) . 52
5.5 Darstellung der Herausforderung bei der Erkennung des Bauteils: Punktwolke des Bauteils (links) und ein Foto des realen Objektes (rechts) . . . 53



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.6 Vorgeschlagene Architektur für das Spatial Augmented Reality Assistenzsystem in der Pilotfabrik der TU Wien . . . . . 57
5.7 Globale Pipeline des Teilsystems „Mensch“ . . . . . 60
5.8 Vergleich verschiedener Thresholding-Werte: links – 1800, mittig – 550, rechts – 200 . . . . . 61
5.9 Multi-stage Architektur von OpenPose [97] . . . . . 62
5.10 Ergebnisse von Mask R-CNN: links die ermittelte Bounding Box, rechts die Maske . . . . . 63
5.11 COCO-Modell: Übersicht der Keypoints (links), abgeleitetes räumliches Modell der Maßlängen des menschlichen Körpers (rechts) . . . . . 64
5.12 Vergleich der zweidimensionalen Pose: links ohne das Segmentierungsnetzwerk, rechts mit dem Segmentierungsnetzwerk . . . . . 64
5.13 Veranschaulichung der Pose anhand der segmentierten Punktwolke: 3D-Projektion ohne Anpassung und Verifizierung (links) Projektion der 2D-Pose mit Anpassung und Verifizierung . . . . . 66
5.14 Veranschaulichung der Geste und der zugehörigen Winkel . . . . . 68
5.15 Nicht-okkludierte, vollständige Darstellung des Skelets und Erkennung der Geste . . . . . 69
5.16 Abgelehnte Geste („Arbeiten am Werkstück“). Okklusion durch das Werkstück führt zur Nicht-Erkennung der Beine . . . . . 69
5.17 Übersicht der globalen Datenpipeline des Teilsystem „Bauteil“ . . . . . 70
5.18 Datenpipeline von Preprocessing . . . . . 70
5.19 Darstellung der Verschmelzung der rohen Punktwolken der drei Kameras (links) und die abgegrenzte ROI (rechts) . . . . . 71
5.20 Veranschaulichung von Denoising (Darstellung bei dreifacher Punktgröße): Ausreißer (links) und gefilterte Punktwolke (rechts) . . . . . 72
5.21 Darstellung von MLS anhand des segmentierten Bauteils (Anzahl der Punkte auf 20% reduziert): rohe Oberfläche (schwarz) samt Oberflächennormalen dargestellt in rot (links) und rekonstruierte Oberfläche mit gleichmäßigen Normalen (rechts) . . . . . 73
5.22 Veranschaulichung der Extraktion der Punktwolke des Bodens: Punktwolke mit Boden (links) und Ergebnis der Szene nach Preprocessing (rechts) 74
5.23 Datenpipeline zur Erkennung der Pose des Bauteils . . . . . 74
5.24 Veranschaulichung der Normalenberechnung: nicht-konsistente Ausrichtung der Oberflächennormalen (links) und übliche Behebung durch die Kenntnis des Aussichtspunkts (rechts) . . . . . 75
5.25 Outline der Methode von [118] zur Erkennung der Pose mittels PPF und anschließender Verfeinerung . . . . . 76
5.26 Veranschaulichung der Suche nach korrespondierenden Punktpaaren [95] 77
5.27 Erkannte Pose des Bauteils . . . . . 79
5.28 Schwer erkennbare Posen: starke Okklusion durch andere Objekte betreffend insbesondere Kamera 1 und Kamera 2 (links) und Selbstokklusion beim Heben einer Kiste (rechts) . . . . . 80



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.1 Verwendete Methoden und Parameter beim Preprocessing von Bauteil . . 108
6.2 Verwendete Methoden und Parameter beim Preprocessing von Bauteil . . 108
6.3 Verwendete Methoden und Parameter beim Teilsystem Mensch . . . . . 111

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Tabellenverzeichnis


2.1 Auflistung der Rezeptoren und Responder des menschlichen Körpers und Beispiele für Anwendungen im Rahmen der MMI (in Anlehung an [18]) . 11
2.2 Mathematische Operation zur Manipulation von dreidimensionalen Koordinaten [38] . . . . . 24

4.1 Unterteilung der Motion Capture Systeme (in Anlehung an [6][73]) . . . . 40

5.1 Aufteilung der allgemeinen Herausforderungen von markerlosem Tracking mit einem ToF-Multi-Kamerasystem . . . . . 51
5.2 Übersicht der Bibliotheken . . . . . 58
5.3 Übersicht der Zeiten für das Teilsystem **Mensch** . . . . . 81
5.4 Übersicht der Zeiten für das Teilsystem **Bauteil** . . . . . 81



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Literaturverzeichnis


[1] T. Bauernhansl, M. Ten Hompel, and B. Vogel-Heuser, *Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung-Technologien-Migration*. Springer, 2014.

[2] A. Botthof and E. Andreas Hartmann, *Zukunft der Arbeit in Industrie 4.0*. Springer Nature, 2015.

[3] D. Romero, J. Stahre, T. Wuest, O. Noran, P. Bernus, Å. Fast-Berglund, and D. Gorecky, "Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies," in *proceedings of the international conference on computers and industrial engineering (CIE46), Tianjin, China*, 2016, pp. 29–31.

[4] D. Wee, R. Kelly, J. Cattel, and M. Breunig, "Industry 4.0-how to navigate digitization of the manufacturing sector," *McKinsey & Company*, vol. 58, 2015.

[5] F. Longo, L. Nicoletti, and A. Padovano, "Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context," *Computers & industrial engineering*, vol. 113, pp. 144–159, 2017.

[6] M. Bortolini, M. Gamberi, F. Pilati, and A. Regattieri, "Automatic assessment of the ergonomic risk for manual manufacturing and assembly activities through optical motion capture technology," *Procedia CIRP*, vol. 72, pp. 81–86, 2018.

[7] J. Deuse and F. Busch, "Zeitwirtschaft in der montage," in *Montage in der industriellen Produktion*. Springer, 2012, pp. 79–107.

[8] P. Rupprecht, H. Kueffner-Mccauley, and S. Schlund, "Information provision utilizing a dynamic projection system in industrial site assembly," *Procedia CIRP*, vol. 93, pp. 1182–1187, 2020.

[9] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (tof) cameras: A survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, 2011.

[10] Z. Zhang, Y. Dai, and J. Sun, "Deep learning based point cloud registration: an overview," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 222–246, 2020.

[11] U. Sendler, "Industrie 4.0–beherrschung der industriellen komplexität mit syslm (systems lifecycle management)," in *Industrie 4.0*. Springer, 2013, pp. 1–19.




Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[12] T. Keller, C. Bayer, P. Bausch, and J. Metternich, "Benefit evaluation of digital assistance systems for assembly workstations," *Procedia CIRP*, vol. 81, pp. 441–446, 2019.

[13] R. Müller, L. Hörauf, and A. Bashir, "Cognitive assistance systems for dynamic environments," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2019, pp. 649–656.

[14] W. Gerke, *Technische Assistenzsysteme: vom Industrieroboter zum Roboterassistenten*. Walter de Gruyter GmbH & Co KG, 2014.

[15] B. Ludwig, *Planbasierte Mensch-Maschine-Interaktion in multimodalen Assistenzsystemen*. Springer, 2015.

[16] S. Röttger, K. Bali, and D. Manzey, "Do cognitive assistance systems reduce operator workload?" *Human factors in complex system performance*, pp. 351–360, 2007.

[17] W. Apt, M. Schubert, and S. Wischmann, "Digitale assistenzsysteme," *Perspektiven und Herausforderungen für den Einsatz in Industrie und Dienstleistungen. Hg. v. Institut für Innovation und Technik in der VDI/VDE Innovation+ Technik GmbH (iit). Berlin. Online verfügbar unter https://www. iit-berlin. de/de/publikationen/digitale-assistenzsysteme, zuletzt geprüft am*, vol. 14, p. 2018, 2018.

[18] I. S. MacKenzie, *Human-computer interaction: An empirical research perspective*. Newnes, 2012.

[19] B. Preim and R. Dachselt, *Interaktive Systeme: Band 2: User Interface Engineering, 3D-Interaktion, Natural User Interfaces*. Springer-Verlag, 2015.

[20] D. Gorecky, M. Schmitt, M. Loskyll, and D. Zühlke, "Human-machine-interaction in the industry 4.0 era," in *2014 12th IEEE international conference on industrial informatics (INDIN)*. IEEE, 2014, pp. 289–294.

[21] Y. G. Kim, K. Little, B. Noronha, M. Xiloyannis, L. Masia, and D. Accoto, "A voice activated bi-articular exosuit for upper limb assistance during lifting tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 66, p. 101995, 2020.

[22] M. Quandt, T. Beinke, and M. Freitag, "User-centered evaluation of an augmented reality-based assistance system for maintenance," *Procedia CIRP*, vol. 93, pp. 921–926, 2020.

[23] F. Loch, F. Quint, and I. Brishtel, "Comparing video and augmented reality assistance in manual assembly," in *2016 12th International Conference on Intelligent Environments (IE)*. IEEE, 2016, pp. 147–150.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[24] X. Wang, S. K. Ong, and A. Y. Nee, "A comprehensive survey of augmented reality assembly research," *Advances in Manufacturing*, vol. 4, no. 1, pp. 1–22, 2016.

[25] S. R. Fussell, L. D. Setlock, and R. E. Kraut, "Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 513–520.

[26] W. Vorraber, J. Gasser, H. Webb, D. Neubacher, and P. Url, "Assessing augmented reality in production: remote-assisted maintenance with hololens," *Procedia CIRP*, vol. 88, pp. 139–144, 2020.

[27] B. G. Mark, E. Rauch, and D. T. Matt, "Study of the impact of projection-based assistance systems for improving the learning curve in assembly processes," *Procedia CIRP*, vol. 88, pp. 98–103, 2020.

[28] L. Rodriguez, F. Quint, D. Gorecky, D. Romero, and H. R. Siller, "Developing a mixed reality assistance system based on projection mapping technology for manual operations at assembly workstations," *Procedia computer science*, vol. 75, pp. 327–333, 2015.

[29] C. Wöhler, *3D computer vision: efficient methods and applications*. Springer Science & Business Media, 2012.

[30] B. Cyganek and J. P. Siebert, *An introduction to 3D computer vision techniques and algorithms*. John Wiley & Sons, 2011.

[31] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, "Time-of-flight and structured light depth cameras," *Technology and Applications*, pp. 978–3, 2016.

[32] Y. Xie, J. Tian, and X. X. Zhu, "A review of point cloud semantic segmentation," *arXiv preprint arXiv:1908.08854*, 2019.

[33] I. Kalová and M. Lisztwan, "Industrial applications of triangulation technique," *IFAC Proceedings Volumes*, vol. 39, no. 21, pp. 258–263, 2006.

[34] S. Song, B. Wang, W. Gong, Z. Chen, X. Lin, J. Sun, and S. Shi, "A new waveform decomposition method for multispectral lidar," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 40–49, 2019.

[35] M. Marghany, *Synthetic Aperture Radar Imaging Mechanism for Oil Spills*. Gulf Professional Publishing, 2019.

[36] L. Shao, J. Han, P. Kohli, and Z. Zhang, *Computer vision and machine learning with RGB-D sensors*. Springer, 2014, vol. 20.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[37] M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, *Time-of-Flight and Depth Imaging. Sensors, Algorithms and Applications: Dagstuhl Seminar 2012 and GCPR Workshop on Imaging New Modalities*. Springer, 2013, vol. 8200.

[38] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[39] M. Weinmann, B. Jutzi, and C. Mallet, "Geometric features and their relevance for 3d point cloud classification," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, p. 157, 2017.

[40] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Science and Information Conference*. Springer, 2019, pp. 128–144.

[41] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, "Comparison of surface normal estimation methods for range sensing applications," in *2009 IEEE international conference on robotics and automation*. IEEE, 2009, pp. 3206–3211.

[42] K. Jordan and P. Mordohai, "A quantitative evaluation of surface normal estimation in point clouds," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4220–4226.

[43] X.-F. Hana, J. S. Jin, J. Xie, M.-J. Wang, and W. Jiang, "A comprehensive review of 3d point cloud descriptors," *arXiv preprint arXiv:1802.02297*, 2018.

[44] M. Bueno, J. Martínez-Sánchez, H. González-Jorge, and H. Lorenzo, "Detection of geometric keypoints and its application to point cloud coarse registration." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[45] R. Hänsch, T. Weber, and O. Hellwich, "Comparison of 3d interest point detectors and descriptors for point cloud fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, no. 3, p. 57, 2014.

[46] N. Zrira, F. Z. Ouadiay, M. Hannat, E. H. Bouyakhf, and M. M. Himmi, "Evaluation of pcl's descriptors for 3d object recognition in cluttered scene," in *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, 2017, pp. 1–6.

[47] J. H'roura, M. Roy, A. Mansouri, D. Mammass, P. Juillion, A. Bouzit, and P. Méniel, "Salient spin images: A descriptor for 3d object recognition," in *International conference on image and signal processing*. Springer, 2018, pp. 233–242.

[48] M. Madry, C. H. Ek, R. Detry, K. Hang, and D. Kragic, "Improving generalization for 3d object categorization with global structure histograms," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 1379–1386.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[49] T. O. Ayodele, "Machine learning overview," *New Advances in Machine Learning*, pp. 9–19, 2010.

[50] A. V. Joshi, *Machine Learning and Artificial Intelligence*. Springer, 2020.

[51] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[52] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[53] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[54] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *European conference on computer vision*. Springer, 2014, pp. 536–551.

[55] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on gpus–a comprehensive review," *Medical image analysis*, vol. 20, no. 1, pp. 1–18, 2015.

[56] Y. Wang, S. Zhang, B. Wan, W. He, and X. Bai, "Point cloud and visual feature-based tracking method for an augmented reality-aided mechanical assembly system," *The International Journal of Advanced Manufacturing Technology*, vol. 99, no. 9-12, pp. 2341–2352, 2018.

[57] C. Sahin, G. Garcia-Hernando, J. Sock, and T.-K. Kim, "A review on object pose recovery: from 3d bounding box detectors to full 6d pose estimators," *Image and Vision Computing*, p. 103898, 2020.

[58] S. A. Bello, S. Yu, C. Wang, J. M. Adam, and J. Li, "deep learning on 3d point clouds," *Remote Sensing*, vol. 12, no. 11, p. 1729, 2020.

[59] Z. Zhang, Y. Dai, and J. Sun, "Deep learning based point cloud registration: an overview," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 222–246, 2020.

[60] R. Vock, A. Dieckmann, S. Ochmann, and R. Klein, "Fast template matching and pose estimation in 3d point clouds," *Computers & Graphics*, vol. 79, pp. 36–45, 2019.

[61] R. Radkowski, "Object tracking with a range camera for augmented reality assembly assistance," *Journal of Computing and Information Science in Engineering*, vol. 16, no. 1, 2016.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[62] K. Alhamzi, M. Elmogy, and S. Barakat, "3d object recognition based on local and global features using point cloud library," *International Journal of Advancements in Computing Technology*, vol. 7, no. 3, p. 43, 2015.

[63] D. Li, N. Liu, Y. Guo, X. Wang, and J. Xu, "3d object recognition and pose estimation for random bin-picking using partition viewpoint feature histograms," *Pattern Recognition Letters*, vol. 128, pp. 148–154, 2019.

[64] Z. Deng and L. Jan Latecki, "Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5762–5770.

[65] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[66] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *arXiv preprint arXiv:1511.06391*, 2015.

[67] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.

[68] F. Hagelskjær and A. G. Buch, "Pointvotenet: Accurate object detection and 6 dof pose estimation in point clouds," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2641–2645.

[69] W. Chen, J. Duan, H. Basevi, H. J. Chang, and A. Leonardis, "Ponitposenet: Point pose network for robust 6d object pose estimation," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2824–2833.

[70] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports medicine-open*, vol. 4, no. 1, p. 24, 2018.

[71] A. C. Bovik, *Handbook of image and video processing*. Academic press, 2010.

[72] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[73] H. Mousavi Hondori and M. Khademi, "A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation," *Journal of medical engineering*, vol. 2014, 2014.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[74] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[75] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[76] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1669–1676.

[77] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2d human pose estimation: A survey," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 663–676, 2019.

[78] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez, "Efficient convolutional neural networks for depth-based multi-person pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[79] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 459–468.

[80] B. Daubney, D. Gibson, and N. Campbell, "Estimating pose of articulated objects using low-level motion," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 330–346, 2012.

[81] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[82] K.-C. Chan, C.-K. Koh, and C. G. Lee, "A 3d-point-cloud feature for human-pose estimation," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1623–1628.

[83] G. Moon, J. Yong Chang, and K. Mu Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 5079–5088.

[84] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1986–1992.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[85] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8417–8426.

[86] Y. Zhou, H. Dong, and A. El Saddik, "Learning to estimate 3d human pose from point cloud," *IEEE Sensors Journal*, vol. 20, no. 20, pp. 12 334–12 342, 2020.

[87] "Webseite der Becom-Group," Eingesehen am 16.06.2020. [Online]. Available: https://www.becom-group.com/becom-systems-time-of-flight-solutions/tof-3d-kameras/argos-kameras/argos3d-p330/

[88] "ISO/IEC 25000:2014(E) – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE," International Organization for Standardization, Geneva, CH, Standard, 15.03. 2014.

[89] "ISO/IEC 25010:2011(E) – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models," International Organization for Standardization, Geneva, CH, Standard, 01.03. 2011.

[90] M.-H. Song and R. I. Godøy, "How fast is your body motion? determining a sufficient frame rate for an optical motion tracking system using passive markers," *PloS one*, vol. 11, no. 3, p. e0150993, 2016.

[91] L. Li, S. Xiang, Y. Yang, and L. Yu, "Multi-camera interference cancellation of time-of-flight (tof) cameras," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 556–560.

[92] V. Castaneda, D. Mateus, and N. Navab, "Stereo time-of-flight," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1684–1691.

[93] D. Peddireddy, X. Fu, H. Wang, B. G. Joung, V. Aggarwal, J. W. Sutherland, and M. B.-G. Jun, "Deep learning based approach for identifying conventional machining processes from cad data," *Procedia Manufacturing*, vol. 48, pp. 915–925, 2020.

[94] F. Liu and Z. Wang, "Polishnet-2d and polishnet-3d: deep learning-based workpiece recognition," *IEEE Access*, vol. 7, pp. 127 042–127 054, 2019.

[95] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*. Ieee, 2010, pp. 998–1005.

[96] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[97] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[98] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[99] P. Abeles, "Boofcv v0.25," http://boofcv.org/, 2016.

[100] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[101] "MATLAB Computer Vision Toolbox," Eingesehen am 28.07.2020. [Online]. Available: https://www.mathworks.com/products/computer-vision.html

[102] K. Zampogiannis, C. Fermuller, and Y. Aloimonos, "Cilantro: A lean, versatile, and efficient library for point cloud data processing," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1364–1367.

[103] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[104] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1–4.

[105] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[106] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[107] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[108] "DIN 33402-2: Ergonomie – Körpermaße des Menschen – Teil 2: Werte," Deutsches Institut für Normung e.V., Berlin, DE, Standard, März 2005.

[109] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, 2017.

[110] Q. Ke, J. Liu, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Computer vision for human–machine interaction," in *Computer Vision for Assistive Healthcare*. Elsevier, 2018, pp. 127–145.

[111] S. Chen and R. R. Yang, "Pose trainer: correcting exercise posture using pose estimation," *arXiv preprint arXiv:2006.11718*, 2020.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[112] X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, "A review of algorithms for filtering the 3d point cloud," *Signal Processing: Image Communication*, vol. 57, pp. 103–112, 2017.

[113] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.

[114] M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva, "State of the art in surface reconstruction from point clouds," 2014.

[115] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C. T. Silva, "Computing and rendering point set surfaces," *IEEE Transactions on visualization and computer graphics*, vol. 9, no. 1, pp. 3–15, 2003.

[116] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, no. 2. Wiley Online Library, 2007, pp. 214–226.

[117] L. Li, F. Yang, H. Zhu, D. Li, Y. Li, and L. Tang, "An improved ransac for 3d point cloud plane segmentation based on normal distribution transformation cells," *Remote Sensing*, vol. 9, no. 5, p. 433, 2017.

[118] "Dokumentation von OpenCV - Modul Surface Matching," Eingesehen am 01.06.2020. [Online]. Available: https://docs.opencv.org/master/d9/d25/group__surface___matching.html

[119] T. Zinßer, J. Schmidt, and H. Niemann, "A refined icp algorithm for robust 3-d correspondence estimation," in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, vol. 2. IEEE, 2003, pp. II–695.

[120] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy, "Geometrically stable sampling for the icp algorithm," in *Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings.* IEEE, 2003, pp. 260–267.

[121] C. Papazov and D. Burschka, "An efficient ransac for 3d object recognition in noisy and occluded scenes," in *Asian Conference on Computer Vision.* Springer, 2010, pp. 135–148.

[122] "Dokumentation von OpenPose - Benchmark," Eingesehen am 30.10.2020. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/speed_up_openpose.md

[123] A. Cano, "A survey on graphic processing unit computing for large-scale data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 1, p. e1232, 2018.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[124] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu, "Pose2seg: Detection free human instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 889–898.

[125] P. Corke, *Robotics, vision and control: fundamental algorithms in MATLAB® second, completely revised*. Springer, 2017, vol. 118.

[126] B. Jutzi and H. Gross, "Nearest neighbour classification on laser point clouds to gain object structures from buildings," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. Part 1, pp. 4–7, 2009.

[127] L. Priese, *Computer Vision: Einführung in die Verarbeitung und Analyse digitaler Bilder*. Springer-Verlag, 2015.

[128] S. G. Hoggar, *Mathematics of digital images: creation, compression, restoration, recognition*. Cambridge University Press, 2006.

[129] "Dokumentation von PCL," Eingesehen am 15.11.2020. [Online]. Available: https://pointclouds.org/documentation/

[130] "Dokumentation von OpenCV," Eingesehen am 15.11.2020. [Online]. Available: https://docs.opencv.org/

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

# Anhang

## Mathematische Grundlagen

### Transformationen

**Translation** wird definiert als eine Bewegung entlang einer oder mehreren Achsen. Mathematisch lässt sich Translation wie folgt beschreiben [38]:

$$\tilde{x} = [I|t] \cdot x = \begin{bmatrix} 1 & 0 & 0 & t \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & t \end{bmatrix} \cdot x \tag{6.1}$$

**Starre Transformation** wird aus Rotation und Translation gebildet. Rotation einer dreidimensionalen Matrix kann beispielsweise durch die Angabe der Rotation um eine der kartesischen Achsen definiert werden [125]:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{6.2}$$

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \tag{6.3}$$

$$R_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.4}$$

Somit lässt sich die starre Transformation wie folgt ausdrücken [38]:

$$\tilde{x} = [R_{x|y|z}|t] \cdot x \tag{6.5}$$

**Ähnlichkeitstransformation** wird aus einem Skalar, Rotation und Translation gebildet. Finden keine Rotation und Translation statt, so werden die Koordinaten um den Faktor s skaliert [38]:

$$\tilde{x} = [s \cdot R_{xyz}|t] \cdot x \tag{6.6}$$

**Eine affine Transformation** bewahrt Kollinearität und Parallelität der Punkte. Die Elemente der Matrix A sind dabei frei wählbar [38]:

$$\tilde{x} = [A] \cdot x = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \end{bmatrix} \cdot x \tag{6.7}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die Kovarianzmatrix einer sphärischen Region mit Radius R besteht im diskreten dreidimensionalen Fall aus neun Trägheitsmomenten und wird durch folgende Gleichung definiert [126][127][39]:

$$m_{ijk} = \frac{\sum_{I=1}^{M}(x_I - \bar{x})^i (y_I - \bar{y})^j (z_I - \bar{z})^k}{R^{i+j+k} \cdot M} \tag{6.8}$$

mit $\bar{x} = \frac{1}{M} \cdot \sum_{I=1}^{M} x_I$ der arithmetische Mittelwert (analog für $y$ und $z$). Die Mittelwerte der Koordinaten $x, y$ und $z$ stellen den geometrischen Schwerpunkt der Punkwolke, oft auch als Zentroid genannt, dar. Durch das Einsetzten der Komponenten aus der Gleichung 6.8 folgt schließlich die Kovarianzmatrix [126]:

$$C = \begin{bmatrix} m_{200} & m_{110} & m_{101} \\ m_{110} & m_{020} & m_{011} \\ m_{101} & m_{011} & m_{002} \end{bmatrix} \tag{6.9}$$

Durch die Kovarianzmatrix kann anschließend der Vektor der Eigenwerte $\lambda$ sowie die Eigenvektoren R ermittelt werden aus [128]:

$$|C - \lambda \cdot I| = 0 \tag{6.10}$$

$$R \cdot C = \lambda \cdot R \tag{6.11}$$

Um bei gleichzeitiger Beibehaltung der Aussagekraft die Dimension der Daten zu reduzieren, wird oft die Hauptkomponentenanalyse (weiters PCA - Principal Component Analysis) verwendet. Dies erfolgt über eine Transformation der Daten in einen neuen, nicht-korrelierten Set der Daten. Unter der Voraussetzung, dass die Kovarianzmatrix symmetrisch und positiv-semidefinit ist, sowie dass die Eigenwerte positiv definit sind und dass die Eigenwerte aufsteigend geordnet sind (d.h. $\lambda_1 > \lambda_2 > \lambda_3$), werden im dreidimensionalen Fall der Vektor der Koordinaten $X$ hinsichtlich der orthonormalen Basis im euklidischen Raum mithilfe der Eigenvektoren seiner Kovarianzmatrix in einen anderen Vektor $Y$ transformiert. Es gilt für die Komponenten des Datensets $Y$ [128]:

$$Y_i = X \cdot R_i \tag{6.12}$$

Demnach wird $M$ als die modale Matrix, die die Eigenvektoren $R_i$ als Spalten beinhaltet, folgend definiert [128]:

$$Y = X \cdot M^T \tag{6.13}$$

Durch die Berechnung der Kovarianz von $Y$ lässt sich überprüfen, dass die Komponenten voneinander unabhängig sind. Es gilt [128]:

$$Cov(Y) = Cov(X \cdot M^T) = M \cdot Cov(X) \cdot M^T = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \tag{6.14}$$



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

## Ergänzende Beschreibung der softwaretechnischen Umsetzung

Dieses Kapitel ergänzt die Beschreibung der Umsetzung der Algorithmen zur Erfassung der Pose des Bauteilss, sowie der Implementierung der Erfassung der Pose des Mitarbeiters.

### Prototyp zur Erfassung der Position des Bauteils

Zunächst wird das Prototyp zur Erfassung der Pose des Bauteils anhand eines Pseudocodes präsentiert, der den Ablauf beschreibt. Dies richtet sich an die Beschreibung der Umsetzung, die bereits im Kapitel 5.7 vorgestellt wurde. Zur Vollständigkeit werden alle verwendeten Parameter anhand der Abbildungen der Pipelines präsentiert (Abbildungen 6.1 und 6.2). Zur Reproduzierbarkeit wird an dieser Stelle auf die offizielle Dokumentation der Bibliotheken PCL [129] sowie OpenCV [130] verwiesen, die die Formale Struktur der Implementierung der Methoden definieren. Dies wird weiter um ausgewählte Abschnitte (blau markiert), deren Beschreibung im Kapitel 5.7 nicht stattgefunden hat, ergänzt. Angemerkt sei hier, dass die Funktion ***RecognisePPF*** sich aus Surface-Matching, Verfeinerung und Verifizierung, wie in Abbildung 6.2) dargestellt, zusammensetzt.

**Algorithmus 3** Genereller Ablauf des Prototyps des Teilsystems Bauteil

```
1:  while PointCloudViewerwasn'tclosed do
2:      for i = 1, 2, 3 do
3:          clouds[i]=CameraFrame(Camera[i])
4:      end for
5:      Scene = clouds[1]+clouds[2]+clouds[3]
6:      Preprocessing(Scene)
7:      Viewer.Display(Scene)
8:      SceneNormals = GetNormals(Scene)
9:      if FirstLoop == TRUE then
10:         Transform PCL to CV(Scene)
11:         Detected Object = RecognisePPF(Model, Scene)
12:         Transform CV to PCL(Detected Object)
13:         Viewer.Display(Detected Object)
14:     else
15:         ICP Tracking(Detected Object, Scene)
16:         Viewer.Display(Detected Object)
17:     end if
18: end while
```

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

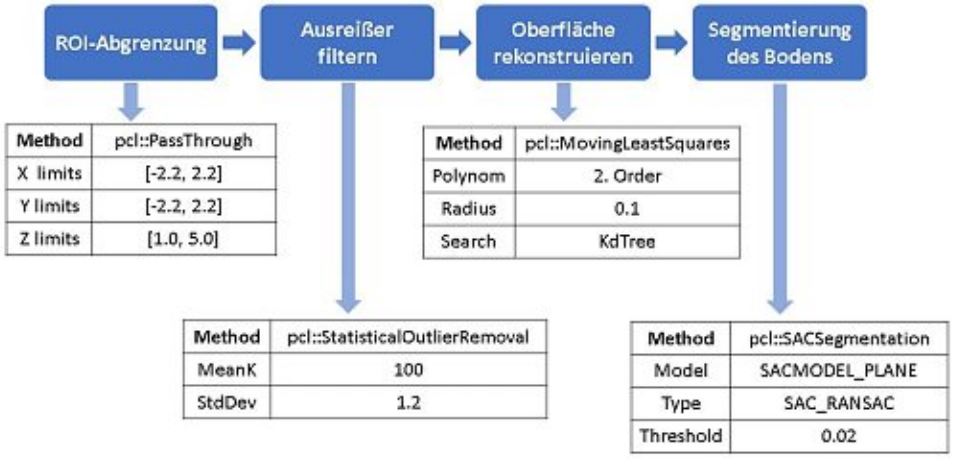



Abbildung 6.1: Verwendete Methoden und Parameter beim Preprocessing von Bauteil

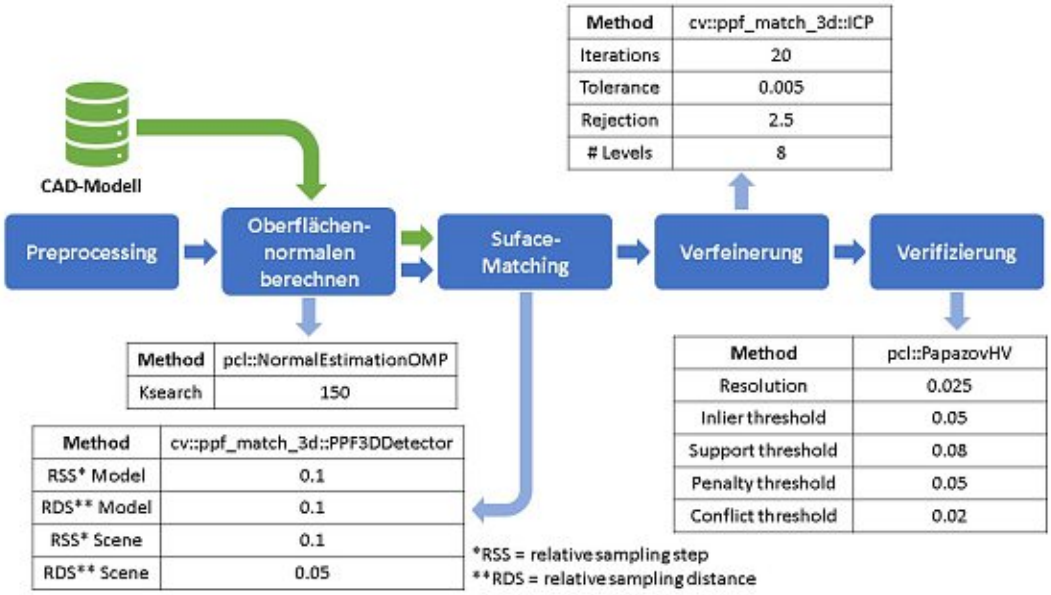



Abbildung 6.2: Verwendete Methoden und Parameter beim Preprocessing von Bauteil

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Listing 6.1: Berechnung und Korrektur der Normalen (**GetNormals**)

```
// Global Definitions
pcl::NormalEstimationOMP<pcl::PointXYZ, pcl::PointNormal> norm_est;

void GetNormals(pcl::PointCloud<pcl::PointXYZ>::Ptr Input, pcl::PointCloud
    <pcl::PointNormal>::Ptr Input_normals) {
    // The function takes an input cloud and an empty normals object. It
        calculates the normals and assigns them to the input_normals object
    norm_est.setKSearch(100);
    norm_est.setNumberOfThreads(2); //multithreading
    norm_est.setInputCloud(Input);
    norm_est.compute(*Input_normals);
    //correct the normals
    for (int i = 0; i < Input_normals->size(); i++) {
        if (Input_normals->points[i].normal_z < 0) {
            // turn around all the normals
            Input_normals->points[i].normal_x = Input_normals->points[i].
                normal_x;
            Input_normals->points[i].normal_y = Input_normals->points[i].
                normal_y;
            Input_normals->points[i].normal_z = Input_normals->points[i].
                normal_z;
        }
    }
}
```

Listing 6.2: Transformation von PCL to Mat(OpenCV) (**TransformPCLtoCV**)

```
void PCDtoMat(pcl::PointCloud<pcl::PointXYZ>::Ptr input, pcl::PointCloud<
    pcl::PointNormal>::Ptr input_normals, cv::Mat output) {
    // The function transfers the point cloud data between the two data
        frames. This enables the recognition by Point Pair Features

    for (int i = 0; i < input->points.size(); i++) {
        //XYZ
        output.at<float>(i, 0) = input->points[i].x;
        output.at<float>(i, 1) = input->points[i].y;
        output.at<float>(i, 2) = input->points[i].z;

        //normals
        output.at<float>(i, 3) = input_normals->points[i].normal_x;
        output.at<float>(i, 4) = input_normals->points[i].normal_y;
        output.at<float>(i, 5) = input_normals->points[i].normal_z;
    }
}
```



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Listing 6.3: Transformation von Mat(OpenCV) in PCL (**TransformCVtoPCL**)

```
void MattoPCD(cv::Mat input, pcl::PointCloud<pcl::PointXYZ>::Ptr output) {
    // The function transfers the point cloud data between the two data
       frames. This serves the further tracking by ICP

    // Resizing
    int rows = input.rows;
    output->width = rows;
    output->height = 1;
    output->points.resize(output->width * cloud_model->height);

    // XYZ transfering
    for (int j = 0; j < input->points.size(); j++) {
        output->points[j].x = input.at<float>(j, 0);
        output->points[j].y = input.at<float>(j, 1);
        output->points[j].z = input.at<float>(j, 2);
    }
}
```

Listing 6.4: **ICP Tracking**

```
// Global Definitions
pcl::IterativeClosestPoint<pcl::PointXYZ, pcl::PointXYZ> icp;
icp.setMaximumIterations(20);
icp.setMaxCorrespondenceDistance(0.25f);

// Inside of void Main()
while (!viewer->wasStopped())     {
    /*
    Previous processing (Frame captue, scene preprocessing, etc.)
    */

    if (loop < 1) {
        RecognisePPF(scene, scene_CV, model, detected);
    }
    else {
        icp.setInputTarget(scene);
        icp.setInputSource(detected); // Transformed model detected by PPF
        icp.align(*detected);
    }
}
```



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

### Teilsystem Mensch

Der Ablauf des Teilsystems Mensch wird weiters anhand eines Pseudocodes präsentiert. Analog zum Teilsystem Bauteil, wird zur Reproduzierbarkeit der Pipeline (Abbildung 6.3) auf die offizielle Dokumentation der beiden Bibliotheken verwiesen. Blau markierte Abschnitte des Codes, die im Rahmen der Beschreibung des Vorgehens im Kapitel 5.6 nicht ausführlich vorgestellt wurden, werden anschließend gelistet.

**Algorithmus 4** Genereller Ablauf der Pipeline des Teilsystems Mensch

```
1:  for i = 1, 2, 3 do
2:      clouds[i]=load("pathCloud[i]")
3:      amplitudes[i]=loadAmplitude("pathAmplitude[i]")
4:      Transform Amp to Image(amplitudes[i], ResultImage)
5:      Preprocessing(ResultImage)
6:      [Mask, Bounding Box]=SegmentationNetwork.detectHuman(ResultImage)
7:      PointCloudWithoutHuman = Extract(clouds[i], Mask)
8:      if i == 3 then
9:          CroppedImage = crop(Resized Bounding Box, ResultImage)
10:         2D Keypoints = PoseNetwork.detect(CroppedImage)
11:         Project 3D Keypoints(clouds[i],2D Keypoints)
12:         Recognise Gesture(3D Keypoints)
13:     end if
14: end for
```

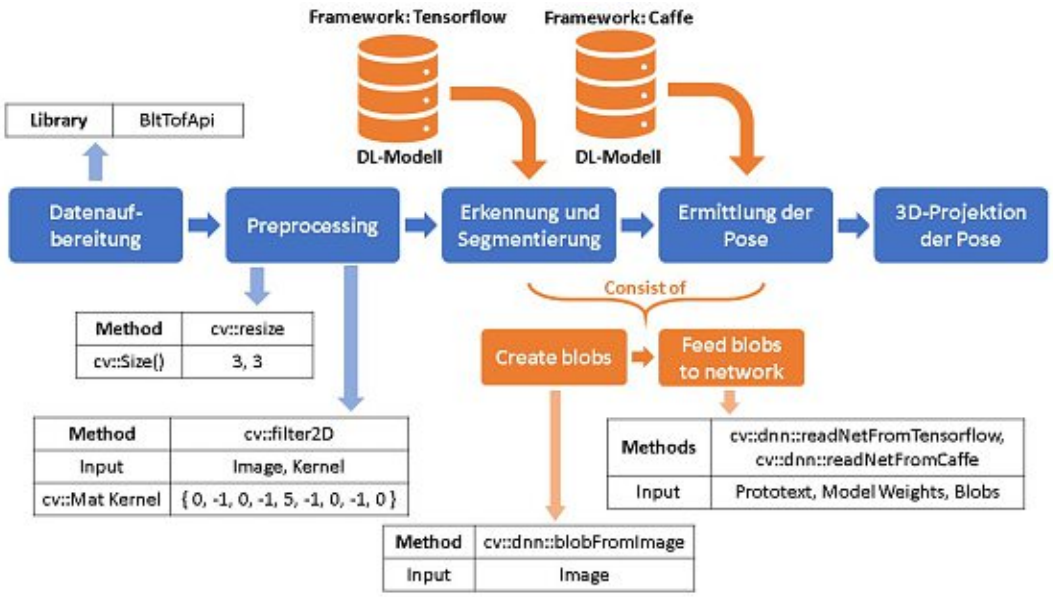



Abbildung 6.3: Verwendete Methoden und Parameter beim Teilsystem Mensch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Listing 6.5: Transformation von der Kamera-Aufnahme (Punktwolke mit Amplituden) in ein Grayscale-Bild (**Transform Amp to Image**)

```
// Global Definitions
float threshold = 550.0f;

void AmptoIm(pcl::PointCloud<pcl::PointXYZI>::Ptr input, cv::Mat output,
    float threshold) {
    for (int i = 0; i < input->points.size(); i++) {
        output.at<float>(i) = input->points[i].intensity / threshold;
    }
}
```

Listing 6.6: Extraktion des Menschen aus der Punktwolke (**PointCloudWithoutHuman**)

```
// Global Definitions
pcl::PointCloud<pcl::PointXYZ>::Ptr sceneSeg(new pcl::PointCloud<pcl::
    PointXYZ>());
sceneSeg->width = clouds[l]->width;
sceneSeg->height = clouds[l]->height;
sceneSeg->points.resize(sceneSeg->width * sceneSeg->height);

// indices of the mask in the point cloud
pcl::PointIndices::Ptr inliers(new pcl::PointIndices());

pcl::ExtractIndices<pcl::PointXYZ> extract; // indices extractor

// loop over 3 cameras
for (int l = 0; l < 3; l++) {
    /*
    Previous processing (preprocessing, segmentation, etc.)
    */

    // transform the local coorginates of the mask into the global
       coordinates using the bounding box (output of the segmentation)
    int StartingX = (BoundingBox.x / 3);
    int EndingX = StartingX + (BoundingBox.width / 3);
    int StartingY = (BoundingBox.y / 3);
    int EndingY = StartingY + (BoundingBox.height / 3);

    int j = 0;
    for (int i = 0; i < MaskLocations.size(); i++) {
        // coordinate transformation
        int Xcoor = StartingX + (MaskLocations[i].x) / 3;
        int Ycoor = StartingY + (MaskLocations[i].y) / 3;
        int value = Xcoor + (Ycoor * sceneSeg->width);
        if (i > 0) {
            if (value == inliers->indices[j]) {
                // repeating values
                continue;
            }
        }
```



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

```
            inliers->indices.push_back(value);
            j++;
        }
        // extract the mask from
        extract.setInputCloud(clouds[l]);
        extract.setIndices(inliers);
        extract.setNegative(true);
        extract.filter(*sceneSeg);

        /*
        Further processing
        */

        *clouds[l]=*sceneSeg; // after the keypoint assignment, the cloud can
            be finally overwritten
}
```

Listing 6.7: Winkel-basierte Erkennung der Pose (**Recognise Gesture**)

```
bool angleMMI(int l, bool keypointsBool[3][18], pcl::PointXYZ keypointsXYZ
    [3][18] ,int i, int j, int k, double limit) {
    // The functions takes coordinates as inputs, calculates the angles
        and returns TRUE, if the pose was recognised

    // if the keypoints were recognised (if 3D coordinates exist)
    if (keypointsBool[l][i] == true && keypointsBool[l][j] == true &&
        keypointsBool[l][k] == true) {
        // define the vectors between the keypoints
        Eigen::Vector3f Vecij(keypointsXYZ[l][i].x - keypointsXYZ[l][j].x,
            keypointsXYZ[l][i].y - keypointsXYZ[l][j].y, keypointsXYZ[l][i
            ].z - keypointsXYZ[l][j].z);
        Eigen::Vector3f Veckj(keypointsXYZ[l][k].x - keypointsXYZ[l][j].x,
            keypointsXYZ[l][k].y - keypointsXYZ[l][j].y, keypointsXYZ[l][k
            ].z - keypointsXYZ[l][j].z);

        // calculate the angle between the vectors
        double angle = std::atan2(Vecij.cross(Veckj).norm(), Vecij.dot(
            Veckj));

        // if the angle falls in the limits, the pose is recognised
        if (angle > limit * 0.7 && angle < limit * 1.3) {
            return true;
        }
        else
            return false;
    }
    else
        return false;
}
```



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN