UNIVERSITÄT
KOBLENZ · LANDAU
Fachbereich 4: Informatik

Personendetektion unter Verwendung von Tiefendaten

Masterarbeit
zur Erlangung des Grades
Master of Science
im Studiengang Informatik

vorgelegt von

Michael Kusenbach

Betreuer: Dipl.-Inform. Viktor Seib, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau
Erstgutachter: Prof. Dr.-Ing. Dietrich Paulus, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau
Zweitgutachter: Dipl.-Inform. Viktor Seib, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau

Koblenz, im April 2013

Kurzfassung

Die Personendetektion spielt eine wichtige Rolle in der Interaktion zwischen Mensch und Maschine. Immer mehr Roboter werden in menschlichen Umgebungen eingesetzt und sollen auf das Verhalten von Personen reagieren. Um das zu ermöglichen, muss ein Roboter zunächst in der Lage sein, die Person als solche zu erkennen. Diese Arbeit stellt ein System zur Detektion von Personen und ihrer Hände mittels einer RGBD-Kamera vor. Um eine Person zu erkennen werden zu Beginn modellbasierte Hypothesen über mögliche Personenpositionen aufgestellt. Anhand des Kopfes und Oberkörpers werden neu entwickelte Merkmale extrahiert, welche auf dem Relief und der Breite von Kopf und Schultern einer Person basieren. Durch die Klassifikation der Merkmale mit Hilfe einer Support Vector Machine (SVM) werden die Hypothesen überprüft und somit gültige Personenpositionen ermittelt. Dabei werden sowohl stehende, wie auch sitzende Personen anhand ihres sichtbaren Oberkörpers in verschiedenen Posen detektiert. Darüber hinaus wird ermittelt, ob die Person dem Sensor zugewandt oder abgewandt ist. Bei einer zugewandten Person werden zusätzlich, mit Hilfe der Farbinformation und der Entfernung zwischen Hand und Körper, die Positionen der Hände der Person bestimmt. Diese Information kann dann im nächsten Schritt zur Gestenerkennung genutzt werden.

Abstract

Human detection is a key element for human-robot interaction. More and more robots are used in human environments, and are expected to react to the behavior of people. Before a robot can interact with a person, it must be able to detect it at first. This thesis presents a system for the detection of humans and their hands using a RGB-D camera. First, a model based hypotheses for possible positions of humans are created to recognize a person. By using the upper parts of the body are used to extract, new features based on relief and width of a person's head and shoulders are extracted. The hypotheses are checked by classifying the features with a support vector machine (SVM). The system is able to detect people in different poses. Both sitting and standing humans are found, by using the visible upper parts of the person. Moreover, the system is able to recognize if a human is facing or averting the sensor. If the human is facing the sensor, the color information and the distance between hand and body are used to detect the positions of the person's hands. This information is useful for gestures recognition and thus can further enhances human-robot interaction.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Vereinbarung der Arbeitsgruppe für Studien- und Abschlussarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	ja ☐	nein ☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	ja ☐	nein ☐

Koblenz, den 18. April 2013

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In diesem Kapitel werden zunächst in dem ersten Abschnitt die Motivation und im zweiten Abschnitt die Ziele vorgestellt. Der dritte Abschnitt enthält eine Übersicht über den Aufbau der Arbeit.

1.1 Motivation

Immer mehr Roboter werden in menschlichen Umgebungen eingesetzt. Den Anfang machen dabei kleine Roboter wie Roomba, die autonom den Boden staubsaugen und bereits jetzt in Haushalten eingesetzt werden. Die Entwicklung geht dahin, dass die Roboter immer mehr Aufgaben übernehmen können.

Der Wettbewerb RoboCup@Home demonstriert diese Entwicklung. Verschiedene Forschergruppen aus der ganzen Welt zeigen die Einsatzfähigkeit ihrer Roboterprototypen im häuslichen Bereich. Neben vielen Aufgaben, wie beispielsweise Navigation, ist eine der wichtigsten Bereiche die Mensch-Roboter-Interaktion. Die Grundlage jeder Interaktion mit einer Person beruht auf der Detektion des Menschen. Zudem hängt die erfolgreiche Durchführung vieler Aktionen von der erfolgreichen Detektion der Person ab. Nur wenn eine Person korrekt detektiert wird kann z.B. die Person verfolgt werden, oder Gegenstände der Person übergeben werden (Abbildung 1.1).

Die Verwendung von Tiefeninformation zur Detektion der Person ist ein schon praktizierter Ansatz. Zur Personendetektion werden beispielsweise im Bereich des RoboCup@Home häufig 2D-Laserdaten verwendet, welche die Beine der Person detektieren. Das Problem solcher Verfahren ist oft die Genauigkeit. Objekte werden fälschlicherweise als Person detektiert und auf der Detektion aufbauende Aufgaben werden nicht erwartungsgemäß erfüllt.

Neben der Detektion der Person selbst sind weitere Informationen für eine erfolgreiche Interaktion von Bedeutung. Werden beispielsweise mehrere Personen

Abbildung 1.1: Der Roboter Lisa übergibt ein Objekt an eine Person.

detektiert ist oft eine Interaktion mit der nächstgelegenen zugewandten Person erwünscht. Die Information, ob eine Person zugewandt oder abgewandt, ist ist auch von Bedeutung wenn die Hand der Person detektiert werden soll.

Die Detektion der Hand stellt neben der Detektion der Person einen wichtigen Bestandteil für eine Interaktion dar. Die detektierte Hand kann als Teil einer Geste interpretiert werden und zur Kommunikation dienen.

Eine wichtige Information ist die Position der Person und ihrer Hand im dreidimensionalen Raum. Durch die Detektion einer Person mittels Tiefendaten liefert eine Detektion zugleich die Koordinaten der Person im Raum. Neben diesem Vorteil sind Tiefendaten zudem im Vergleich zu 2D-Kamerabildern robuster gegenüber Beleuchtungsänderungen. Bei der Verwendung von Tiefendaten ergeben sich neue Herausforderungen. Die Tiefeninformationen müssen beispielsweise in Form einer Punktewolke verarbeitet werden. Einzelne Bereiche dieser Punktewolke müssen analysiert werden um letztlich eine Person innerhalb der Tiefeninformation zu detektieren.

1.2 Zielsetzung

Das Ziel dieser Arbeit ist eine Personen- und Handdetektion zu entwickeln. Aufgrund der zuvor genannten Vorteile von Tiefendaten wird als Tiefensensor eine Microsoft Kinect verwendet. Dieser Sensor liefert eine Punktewolke, indem jeder einzelne Punkt zugleich auf ein RGB-Bild einer integrierten Farbkamera abgebildet wird.

Im späteren Anwendungsfall soll der Algorithmus beispielsweise auf dem Roboter Lisa der Universität Koblenz-Landau genutzt werden können. Die Personendetektion soll die Person unabhängig von ihrer Ausrichtung, wie beispielsweise

frontal, seitlich oder abgewandt zum Sensor, detektieren können. Indem für die Detektion nur der Kopf und Oberkörper der Person verwendet werden, soll es neben stehenden Personen ermöglicht werden sitzende Personen zu detektieren. Steht eine Person zugewandt zum Sensor sollen zudem noch die Hände der Person ermittelt werden. Da am Körper anliegende Hände in den Tiefendaten der Kinect nicht hinreichend gut sichtbar sind, sollen Hände nur detektiert werden, wenn sich diese vom Körper abheben.

Die einzelnen Ziele dieser Arbeit sind:

1. Auswahl möglicher Merkmale zur Detektion von Personen, Kopf und Händen
2. Entwurf eines geeigneten Softwaredesigns in ROS für die hier zu entwickelnden Algorithmen
3. Implementierung eines Algorithmus zur Detektion von Personen
4. Implementierung eines Algorithmus zur Positionsbestimmung von Kopf und Händen von detektierten Personen
5. Dokumentation und Evaluation der Ergebnisse

1.3 Aufbau der Arbeit

Die Arbeit ist wie folgt aufgebaut: In Kapitel 2 wird ein Überblick über bereits bestehende Verfahren zur Personen- und Handdetektion gegeben. Es werden zunächst in Abschnitt 2.1 Verfahren zur Personendetektion vorgestellt. Neben den wichtigsten Verfahren zur Ermittlung der Person auf 2D-Bildern liegt der Schwerpunkt dieses Kapitels in der Personendetektion unter Verwendung von Tiefendaten. Abschnitt 2.2 gibt einen Überblick über unterschiedliche Verfahren zur Handdetektion.

Kapitel 3 beschreibt die Vorgehensweise der Personendetektion in dieser Arbeit. Nach einer kurzen Systemübersicht (Abschnitt 3.1) werden die wichtigsten Bestandteile des Systems beschrieben. Hierzu wird das Verfahren der Kandidatensuche beschrieben um den Suchraum der Tiefendaten einzuschränken. Weiterhin werden die neu entwickelten Merkmale des Reliefs- und Breitenmerkmals eingeführt und die Klassifikation dieser Merkmale beschrieben.

Die nach erfolgreicher Personendetektion durchgeführte Handdetektion ist in Kapitel 4 erklärt. Auf Basis der Entfernung der Hand zum Körper und der Hautfarbe der Person werden die Hände der Person detektiert.

Im folgenden Kapitel 5 ist die Evaluation der Personen- und Handdetektion zu finden.

In Kapitel 6 wird die Arbeit zusammengefasst und ein Ausblick gegeben.

Kapitel 2

Verwandte Arbeiten

Dieses Kapitel bietet einen Überblick über bereits bestehende Verfahren zur Personen- und Handdetektion. Im ersten Abschnitt werden Verfahren zur Personendetektion vorgestellt. Der zweite Abschnitt beschreibt Verfahren zur Detektion der Hände.

2.1 Personendetektion

In der Literatur sind viele Ansätze zur Detektion von Personen vorhanden. Die verschiedenen Ansätze unterscheiden sich je nach verwendeten Sensoren und Aufgabenstellungen.

Im Folgenden werden verschiedene Verfahren zur Personendetektion kurz vorgestellt. Zunächst werden in Abschnitt 2.1.1 Verfahren erläutert, welche auf 2D-Bildern arbeiten. Anschließend werden Verfahren erläutert, welche auf 3D-Kameradaten (Abschnitt 2.1.2) und auf Laserdaten (Abschnitt 2.1.3) basieren.

Die Verfahren folgen einem gemeinsamen Grundschema: Vorverarbeitung, Merkmalsextraktion und Klassifikation (Abbildung 2.1). Sie unterscheiden sich je nach Art der Eingabedaten in der Art und Weise der Durchführung dieser Schritte.

2.1.1 Verfahren für 2D-Bilder

Die bildbasierten Verfahren arbeiten mit 2D-Bildern von Fotos bzw. Videos einer Kamera. Zur Personendetektion werden oft abgewandelte Verfahren aus der

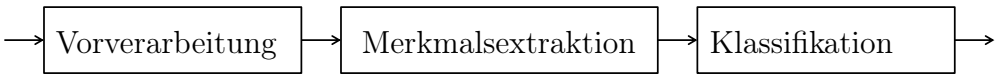


Abbildung 2.1: Allgemeiner Ablauf der Verfahren zur Personendetektion.

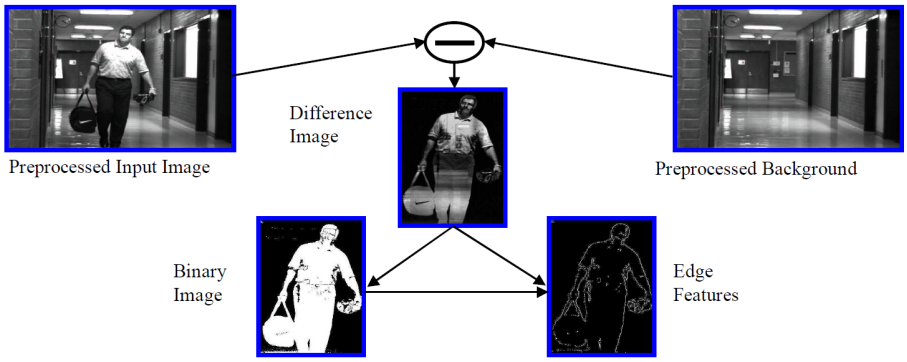


Abbildung 2.2: Merkmalsextraktion durch Differenzbilder aus [LZTS04]

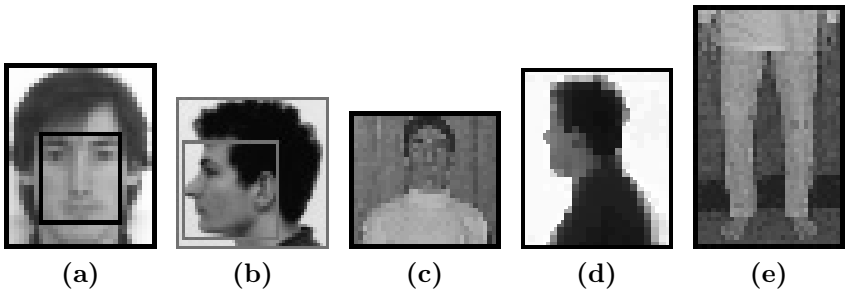
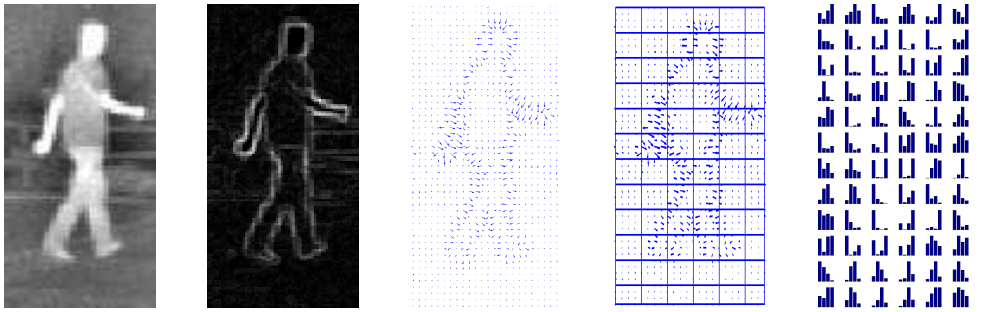


Abbildung 2.3: Verschiedene Teilbereiche einer Person aus [MSZ04]

Objektdetektion eingesetzt. Dabei werden Eigenschaften der Person genutzt, beispielsweise, dass sich eine Person bewegt. So verwendet Lee et al. für die Vorverarbeitung die Subtraktion des Hintergrundbildes [LZTS04]. Anhand des Differenzbildes lassen sich weitere Merkmale wie Kanten extrahieren (siehe Abbildung 2.2) und eine Kontur bestimmen.

Eine extrahierte Kontur kann anschließend mit bekannten Konturen aus einem Datenbestand verglichen werden, um eine Kontur als Person zu klassifizieren und von anderen beweglichen Objekten abzugrenzen. Daneben gibt es viele weitere Verfahren zur Detektion von Personen. Beispielsweise Verfahren aus dem Bereich des *Shape Matchings* [BMP01] oder unter Verwendung von Wavelets [VJS05]. Eine Methode die sich hierbei bewährt hat ist die Detektion von Teilen einer Person. Anstelle eine Person als Ganzes zu detektieren, detektieren Mikolajczyk et al. [MSZ04] einzelne Personenteile. Abbildung 2.3 zeigt Beispiele solcher Teile, wie das Gesicht, den Oberkörper oder die Beine.

Die Idee anhand einzelner Hinweise auf die Existenz einer Person zu schließen wird bei der Verwendung des *Implicit Shape Models* [LLS04] aufgegriffen. Hierbei wird ein Codebuch erlernt, das die lokalen Erscheinungen, die in einem Objekt vorkommen, speichert. Dies ist vergleichbar mit der Idee von Mikolajczyk et al.



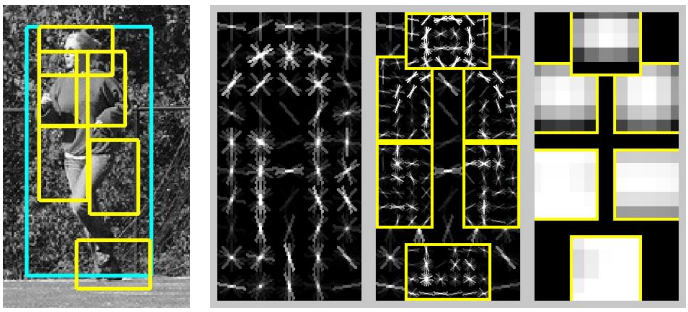
(a) Originalbild **(b)** Gradientennorm **(c)** Gradientenorientierung **(d)** Aufteilung in Zellen **(e)** Histogramm der Zellen

Abbildung 2.4: Veranschaulichung der Entstehung der Gradientenhistogramme [BBR+07].

Im Gegensatz zu diesem Ansatz wird den Einträgen im Codebuch keine Semantik, wie z.B. Bein oder Kopf, zugeordnet. Vereinfacht ausgedrückt speichert das Verfahren, das aus der Objekterkennung stammt, das lokale Aussehen verschiedener Abschnitte eines Objektes. Werden mehrere Abschnitte eines Objektes detektiert und stehen die detektierten Abschnitte in richtiger Relation zueinander, wird auf ein Objekt, in diesem Falle auf eine Person, geschlossen [LSS05].

Bei der Detektion von Personen auf 2D-Bildern hat sich vor allem der Ansatz der *Histogram of Oriented Gradients* (HOG) als sehr geeignet erwiesen [DWS+09]. Dalal et al. [DT05] verwenden diese Gradientenhistogramme zur Personendetektion. Ein zu klassifizierender Bildausschnitt wird gleichmäßig in Zellen zerlegt und für jede Zelle wird die Gradientenrichtung in ein 1D-Histogramm gespeichert (siehe Abbildung 2.4). Mehrere benachbarte Zellen werden zu einem Block zusammengeschlossen. Innerhalb eines Blocks werden Normalisierungen durchgeführt. Die Verkettung der Histogramme aller Blöcke ergibt einen Merkmalsvektor. Dieser Merkmalsvektor wird mit einer SVM klassifiziert. Zur Detektion auf einem Bild wird ein Detektionsfenster mit unterschiedlicher Skalierung über das Eingangsbild geschoben und jeweils klassifiziert. Auf dieser Art der Personendetektion bauen weitere Detektoren auf. Bertozzi et al. [BBR+07] verwenden zusätzlich zu einer normalen Kamera eine Infrarotkamera. Felzenszwalb et al. [FMR08] kombinieren den Ansatz der abschnittsweisen Detektion mit dem des HOG-Detektors (siehe Abbildung 2.5).

Zusammenfassend lassen sich für die Personendetektion auf 2D-Bildern zwei Grundideen festhalten. Die erste Idee besteht darin eine Person aufgrund von Teilabschnitten zu detektieren. Die zweite Vorgehensweise verwendet Bildausschnitte auf Basis eines Detektionsfensters, beispielsweise mit HOG, zur Personendetektion.



(a) (b)

Abbildung 2.5: HOG auf Abschnitte aus [FMR08]

2.1.2 Verfahren für 3D-Kameradaten

3D-Kameradaten können auf unterschiedliche Weise erstellt werden. Die drei gebräuchlichsten Möglichkeiten sind folgende:

- Stereokamerasysteme
- Systeme zur Messung der Laufzeit des Lichtes
- Systeme unter Verwendung von strukturiertem Licht

Bei Stereokamerasystemen wird eine Szene mit zwei versetzten Kameras aufgenommen. Anhand des vertikalen Versatzes beider Bilder ist die Bestimmung der Tiefe möglich. Neben den normalen Farbbildern der einzelnen Kameras steht die Entfernung einzelner Pixel zu Verfügung.

Der bekannteste Vertreter der Laufzeitverfahren ist die *Time-of-Flight-Kamera* (TOF). Die Kamera misst für jeden Pixel die Zeit, die benötigt wird bis ein ausgesendeter Lichtimpuls wieder auf den Sensor trifft.

Die Verwendung von strukturiertem Licht nutzt beispielsweise die Kinect. Ein ausgesendetes Infrarotmuster wird von einer Infrarotkamera aufgenommen. Anhand der Verzerrung des Musters wird die Entfernung jedes Pixels berechnet.

Die Gemeinsamkeit dieser Verfahren ist, dass die Tiefendaten auf zwei Arten interpretiert werden können. Die Tiefendaten können als 2D-Bild, Tiefenbild genannt, verarbeitet werden, indem an jedem Pixel die Entfernung zum Objekt dargestellt wird. Es ist ebenfalls möglich anhand bekannter Kameraparameter für jeden Pixel die Position im 3D-Raum zu berechnen. Die 3D-Punktewolke besteht aus 3D-Punkten mit x, y und z-Koordinaten.

Bei der Detektion von Personen in Tiefendaten werden beide Interpretationsmöglichkeiten verwendet.

Insgesamt lässt sich die Detektion von Personen auf kamerabasierten Tiefendaten in 3 Verfahren aufteilen:

1. Verfahren, die die Algorithmen aus 2D-Bildern (z.B. HOG) auf Tiefendaten übertragen
2. Verfahren, die die Form einer Person verwenden
3. Verfahren, die die Eigenschaften der Punktewolke nutzen

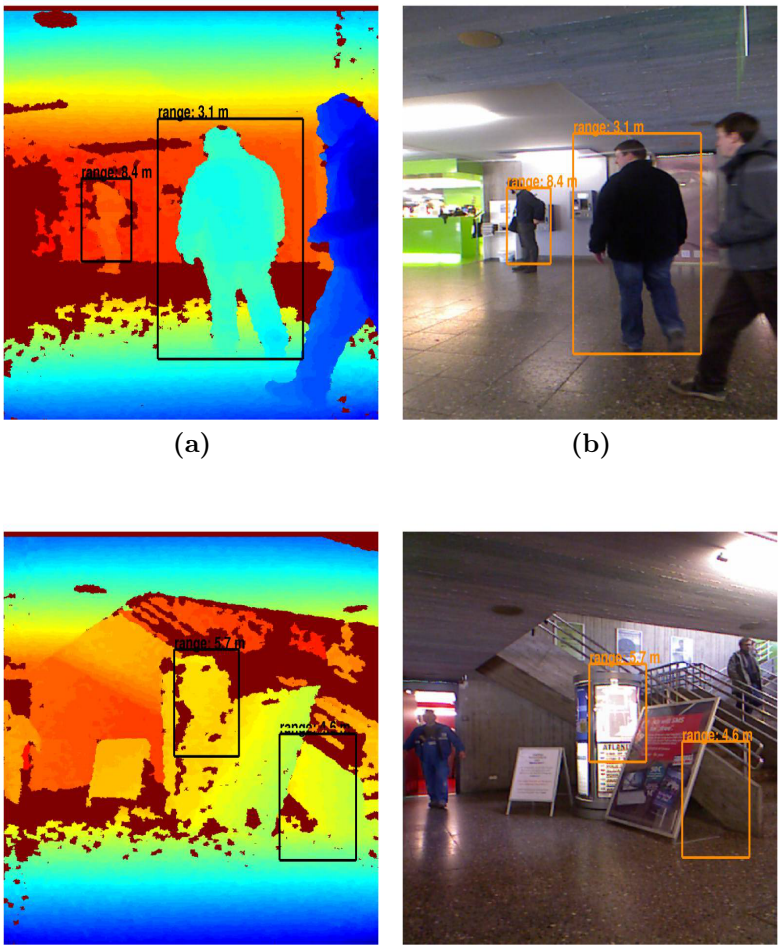
Verwendung von 2D-Bild-Verfahren

Zu den Verfahren, die bereits verwendete Algorithmen aus 2D-Bildern zur Personensuche auf Tiefendaten einzusetzen, gehören die Ideen von Spinello et al. [SA11] und Ikemura et al. [IF11]. Spinello et al. verwenden das HOG-Merkmal (*Histogram of Oriented Gradients*) auf Tiefendaten und erstellen ein HOD-Merkmal (*Histogram of Orinted Depths*). Beim HOD-Merkmal wird vergleichbar zum HOG-Merkmal vorgegangen. Anstelle der Gradienten werden nun die Tiefenwerte verwendet. Das Tiefenbild wird in einzelne Zellen eingeteilt und für jede Zelle ein 1D-Histogramm erstellt. Analog zum HOG-Detektor Verfahren wird mittels einer SVM klassifiziert.

Die verwendete Kinect als Tiefensensor liefert zusätzlich auch ein RGB-Farbbild. Auf diesem wird zusätzlich mit dem Standard HOG-Detektor nach Personen gesucht [SA11]. Aus einer Kombination von HOG- und HOD-Detektoren können die Vorteile der Tiefendaten genutzt werden. Durch die Verwendung der Tiefendaten ist der Algorithmus weniger anfällig gegenüber Beleuchtungsänderung, da beispielsweise Schattenwürfe in den Tiefendaten keine Auswirkungen haben. Der Nachteil des Ansatzes ist, dass es bei unaufgeräumten Umgebungen zu Fehlerkennungen kommen kann. Abbildung 2.6 zeigt ein solches Beispiel, in dem durch die Beschriftung, z.B. auf der Litfaßsäule, viele Gradienten entstehen und außerdem der Vergleich mit dem Histogramm auf den Tiefenwerten fehlschlägt.

Ikemura et al. [IF11] erweitern die Idee von HOG und entwickeln ein eigenes Merkmal. Wie beim HOD wird der Bildabschnitt in einzelne Zellen zerlegt und für jede Zelle ein Histogramm der Tiefenwerte erstellt. Zwischen den normalisierten Histogrammen der Tiefenwerte werden Ähnlichkeiten berechnet. Das erstellte *Relational Depth Similarity Feature* (RDSF) berechnet sich anhand der relativen Ähnlichkeit einzelner Zellen. Innerhalb eines zu untersuchenden Bildausschnittes werden alle Kombinationsmöglichkeiten der verschiedenen Zellen berechnet. Es werden auch Zellabschnitte unterschiedlicher Größe verglichen. Beispielsweise ein Zellabschnitt der Größe 2x2 mit einem Zellabschnitt der Größe 1x1 (zu sehen in Abbildung 2.7).

Aufgrund der enormen Anzahl der Kombinationsmöglichkeiten der Zellen ergeben sich viele Merkmale. Bei der Klassifikation wird deshalb Adaboost eingesetzt. Adaboost bewertet die einzelnen Merkmale und wählt die Merkmale, welche die beste Trennung in eine positive und negative Klasse ergeben. Es werden somit



(a) **(b)**

(c) **(d)**

Abbildung 2.6: Beispiele einer Personendetektion aus [SA11]: (a) und (b) zeigen eine korrekte Erkennung, (c) und (d) Fehlerkennungen

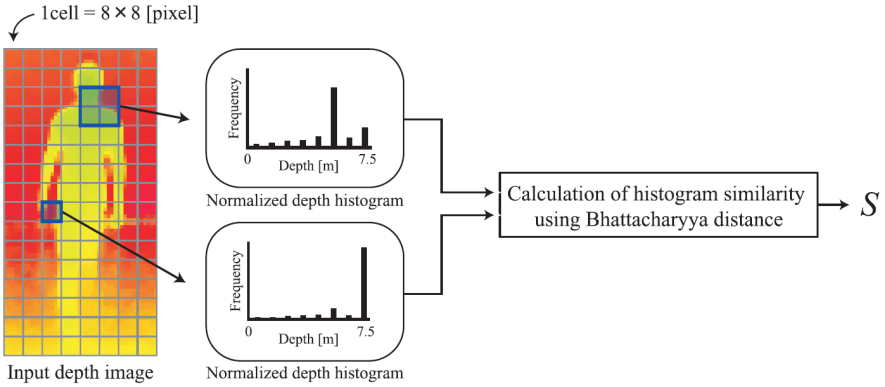


Abbildung 2.7: Berechnung des Relational Depth Similarity Feature (RDSF) aus [IF11]

automatisch die Teilbereiche im Detektionsfenster ausgewählt, die sich bei einem Vergleich der Tiefendatenhistogramme gut für eine Klassifikation eignen.

Zur Suche einer Person muss, wie bei HOG, ein Detektionsfenster über das gesamte Bild geschoben werden. Im Gegensatz zu HOG muss keine Bildpyramide mit unterschiedlicher Skalierung aufgebaut werden, da die Tiefe und somit die Größe des Detektionsfensters bekannt ist. Das Detektionsfenster kann, wie in Abbildung 2.8 dargestellt, optimal zur Tiefe gewählt werden.

Für eine Person können mehrere Detektionsfenster ein positives Resultat einer Personendetektion liefern. Im Gegensatz zu den Verfahren, die mit 2D-Bildern arbeiten, ergeben sich nun neue Möglichkeiten der Auswertung. Die Mittelpunkte der Detektionsfenster können als 2D-Koordinate des Kamerabildes oder als 3D-Koordinate als Punkte im Raum aufgefasst werden (Abbildung 2.9). Der Vorteil der Auffassung der Punkte in 3D ist, dass mit Hilfe des *Mean-Shift-Clustering* auch Personen entdeckt werden können, welche räumlich hinter anderen Personen stehen.

Verwendung der Form einer Person

Die bisher vorgestellten Ansätze verwenden die Form einer Person eher indirekt, indem beispielsweise die Form Einfluss auf Merkmalsvektoren von Histogrammen hatte. Die Silhouette einer Person kann auch als Form verarbeitet werden. Hierzu werden Schablonen (engl. Templates) eingesetzt [SM09] [XCA11] oder Verfahren aus dem *Shape-Matching* verwendet [HK10].

Satake et al. [SM09] verwenden Schablonen zur Personendetektion. Die drei Arten von Schablonen *Frontal*, *Links* und *Rechts* sind in Abbildung 2.10 dargestellt. Das Tiefenbild, in diesem Fall das Tiefenbild einer Stereokamera, wird in einer Höhe von 0,7 bis 2 Metern nach der Schablone durchsucht. Für die Berechnung

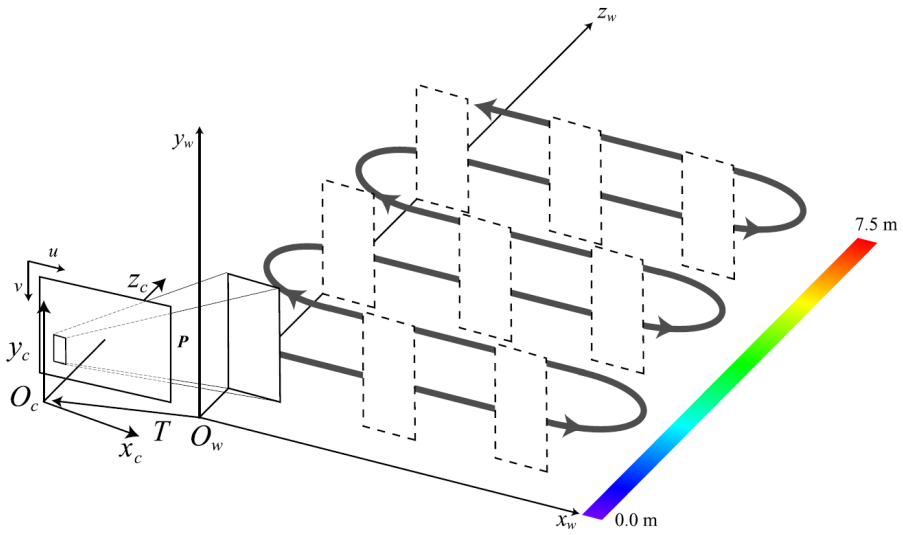


Abbildung 2.8: Nutzen der Tiefeninformation für das Detektionsfenster [IF11]

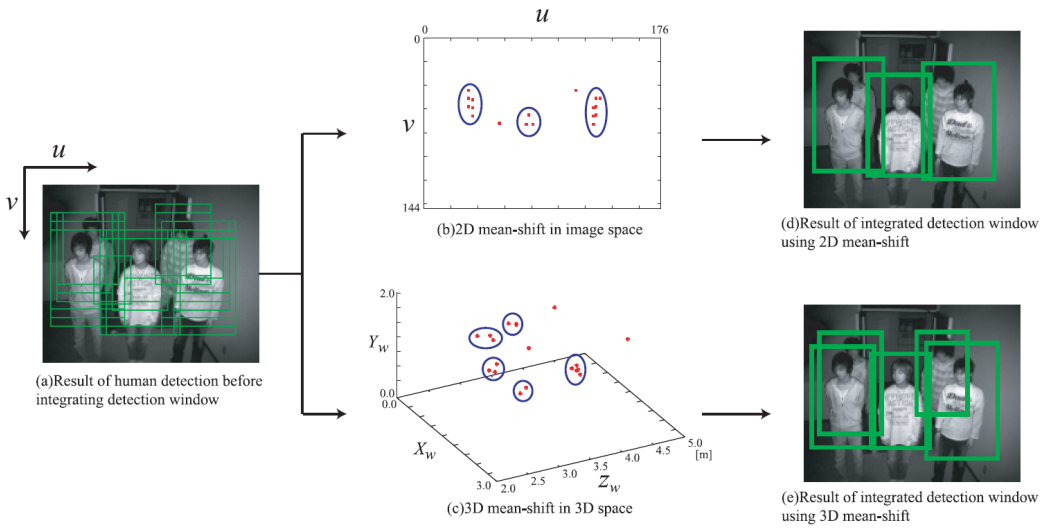


Abbildung 2.9: Integration des Detektionsfensters mit Hilfe von *Mean-Shift-Clustering* aus [IF11]

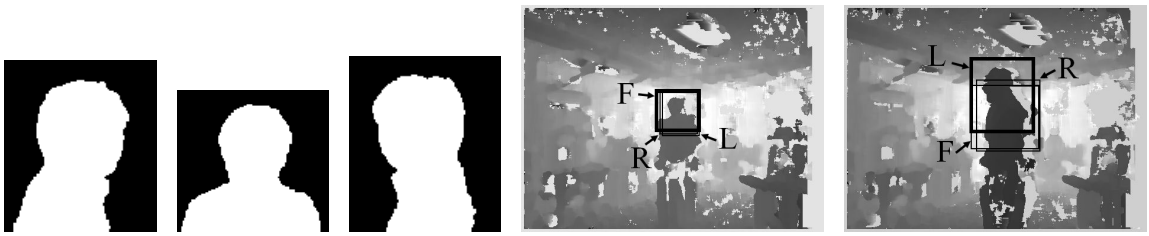


Abbildung 2.10: Schablonen (a),(b),(c) und die Detektion im Tiefenbild (d), (e) aus [SM09]

der Ähnlichkeit eines Bildabschnittes mit der Schablone wird die Summe der quadratischen Distanzen verwendet. Wie auf Abbildung 2.10d und 2.10e zu erkennen, können alle drei Arten von Schablonen für eine Person gefunden werden. Die Suche des Schablonen ist deshalb nicht sehr genau und enthält eine hohe Anzahl an Fehlerkennungen, weshalb diese oft in Kombination mit weiteren Verfahren angewandt wird. Satake et al. untersuchen deshalb den Bildausschnitt des RGB-Bildes der Kamera weiter um die Fehlerkennungen zu reduzieren.

Schablonen können auf unterschiedliche Arten genutzt werden. Während Satake et al. die Form als Schablone verwenden, verwenden Xia et al. [XCA11] die Kanten des Kopf- bzw. Schulterbereiches. Mit Hilfe eines Medianfilters wird zunächst das Rauschen auf den Tiefendaten reduziert. Auf dem vorbereiteten Tiefenbild werden mit Hilfe des Canny-Algorithmus zur Kantendetektion die stärksten Kanten extrahiert (2.11b). Zur Detektion möglicher Personenkandidaten wird der *Chamfer Matching-Algorithmus* eingesetzt. Es handelt sich um ein Verfahren aus dem Bereich des Shape Matchings [TSTC03]. Die Schablone in Abbildung 2.11c wird verwendet um ähnliche Bildbereiche im Tiefenbild zu detektieren. Die Mittelpunkte des Detektionsfensters, an denen die Schablone Ähnlichkeiten zum Bildbereich aufgewiesen hat, sind in Abbildung 2.11d gelb eingezeichnet.

Der Einsatz von Schablonen zu Personensuche muss sich nicht auf das 2D-basierte Vergleichen der Form beschränken. So verwenden Xia et al. im Anschluss an die 2D-Schablonen-Suche einen 3D-Schablonen-Vergleich um die Detektion zu verbessern. Das Problem liegt darin, dass sich die Form einer Person, abhängig vom Betrachtungswinkel verändert (Abbildung 2.12a). Die Idee ist anstelle eines komplexen Modells das Modell möglichst einfach zu halten, um Ähnlichkeiten bei verändertem Betrachtungswinkel zu erhalten. Xia et al. verwenden anstelle eines komplexen Kopfmodells lediglich eine Halbkugel als Modell für den Kopf. Die aus dem 2D-Vergleich entstehenden Kandidaten (Abbildung 2.11d) können so reduziert werden (Abbildung 2.11e). Für den Vergleich wird der quadratische Fehler zwischen dem Kopfbereich der Person und der Halbkugel berechnet. Ist

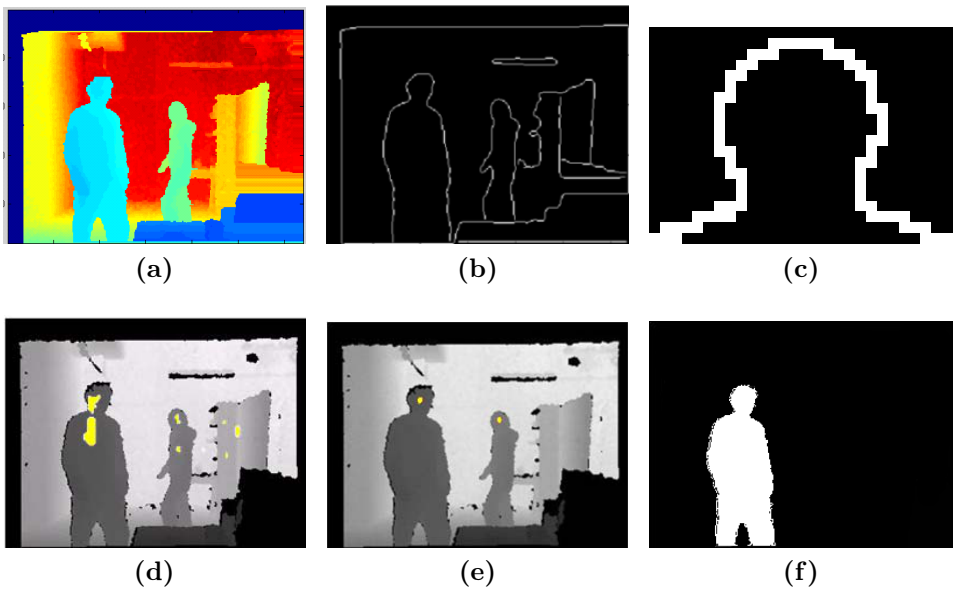


Abbildung 2.11: Ablauf nach Xia [XCA11]

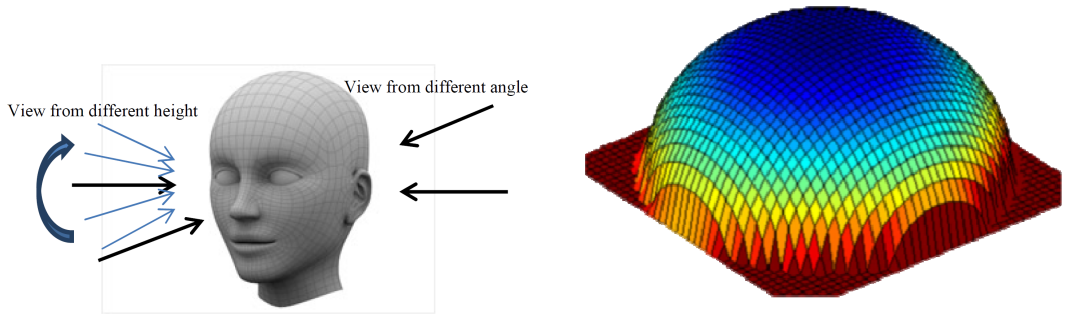
der Bildabschnitt eines Kopfs durch die zwei verschiedenen Schablonen bestätigt kann anschließend die Person als Ganzes detektiert werden. Ein Region-Growing-Algorithmus gestartet im Kopfbereich separiert die Person (Abbildung 2.11f).

Es ist auch möglich die Form der Person zu nutzen, ohne eine feste Schablone zu verwenden. Hordern et al. [HK10] verwenden ein Verfahren um die 2D-Silhouette einer Person zu erlernen und anschließend zu klassifizieren.

Der von Hordern et al. entwickelte Ansatz extrahiert aus der Punktewolke eine 2D-Silhouette, die anhand ihrer Kontur klassifiziert wird. Im Folgenden wird der Ablauf, der in Abbildung 2.13 skizziert ist, weiter erläutert.

Im ersten Schritt, der Silhouettenextraktion, wird die 3D-Punktewolke auf die horizontale Ebene projiziert [HK10]. Eine solche Projektion ist in Abbildung 2.14a visualisiert. Die Darstellung ähnelt der Draufsicht auf die Szene aus der Vogelperspektive. Bei der Projektion wird ein 2D-Histogramm erstellt, indem die Häufigkeit der Punkte auf der 2D-Ebene gespeichert wird. Personen und andere Objekte bilden Anhäufungen in diesem Histogramm. Mit Hilfe einer Blob-Extraktion werden diese Anhäufungen lokalisiert.

Regionen, welche Anhäufungen im 2D-Histogramm bilden, werden anschließend vertikal projiziert. Abbildung 2.14b zeigt eine solche Projektion der 3D-Punkte in Richtung der Sensorblickrichtung. Punkte, die nicht innerhalb der Anhäufungen liegen, beispielsweise die Punkte der Wand im Hintergrund, werden verworfen. Das Ergebnis dieser Projektion ist ein 2D-Projektionsbild, die Silhouette, welche den Kopf mit Schultern, Arme, Torso und evtl. die Beine enthält (Abbildung 2.15a).



(a) (b)

Abbildung 2.12: Verwendetes Modell einer Halbkugel als 3D-Kopfmodell [XCA11]

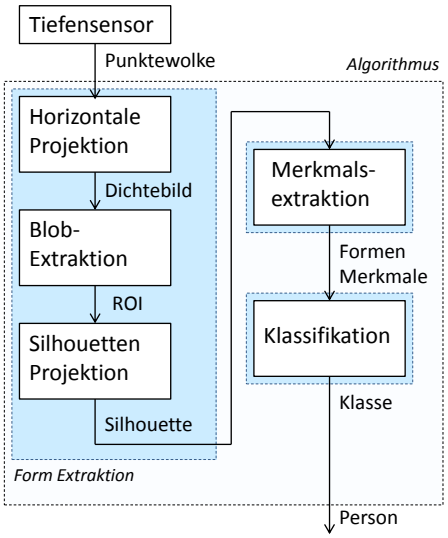
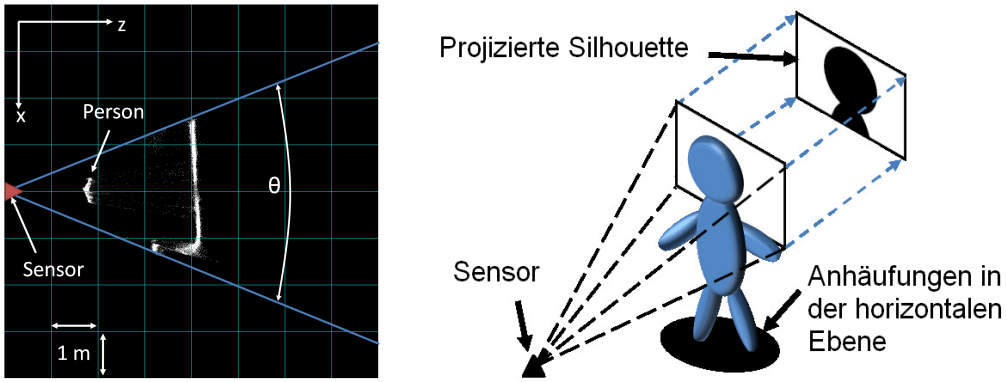
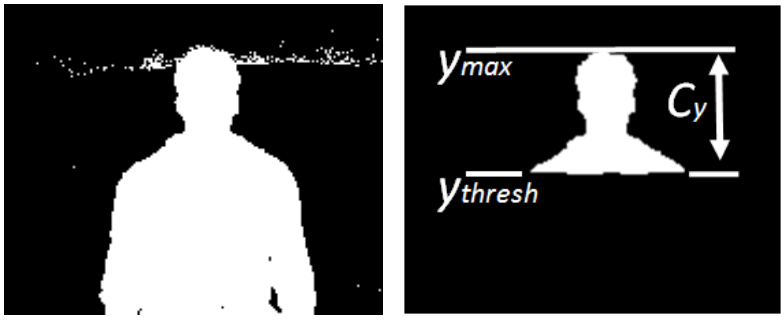


Abbildung 2.13: Ablauf und Informationsfluss der Personendetektion vgl. [HK10]



(a) Dichtebild **(b)** Projektion der Silhouette

Abbildung 2.14: Horizontale Projektion der 3D-Punktewolke [HK10]



(a) Personensilhouette entstanden durch vertikale Projektion der ROI **(b)** Reduzierte Silhouette

Abbildung 2.15: Extraktion der Silhouette [HK10]

Bereiche der Silhouette, vor allem die Arme, können je nach Körperstellung verschiedenste Formen annehmen. Aus diesem Grund wird die Silhouette auf den Bereich des Kopfes beschränkt. Vom höchsten Punkt der Silhouette y_{max} in Metern wird eine Konstante von $C_y = 0.4m$ subtrahiert und alle Punkte unterhalb des Schwellwertes y_{thresh} verworfen.

$$y_{thresh} = y_{max} - C_y \tag{2.1}$$

Ein positiver Nebeneffekt ist, dass die entstehende Silhouette, dargestellt in 2.15b, invariant zur Größe der Person ist. Das Problem der Personendetektion wird auf das Problem der Erkennung einer Form reduziert.

Bei der Merkmalsextraktion werden anhand der Formen Merkmale erzeugt. Um die Merkmale zu klassifizieren werden Fourierdeskriptoren verwendet, da diese sich bereits bei Vergleichen von Formen bewährt haben [ZL02]. Zur Merkmalsextraktion einer Form werden folgende Schritte durchgeführt [HK10]:

1. Die Kontur $c(t)$ wird extrahiert.

2. Die Kontur wird gleichmäßig an N Stellen abgetastet und auf wenige Punkte reduziert (Abbildung: 2.16a):

$$c(t) \rightarrow s(t) \text{ mit } s(t) = (x(t), y(t)) \tag{2.2}$$

3. Der Schwerpunkt der Kontur wird bestimmt:

$$c(t) \rightarrow (x_c, y_c) \tag{2.3}$$

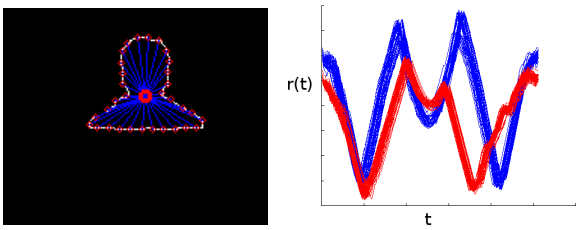
$$\text{mit } x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t) \text{ und } y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t) \tag{2.4}$$

4. Die euklidischen Distanzen zwischen dem Schwerpunkt und den diskreten Konturstellen werden bestimmt. Alle Distanzen einer Form ergeben angeordnet die Signatur $r(t)$ der Form:

$$s(t) \rightarrow r(t) \tag{2.5}$$

$$\text{mit } r(t) = \sqrt{([x_t - x_c]^2) + ([y_t - y_c]^2)} \tag{2.6}$$

$$t = 0, 1, ..., N-1 \tag{2.7}$$



(a) Beispielkontur **(b)** Formensignaturen

Abbildung 2.16: Extraktion der Signatur (Abb. 2.16b) aus einer Beispielkontur (Abb. 2.16a) [HK10]

Abbildung 2.16b zeigt 50 solcher Signaturen, welche von drei verschiedenen Personen erstellt wurden. Die Aufnahmen wurden frontal (blau) und im Profil (rot) durchgeführt.

5. Mit Hilfe der diskreten Fourier-Transformation (DFT) (2.8) wird die Signatur $r(t)$ der Form zu einem Merkmalsvektor $\mathbf{f}$.

$$s_k = \frac{1}{N} \sum_{t=0}^{N-1} r(t) e^{-j2\pi nk/N}, k = 0, 1, ..., N-1 \tag{2.8}$$

Indem der Vektor durch den Gleichspannungsanteil geteilt wird, wird der Vektor skalenunabhängig (2.9). Da die Signatur nur aus reellen Zahlen besteht gibt es nur $N/2$ verschiedene Frequenzen. Die Größe der Form und somit die Entfernung der Person beeinflussen daher nicht den Merkmalsvektor.

$$\mathbf{f}_{k-1} = \frac{|s_k|}{|s_0|}, k = 1, ..., N/2 \tag{2.9}$$

Der Merkmalsvektor $\mathbf{f}$ beschreibt die Form. Um eine Aussage darüber zu treffen, ob es sich um eine Person handelt, wird im nächsten Schritt die Form klassifiziert.

Zu Silhouettenklassifizierung wird eine Support Vector Machine (SVM) eingesetzt. Diese wird zuvor mit bekannten Klassen trainiert. Die Klassen sind Formen einer frontal stehenden Person, einer um 90° gedrehten seitlich stehenden Person und die Klasse der Objekte. Der von Hordern et al. vorgestellte Ansatz kommt ohne Schablonen aus und erlernt die Form einer Person.

Verwendung der Eigenschaften der Punktewolke

Zusammenfassend für die Verfahren, welche die Form einer Person verwenden, lässt sich festhalten, dass der Oberkörper, genauer der Kopf- und Schulterbereich, für

die Detektion verwendet werden. Das Prinzip der Reduzierung der komplexen 3D-Daten auf einfache 2D-Silhouettenformen oder 3D-Halbkugeln hat sich bewährt.

Daneben gibt es Verfahren welche die Eigenschaften der Punktewolke verwenden. Hierzu zählen die Ansätze von Bajracharya et al. [BMH+09] und Hegger et al. [HHKP12].

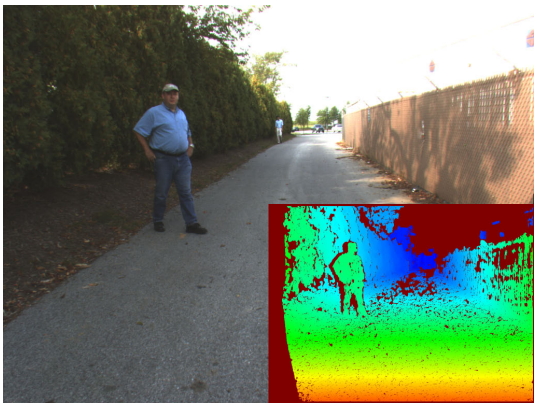
Bajracharya et al. geht zunächst ähnlich wie Hordern et al. vor und betrachtet die 3D-Punkte aus der Vogelperspektive. Die Tiefendaten, in diesem Falle erzeugt durch ein Stereosystem, werden auf die horizontale Ebene projiziert. Lokale Maxima innerhalb der Ebene stellen mögliche Personenpositionen dar (Abbildung 2.17b). Eine gradientenbasierte Suche auf der horizontalen Projektion ermittelt lokale Maxima und segmentiert Bereiche möglicher Personen. Für die Segmentierung wird die zu erwartende Breite und Tiefe einer normalen Person verwendet. Segmente, welche zu klein für eine Person sind, werden zu größeren Segmenten verschmolzen. Segmente, die zu groß für eine Person sind, werden verworfen und so die Anzahl der zu klassifizierenden Segmente reduziert. Anhand eines ermittelten Segmentes lässt sich eine 3D-Region bestimmen in der sich Punkte möglicher Personen befinden.

Zur Merkmalsextraktion werden verschiedene geometrische Merkmale auf den Punkten innerhalb der 3D-Region berechnet. Zu diesen Merkmalen zählen unter anderem die Momente wie beispielsweise die Varianz der Punkte oder die Eigenwerte der Streumatrix der 3D-Punkte. Zusätzlich werden noch Eigenschaften, welche Breite, Höhe, Tiefe oder das Volumen betreffen, mit als Merkmal aufgenommen und anschließend klassifiziert.

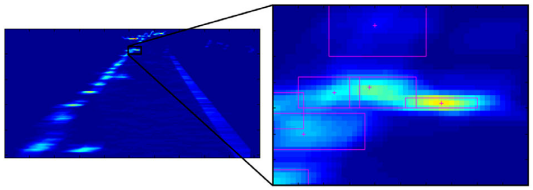
Die Eigenschaften der Punktewolke werden je nach Einsatzgebiet unterschiedlich verwendet. Während Bajracharya et al. mit Hilfe der Merkmale, wie den Momenten und Volumen eines Objektes im Außenbereich Fußgänger von z.B. Bäumen trennt, gibt es im Innenbereich eine andere Situation. Hier kommt es darauf an, Personen von z.B. Tischen, Wänden, Decken usw. zu unterscheiden.

Hegger et al. [HHKP12] verwendet aus diesem Grund lokale Oberflächennormalen (*Local Surface Normals, LSN*) einer 3D-Punktewolke um Personen zu detektieren.

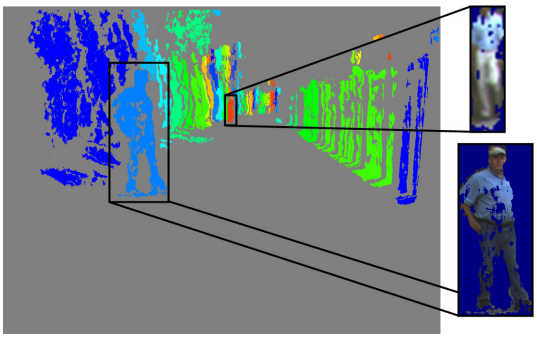
Die in Abbildung 2.18 gezeigte Vorgehensweise in 4 Schritten wird im Folgenden zusammengefasst. Im ersten Schritt, der Vorverarbeitung, wird die Anzahl der 3D-Punkte reduziert. Punkte oberhalb von 2 Metern werden verworfen und die Punktewolke ausgedünnt. Zudem werden zu jedem Punkt der verbleibenden Punktewolke lokale Oberflächenormalen bestimmt. Die Oberflächennormale eines Punktes wird ermittelt, indem eine Ebene durch die k-nächsten Punkte bestimmt wird. Die Normale der am besten passenden Ebene stellt die Normale des Punktes dar. Ursprünglich entwickelt wurde dieses Verfahren um Ebenen in RGBD-Punktewolken zu detektieren [HHRB11].



(a) Eingabedaten



(b) Projektionsansicht



(c) Segmentierte Bereiche der Tiefendaten

Abbildung 2.17: Segmentierung der Tiefendaten eines Stereosystems [BMH+09].

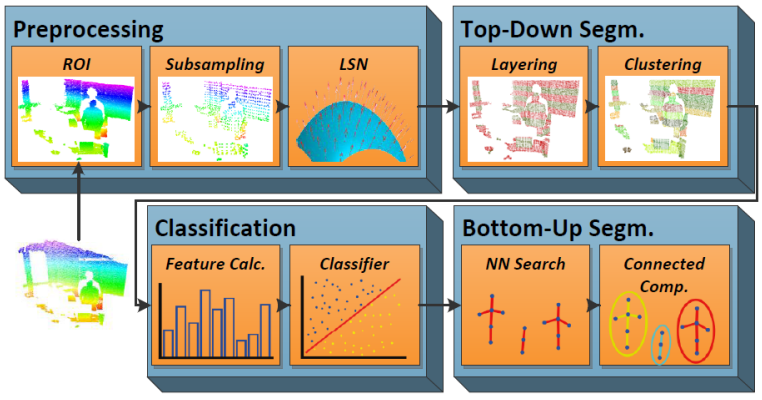
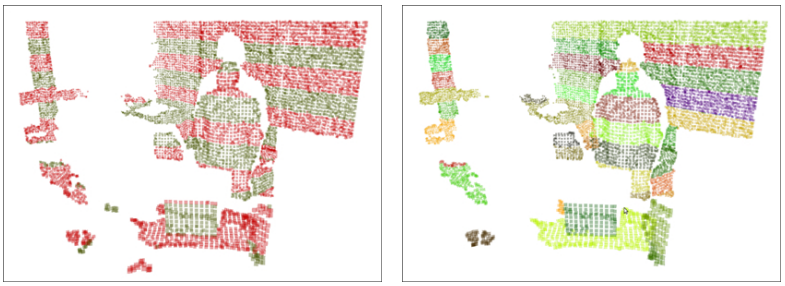


Abbildung 2.18: Ablaufplan zur Personendetektion nach Hegger et al. [HHKP12]



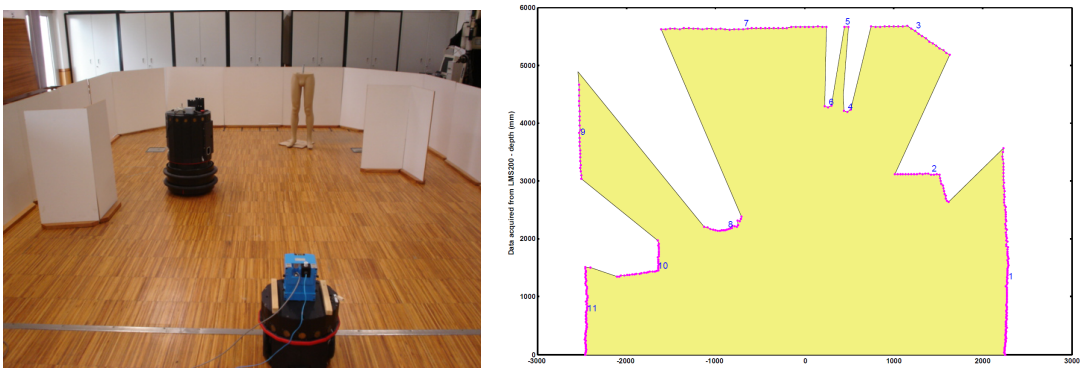
(a) Zerlegung in Schichten **(b)** Cluster jeder Schicht

Abbildung 2.19: Segmentierung der Punktewolke in Cluster nach Hegger et al. [HHKP12]

Im zweiten Schritt wird die Punktewolke mit Hilfe einer Top-Down Segmentierung von oben nach unten in 8 Schichten zu je 25 cm zerlegt (Abbildung 2.19a). Auf jeder Schicht wird anschließend ein euklidisches Clustering durchgeführt (Abbildung 2.19b). Liegt die Distanz zwischen zwei Punkten innerhalb eines gewissen Schwellwertes wird der Punkt zum Cluster hinzugefügt.

Für den dritten Schritt der Klassifikation wird auf jedem erzeugten Cluster ein Histogramm der lokalen Oberflächennormalen erstellt. Die ermittelte Normale in jedem Punkt innerhalb eines Clusters wird in ein Histogramm eingetragen. Als Merkmal jedes Clusters wird das Histogramm und die Breite und Höhe des Clusters gespeichert. Das Merkmal wird mit Hilfe eines *Random Forest* Klassifikationsverfahrens klassifiziert. Dieses Verfahren des überwachten Lernens nutzt die zuvor bekannten Trainingsdaten um einen Wald aus mehreren Entscheidungsbäumen zu erstellen.

Im vierten und letzten Schritt wird eine Bottom-Up Segmentierung durchgeführt. Es werden Cluster betrachtet, welche als Person klassifiziert wurden. Weisen die Cluster eine geringe euklidische Distanz zueinander auf, werden sie zu einer Person gehörend gespeichert. Eine Person wird als erfolgreich detektiert betrachtet, wenn mindestens 3 Cluster zur Person gehören.



(a) **(b)**

Abbildung 2.20: Segmentierung der Laserdaten [PN05]

2.1.3 Verfahren für Laserentfernungsdaten

Bei den laserbasierten Verfahren wird die Entfernung eines Objektes zum Sensor mit Hilfe eines Lasers gemessen. Bei 2D-Laserscannern wird der Laserstrahl durch einen sich drehenden Spiegel abgelenkt. Die erzeugten Tiefendaten liegen auf einer 2D-Ebene und decken typischerweise einen Bereich zwischen 180 bis 270 Grad ab. Die Messdaten entsprechen keinem geschlossenen Tiefenbild, sondern einer Reihe aufeinanderfolgender Punkte zu denen Abstand und Winkel im Verhältnis zum Sensor bekannt sind.

Aufgrund dieser Eigenschaft des Sensors verwenden die laserbasierten Verfahren zuerst eine Segmentierung der Tiefendaten. Zu diesem Zweck gibt es verschiedene Verfahren [PN05]. Das am häufigsten verwendete Verfahren ist das *Jump Distance Clustering* (JDC) [PN05]. Der Abstand zwischen den aufeinanderfolgenden Punkten wird bestimmt. Sobald der Abstand größer als ein gewisser Schwellwert ist, wird ein neues Segment erzeugt. In Abbildung 2.20 sind die Segmente nummeriert dargestellt.

Die verschiedenen Verfahren zur laserbasierten Personenerkennung lassen sich grob in drei Vorgehensweisen einteilen.

1. Die Person wird als lokales Minimum detektiert.

2. Die Beine einer Person werden unter Verwendung geometrischer Modelle (Halbkreis) detektiert.

3. Die Merkmale der Lasersegmente werden erlernt und anschließend klassifiziert.

Detektion mit Hilfe der lokalen Minima

Die einfachste Methode zur laserbasierten Personenerkennung ist die Person als lokales Minimum zu detektieren [SBFC03, BBCT05, FHM02]. Verwendet wird dazu ein Laserscanner, der einen 2D-Scan auf Höhe der Beine durchführt. Aus Abbildung 2.20 wird ersichtlich, dass die Segmente der beiden Beine der Puppe, nummeriert mit 4 und 6, lokale Minima bilden.

Nach der Segmentierung von rechts nach links sind auch andere Objekte, beispielsweise die Ecke (Segment 2), ein lokales Minimum. Um die Detektion zu verbessern werden nun oft weitere Bedingungen an eine Person gestellt, um sie als Person zu klassifizieren. Bennewirtz et al. [BBCT05] verwenden zusätzlich ein Tracking der Minima um anhand der Bewegung auf eine Person schließen zu können. Die Eigenschaft, dass eine Person oft durch zwei Beine, die nahe beieinander stehen im Laserscan zu vertreten ist, wird ebenfalls genutzt [TC05]. Eine Möglichkeit ist die Größe des Segmentes zu benutzen um falsche Detektionen zu verringern. Segmente der Wände sind beispielsweise zu groß für ein Bein und können ausgeschlossen werden.

Detektion mit Hilfe geometrischer Modelle

Die Verwendung eines geometrischen Modells ist die Fortführung des Grundgedankens der Verwendung der Größe des Segmentes. Ein Beinabschnitt wird nicht nur durch seine Größe, sondern ebenfalls durch seine geometrischen Eigenschaften bestimmt. Eines der effektivsten Methoden ist die Klassifikation der Segmente in Halbkreisen bzw. Bögen und Geraden. Xavier et al. [XPCR05] nutzen die Varianz des inneren Winkels eines Segments (*Internal Angle Variance, IAV*). Bei einem Segment, bestehend aus mehreren Punkten, wird jeweils der Winkel zwischen Anfangs- und Endpunkt des Segments und den jeweiligen Punkten innerhalb des Segments gemessen. Abbildung 2.21a demonstriert dies anhand eines Beispiels mit vier Punkten wobei P1 der Anfangspunkt und P4 der Endpunkt des Segments ist. Für jeden Punkt innerhalb des Segments (hier P2 und P3) ergibt sich ein Winkel. Liegt der Winkel durchschnittlich zwischen 90° und 135° liegt bei dem Segment ein Halbkreis bzw. ein bogenförmiges Segment vor. Bei einem geraden Segment, dargestellt in Abbildung 2.21b liegen die inneren Winkel bei ca. 180°. Anhand von festgelegten Bedingungen wie beispielsweise, dass ein Beinsegment eine bogenförmige Form einer bestimmten Größe haben sollte, lässt sich auf eine mögliche Person schließen.

Das Problem bei dieser Art der Detektion ist, dass die Beine im Laserscan oft sehr unterschiedlich aussehen und nicht immer auf das verwendete Modell eines Bogens bzw. Halbkreises passen. Abbildung 2.22 zeigt die Verschiedenartigkeit von Segmenten unterschiedlicher Beine. Aus diesem Grund gibt es Verfahren, welche

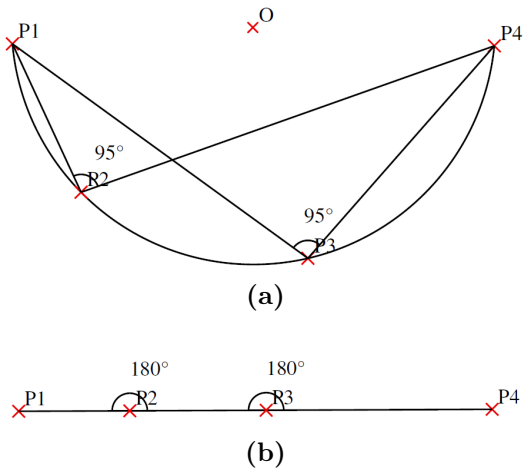


Abbildung 2.21: Winkel innerhalb eines Bogens [XPCR05]

Abbildung 2.22: Unterschiedliches Aussehen von Beinsegmenten [AMB07]

überwachtes Lernen einsetzen um anhand von Trainingsdaten Segmente von Personen zu erlernen und anschließend zu klassifizieren. Diese werden im nächsten Abschnitt vorgestellt.

Detektion mit Hilfe von überwachtem Lernen

Im Kern verwenden die Verfahren dasselbe Grundprinzip für die Klassifikation eines Segments [AMB07, MKH09, SATS10, STS10, Spi08, ZK07]. Die Autoren berechnen auf Basis der Punkte eines Segments für jedes Segment einen Merkmalsvektor. Der Merkmalsvektor besteht aus einer Reihe einzelner Deskriptoren. Hierzu gehören einfache Deskriptoren, wie beispielsweise die Breite des Segments, oder die Anzahl der enthaltenen Punkte. Komplexere Deskriptoren, wie die Standardabweichung oder die Wölbung der Punktereihe im Bezug zum Schwerpunkt werden je nach Autor ebenfalls verwendet.

Eine Übersicht der am häufigsten verwendeten Deskriptoren findet sich in [SATS10] und [AMB07]. Jeder einzelne Deskriptor ist unterschiedlich gut geeignet um Personen zu klassifizieren. So ist der Deskriptor, welcher die Breite eines Segments misst evtl. besser für eine Klassifikation geeignet als ein Deskriptor, welcher die Anzahl Punkte eines Segments enthält, da die Anzahl der Punkte von der Entfernung abhängig ist. Die Information der Punkteanzahl kann aber den-

noch zur Klassifikation genutzt werden. Aus diesem Grund wird zu Klassifikation Adaboost verwendet.

Die allgemeine Formel des Adaboost Klassifikators $H(\mathbf{f})$ ist in Formel 2.10 zu finden.

$$H(\mathbf{f}) = sigmoid\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{f})\right) \tag{2.10}$$

Die Grundidee von Adaboost ist aus mehreren schwachen Klassifikatoren $h_t(\mathbf{f})$ einen starken Klassifikator $H(\mathbf{f})$ zu erstellen, in dem die Gewichtung des Vorfaktors α_t bestimmt wird. Die schwachen Klassifikatoren sind in diesem Fall *Decision Stumps* [AL92]. Diese einstufigen Entscheidungsbäume entscheiden aufgrund eines Merkmals, ob das Segment zu einer Person gehört oder nicht.

Auf die gewichtete Summe der T-vielen schwachen Klassifikationen wird eine Sigmoidfunktion angewendet um einen Wertebereich zu erhalten, indem das Ergebnis als Wahrscheinlichkeitswert interpretiert werden kann.

Der Vorfaktor α_t wird von Adaboost anhand der Trainingsdaten gewählt. Ist ein Deskriptor wie z.B. die Größe des Segments in Metern gut für die Trennung der Klassen geeignet wird ein hoher Vorfaktor vergeben. Sollte ein Deskriptor sich schlecht zur Klassifikation eignen wird ein geringer Vorfaktor vergeben und der Deskriptor hat so weniger Einfluss auf die Klassifikation.

Die Schwellwerte des Decision Stumps werden anhand der Trainingsdaten automatisch bestimmt. Ein Entscheidungsstumpf $h_1(\mathbf{f})$ könnte z.B. den Deskriptor der Größe des Segments nutzen. Eine Person wird klassifiziert, wenn beispielsweise die Größe unterhalb des Segments von 20 cm liegt. Ein anderer Entscheidungsstumpf $h_2(\mathbf{f})$ entscheidet anhand des Wölbungsgrades. Durch die Verwendung von Adaboost werden die Eigenschaften eines Segments einer Person anhand von Trainingsdaten erlernt, ohne ein explizites geometrisches Modell angeben zu müssen.

Aufbauend auf diesen Grundideen gibt es verschiedene Erweiterungen. Eine Erweiterung ist es mehrere 2D-Scanlinien zu verwenden. Mozos et al. [MKH09] verwenden mehrere 2D-Laserscanner in unterschiedlichen Höhen. Die entstehenden 2D-Segmente, zu sehen in Abbildung 2.23, beziehen sich nicht nur auf die Beine, sondern auch z.B. auf den Kopf und Schulterbereich. Hierbei wird der Vorteil der Nutzung eines überwachten Lernens deutlich. Es müssen nicht mehr einzelne feste Werte für z.B. die Segmente auf Kopf- oder Schulterhöhe angegeben werden. Die Eigenschaften der unterschiedlichen Segmente werden mit Hilfe von Adaboost erlernt und anschließend zur Klassifikation eingesetzt. Die Herausforderung ist die Integration der Ergebnisse der einzelnen Klassifikatoren auf den unterschiedlichen Ebenen. Mozos et al. berechnen die euklidische Distanz zwischen gefundenen Personensegmenten und berechnen anschließend die Wahrscheinlichkeit mit der es sich um eine Person handelt.

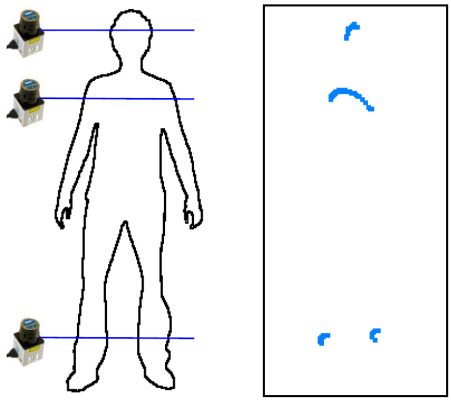


Abbildung 2.23: Verwendung mehrerer 2D-Laser auf unterschiedlichen Arbeitshöhen [MKH09]

Je mehr Lasersegmente verwendet werden um eine Person zu detektieren desto komplexer wird die Integration der Daten. Spinello et al. [SATS10] verwendet einen Velodyne HDL-64E Laserscanner. Dieser Laserscanner tastet seine Umgebung mit 64 Laserstrahlen ab, während er sich fortlaufend im Radius von 360° um seine vertikale Achse dreht. Die entstehenden Scanlinien sind ringförmig um den Sensor angeordnet (Abbildung 2.24). Spinello et al. unterteilen die Person in

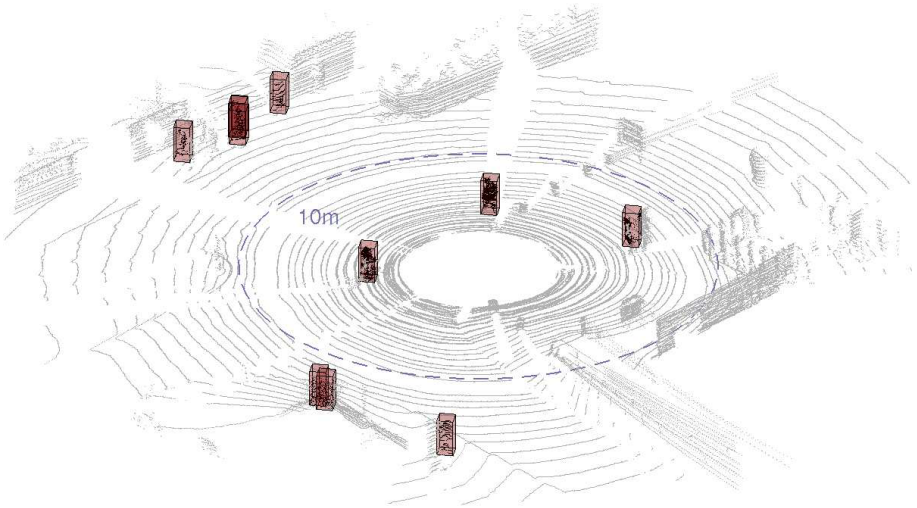
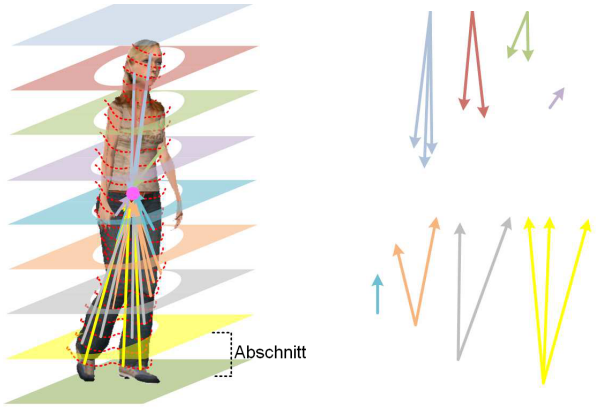


Abbildung 2.24: Scanlinien des Laserscanners. Gefundene Personen sind als Boundingboxes eingezeichnet [SATS10].

K-viele *Abschnitte* (Abbildung 2.25a). In einen Abschnitt können mehrere Scanlinien fallen. Für jeden der Abschnitte wird ein einzelner Klassifikator verwendet.



(a) Abstimmungsmodell zum einlernen des 3D-Personenmodells

(b) Darstellung einzelner Stimmen als Vektor.

Abbildung 2.25: Die einzelnen K-vielen Abschnitte einer Person haben jeweils eine Menge von assoziierten Stimmen [SATS10].

Die einzelnen Segmente eines Abschnittes werden mit Hilfe von Adaboost erlernt. Zu einem gegebenen Segment kann bestimmt werden, wie wahrscheinlich es ist, dass es zu einem bestimmten Abschnitt einer Person gehört. Um die Informationen der einzelnen klassifizierten Abschnitte einer Person zu vereinigen werden mit Hilfe eines *Abstimmungsmodells* (engl. *voting model*), die einzelnen Abschnitte in geometrische Beziehungen gesetzt.

Für diesen Zweck wird anhand der Trainingsdaten der Vektor zwischen dem Schwerpunkt der Person und dem Schwerpunkt eines Segments bestimmt. Dieser Vektor, Stimme[6] (engl. *vote*) genannt, stellt die geometrische Relation zwischen Segmentposition und Person dar. Für jeden Abschnitt einer Person werden die entsprechenden Stimmen bestimmt und mit Hilfe eines agglomerativen Clusterings auf wenige aussagekräftige Stimmen reduziert. Diese Stimmen sind in Abbildung 2.25 für mehrere Abschnitte skizziert. Sind die einzelnen Stimmen eines Abschnittes relativ ähnlich lassen sich durch das Clustering die Anzahl der Stimmen des Segments reduzieren. Variieren die Stimmen stark verbleiben nach dem Clustering mehrere Stimmen pro Abschnitt. Diese Eigenschaft wird genutzt um das Klassifikationsergebnis der einzelnen Abschnitte zu bewerten. Bereiche in denen eine Stimme stark variiert, wie z.B. im Fußbereich, werden entsprechend weniger stark gewichtet.

[6]Der Begriff der Stimme bezieht sich hier auf die Analogie zu Wahlen. Anhand von Anzahl und Gewichtung der Stimmen wird im übertragenen Sinne mit Hilfe des Abstimmungsmodells eine Personenposition „gewählt“.

Wird ein Segment als Abschnitt einer Person detektiert, lässt sich anhand der Stimme auf einen vermuteten Schwerpunkt der Person zurückschließen. Nach der Klassifikation der Segmente zu den einzelnen Abschnitten der Person ergibt sich für jedes Segment ein vermuteter Schwerpunkt der Person. Jedes klassifizierte Segment erstellt eine Hypothese über den Schwerpunkt der Person. Die vermuteten Schwerpunkte in 3D-Koordinaten werden gesammelt und mit Hilfe des *Mean Shift Estimation Algorithmus* [CM02] die Häufungspunkte der Hypothesen bestimmt. Im Anschluss werden die Häufungspunkte mit Hilfe einer Bewertungsformel bewertet. Berücksichtigt werden unter anderem die Anzahl der Abschnitte der Person, die gefunden wurden und welches Segment gefunden wurde.

Der gesamte Ablauf ist in Abbildung 2.26 zusammengefasst.

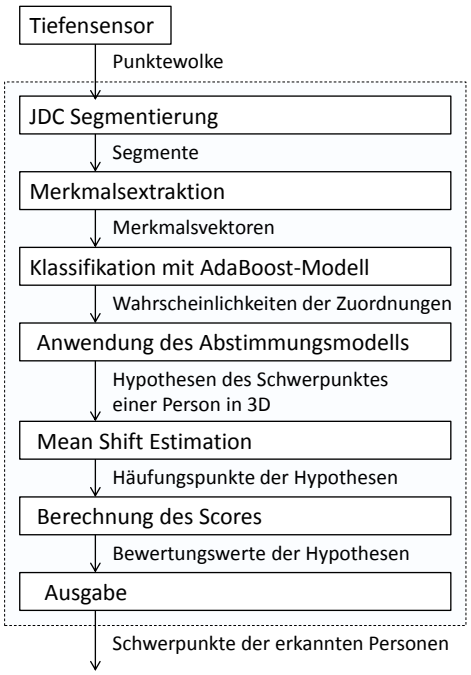


Abbildung 2.26: Ablauf und Informationsfluss der Personendetektion nach [SATS10]

Neben den Verfahren, welche zur Detektion ausschließlich Laserdaten einsetzen, gibt es Verfahren, die Laserdaten mit anderen Sensordaten mischen. Die häufigste Kombination ist die Verwendung von 2D-Bilderkennungsverfahren in Kombination mit Laserdaten [STS10, Spi08, ZK07].

Es ergeben sich hauptsächlich zwei Herausforderungen. Zum einen müssen die unterschiedlichen Sensoren untereinander kalibriert werden. Zu diesem Zweck wird beispielsweise das Verfahren nach Zhang eingesetzt [Zha04]. Zum anderen müssen die Ergebnisse der Sensoren miteinander kombiniert werden. Verwendet werden unterschiedliche probabilistische Methoden. Eingesetzt werden hier Verfahren, welche

Abbildung 2.27: 2D-Laserscanner in Kombination mit einer omnidirektionalen Kamera [ZK07]

dem zuvor vorgestellten Abstimmungsmodell ähneln und Hypothesen unterschiedlicher kalibrierter Sensoren ansammeln. Auch die Art des verwendeten Algorithmus ist unterschiedlich. Es werden Verfahren des *Implicit Shape Modells* auf Basis eines Kamerabildes in Kombination mit 2D-Laserscans eingesetzt [STS10]. Weiterhin finden Verfahren in Kombination mit dem HOG-Detektor [Spi08] Verwendung.

Grundidee ist es die Vorteile der einzelnen Sensorarten zu nutzen. Ein 2D-Laserscanner deckt einen höheren Öffnungswinkel ab, als beispielsweise eine Kamera. Um diese Nachteile auszugleichen setzt Zivkovic et al. eine omnidirektionale Kamera in Kombination mit einem 2D-Laserscanner ein (Abbildung 2.27).

2.2 Handdetektion

Die Detektion von Händen ist die Grundlage für viele Anwendungsmöglichkeiten. Nach einer Detektion kann die Position der Hand ausgewertet werden und als statische Geste aufgefasst werden. Alternativ bauen auf einer Detektion Verfahren zur Verfolgung der Handbewegung auf, um die dynamische Bewegung der Hand festzuhalten. Das Anwendungsgebiet ist dabei vielfältig. Die Handdetektion wird eingesetzt um Computersysteme ohne Maus zu bedienen, entweder weil es aus hygienischen Gründen z.B. im medizinischen Bereich nötig ist kontaktlos zu arbeiten [WKSE11], oder weil eine Steuerung mit den Händen beispielsweise auf einem Tisch eine intuitivere Bedienung ermöglichen soll. Daneben werden Systeme zur Steuerung von Augmented Reality Systemen eingesetzt oder zur Kommunikation mit Robotern. Der Stand der Forschung geht soweit, dass es Ansätze gibt Gebärdensprache anhand der Position und Form der Hand automatisch zu verarbeiten [CXL+12]. Aufgrund der Vielzahl der Anwendungsformen gibt es in der Literatur eine Vielzahl unterschiedlicher Möglichkeiten Hände zu detektieren. Im Kern haben diese Verfahren gemeinsam, dass sie eine oder mehrere der folgenden vier Eigenschaften zur Detektion nutzen:

1. Farbe
2. Form
3. Bewegung
4. Position

Im Folgenden werden Verfahren vorgestellt, die diese vier Eigenschaften auf unterschiedliche Art und Weise nutzen.

2.2.1 Farbe

Die ersten und einfachsten Verfahren zur Detektion der Hand beruhen darauf Hände mit Hilfe eines farbigen Handschuhes, der vom Benutzer getragen werden muss, zu detektieren. Indem sich die auffällige Farbe des Handschuhes vom Hintergrund abhebt kann die Hand vom Hintergrund separiert werden.

Um das Tragen eines Handschuhes zu vermeiden verwenden viele Verfahren die Hautfarbe zur Detektion der Hand. Durch die Nutzung der Hautfarbe ergeben sich Herausforderungen, welche zu lösen sind. Hierzu zählt, dass die Hautfarbe je nach Mensch variiert und zudem je nach Beleuchtungssituation unterschiedlich ausfallen kann.

Zur Lösung dieser Probleme werden vor allem zwei Maßnahmen angewendet: Die Erste Maßname ist die Auswahl geeigneter Hautfarben. Die Zweite Maßname ist die Nutzung eines geeigneten Farbmodells.

Auswahl geeigneter Hautfarben

Bei der Wahl geeigneter Hautfarben gibt es 3 Möglichkeiten. Erstens die Extraktion der Hautfarbe aus Trainingsdaten, zweitens die Extraktion der Hautfarbe aus einer Initialisierungsphase und drittens die Extraktion der Hautfarbe aus dem Gesichtsbereich.

Die erste Möglichkeit besteht darin Hautfarben aus annotierten Bildern zu extrahieren. Jones et al. verwenden hierzu tausende Testbilder aus dem Internet, die manuell annotiert wurden [JR02]. In einem Histogramm werden anschließend die Farben gespeichert, die zur Haut gehören und in einem zweiten Histogramm die Farben, welche keine Haut darstellen. Der Vorteil des Ansatzes ist, dass aufgrund der Größe der Testdaten viele Hautfarben vertreten sind. Der Nachteil ist eine hohe Fehlerrate, da viele Hautfarben sehr ähnlich zu Farben der Umgebung sind.

Um diesen Nachteil auszugleichen extrahieren viele Verfahren die Hautfarbe erst im konkreten Anwendungsfall. Ghosh et al. verwenden eine Initialisierungsphase bei der die Handflächen mit gespreizten Fingern vor die Kamera gehalten werden [GZC+10]. Die Hautfarbe des Benutzers wird bestimmt und steht somit für eine weitere Detektion der Hand zu Verfügung. Die Fehlerrate der Erkennung der Hautfarbe wird gesenkt, da keine verallgemeinerte Hautfarbe verwendet wird und die Hautfarbe in der Belichtungssituation der Initialisierung verwendet wird.

Es ist daher von Vorteil die Hautfarbe erst zum Anwendungszeitpunkt zu bestimmen. Das häufigste verwendete Verfahren zur Bestimmung der Hautfarbe verwendet deshalb die Extraktion der Hautfarbe aus dem Gesicht des Nutzers. Die Autoren [SAAS08, ZAA11, FSV07, IMKK04, SBL11, CXL+12] verwenden bekannte Verfahren zu Gesichtsdetektion. Durch Nutzung des Gesichtes ist keine Initialisierungsphase mehr nötig. Da das Gesicht im Laufe der Beobachtung der Person mehrfach detektiert werden kann, kann die Hautfarbe während der Dauer der Beobachtung des Nutzers bestimmt werden. Ändert sich die Beleuchtungssituation kann anhand des Gesichtes die Hautfarbe aktualisiert werden. Je nach Anwendungsfall ergeben sich unterschiedliche Anpassungen. So verwenden Francke et al. die Hautfarbe zuvor erkannter Gesichter, solange kein neues Gesicht erkannt worden ist [FSV07]. Eine weitere Anpassung ist in Abbildung 2.28 dargestellt. So wird nur der mittlere Bereich der Gesichtserkennung verwendet um die Hautfarben zu extrahieren. Es wird somit verhindert, dass Farben des Hintergrundes versehentlich als Hautfarbe aufgefasst werden.

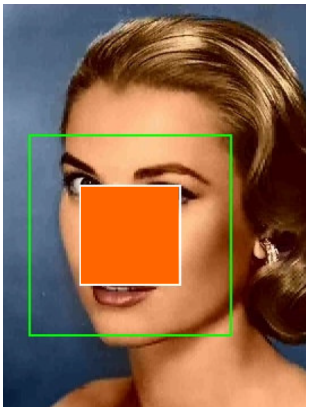


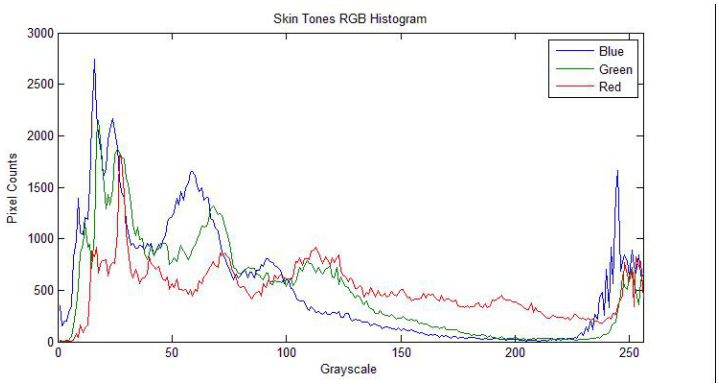
Abbildung 2.28: Ergebnis der Gesichtsdetektion als grünes Rechteck dargestellt. Aus den Farbwerten des Bereich des inneren orangenen Rechteckes wird die Hautfarbe der Person bestimmt. [FSV07]

Auswahl eines Hautfarbenmodelles

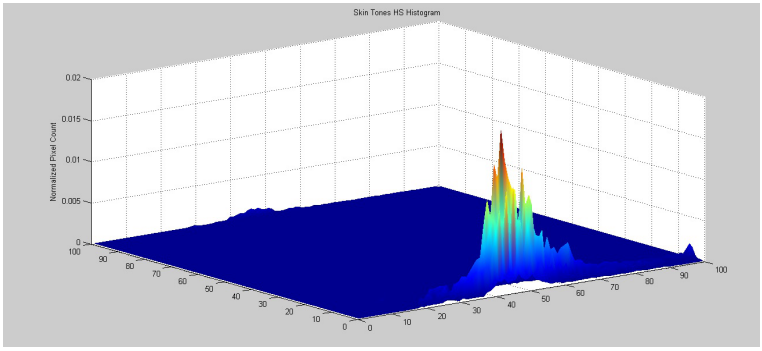
Zur Erstellung eines Farbmodelles der Haut werden verschiedenen Farbräume und Vorgehensweisen verwendet. Zu den häufigsten verwendeten Farbräumen gehören RGB, HSV, YCrCb, L*a*b*.

Ein Hautfarbenmodell ist die Repräsentation der Farben, welche Hautfarben darstellen. Hierzu gibt es drei Verfahren. Das erste Verfahren modelliert die Hautfarbe durch Schwellwerte. Ghosh et al. verwenden beispielsweise Schwellwerte für den Farbwert (H) und die Farbsättigung (S) des HSV Farbraumes [GZC+10]. Cerlinca et al. verwenden ebenfalls Farbwert und Farbsättigung und ignoriert die Helligkeit (V). Befindet sich der Farbwert zwischen einem gesetzten Minimum und Maximum und überschreitet die Farbsättigung ein gegebenes Minimum wird die Farbe als Hautfarbe aufgefasst [CPVC07]. Die Schwellwerte für die Hautfarbe werden aus Beispielregionen für Haut bestimmt. Diese stammen aus einer Initialisierungsphase [GZC+10] oder aus dem detektierten Gesicht der Person [CPVC07].

Die Alternative zu diesem Verfahren ist die Erstellung eines Histogramms der Farbwerte. Farbpixel der hautfarbenen Bereiche z.B. aus dem Gesicht der Person werden in einen geeigneten Farbraum überführt und in die entsprechenden Klassen eines Histogramms eingetragen. Abbildung 2.29a zeigt die Histogramme des Rot-, Grün- und Blaukanals für hautfarbene Pixel. Abbildung 2.29b zeigt die selben hautfarbenen Daten im HSV Farbraum durch Darstellung der Farbwerte (H) und der Farbsättigung (S). Es ist zu erkennen, dass die Repräsentation der Hautfarben im HSV-Farbraum den Teilbereich der Hautfarben im Farbraum gut abgrenzen. Aus diesem Grund verwenden die meisten Verfahren nicht den vom Sensor vorgegebenen RGB-Farbraum sondern alternative Farbräume. Gemeinsam haben die alternativen Farbräume die Eigenschaft die Beleuchtungskomponente einzeln auf-



(a) Histogramm im RGB Farbraum



(b) Histogramm vom Farbwert (H) und der Farbsättigung (S) des HSV Farbraumes

Abbildung 2.29: Histogramme von Hautfarbe in unterschiedlichen Farbräumen [GZC+10].

(a) (b) (c) (d)

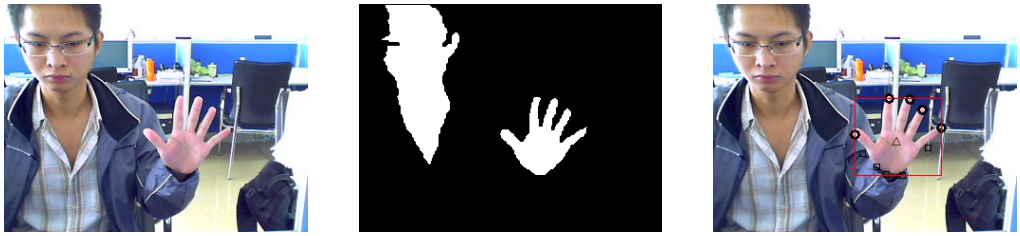
Abbildung 2.30: Handdetektion unter unterschiedlichen Beleuchtungsbedingungen [YFMT08]. Durch Verwendung des L*a*b*-Farbraumes kann die Hautfarbe unabhängig von der Beleuchtung detektiert werden.

zuführen, damit diese verworfen werden kann. Wie bereits erwähnt wird der Hellwert V aus dem HSV-Farbraum verworfen und lediglich der Farbwert (H) und die Farbsättigung (S) verwendet [CPVC07, GZC+10]. Bei Verwendung des YCrCb-Farbraumes wird die Grundhelligkeit Y verworfen und lediglich die zwei Farbkomponenten Cb (Blau Gelb Chrominanz) und Cr (Rot Grün Chrominanz) verwendet [FP09, CXL+12, LC09]. Die Nutzung des L*a*b*-Farbraumes ermöglicht durch das Weglassen der L- Komponente die Entfernung der Helligkeit [YFMT08]. In Abbildung 2.30 ist dies demonstriert, indem eine Hand unter unterschiedlich starker Beleuchtung aufgrund ihrer Farbe detektiert wurde.

Die einzelnen Pixel eines Bildes können unabhängig von der Beleuchtung klassifiziert werden. Durch die Erstellung des Histogramms der Hautfarbe aus dem Gesichtsbereich werden die Klassen (engl. bins) identifiziert, welche Hautfarbe darstellen. Fällt die Farbe eines Pixel in diese Klasse wird das Pixel als Hautfarbe detektiert.

Beispielsweise verwenden Wen et al. diese Vorgehensweise. Anhand eines Farbhistogrammes des CR-Kanals des YCrCb-Farbraums werden die hautfarbenen Pixel bestimmt. Kleine hautfarbene Pixelbereiche werden gelöscht und die Löcher in größeren Bereichen mit Hilfe von morphologischen Operationen (Erosion und Dilatation) gefüllt. Das Ergebnis einer solchen Detektion ist in Abbildung 2.31 zu sehen, in dem sowohl Kopf- und Handbereich als hautfarbene Segmente detektiert wurden. Im Anschluss können dann weitere Untersuchungen der Bereiche durchgeführt werden, um beispielsweise die Finger der Person zu detektierten.

Neben der Verwendung eines Histogramms, welches die Farben in einzelne diskrete Klassen einteilt, gibt es die Möglichkeit ein parameterbasiertes Modell der Hautfarbe zu erstellen. Hierzu gehören die Ansätze von Schiffer et al. [SBL11] und Francke et al. [FSV07]. Schiffer et al. verwenden beispielsweise als Parameter des Modells den Mittelwert und die Standardabweichung der Hautfarben im HSV-Farbraum [SBL11]. Es wird somit eine Gaußkurvenfunktion modelliert, welche für die Hautfarbe steht. Liegt die Abweichung eines zu klassifizierenden Pixels



(a) Original Bild **(b)** Binärbild der Hautfarbenbereiche **(c)** Detektierte Hand und Finger

Abbildung 2.31: Bestimmung der Hand unter Verwendung der Hautfarbendetektion [WZ09].

innerhalb eines festgelegten Schwellenwertes im Bezug zu ermitteltem Mittelwert und Standardabweichung wird die Farbe als Hautfarbe erkannt. Eine Verbesserung dieses Vorgehens ist möglich, indem die Parameter angepasst werden abhängig von der prozentualen Fläche der detektierten Hautfarbe auf dem Bild. Wird in zu vielen Bereichen des Bildes die Hautfarbe detektiert, wird der Schwellwert entsprechend verringert und das Hautfarbenmodell angepasst.

Eine weitere Möglichkeit in der Erstellung eines Modelles der Hautfarbe besteht darin die Farbwerte zu clustern. Yuan et al. verwenden hierzu den L*a*b*-Farbraum [YFMT08]. Der Vorteil des Farbraumes ist, dass die euklidische Distanz zwischen zwei berechneten Farben den visuell wahrgenommenen Farbabständen entspricht. Es ist so möglich den Farbraum in mehrere Cluster für Hautfarben einzuteilen und diese in einer Liste zu speichern. Die einzelnen Bereiche des Farbraumes der ermittelten Cluster repräsentieren somit die zu detektierende Hautfarbe.

2.2.2 Form

Bei der Detektion der Hand über ihre Form ergibt sich das Problem, dass die Hand sehr viele Formen annehmen kann. Dabei ändert sich die Form abhängig von der Stellung der Finger und ebenso abhängig von der Richtung aus der die Hand betrachtet wird. Eine Möglichkeit dieses Problem auszugleichen besteht darin Voraussetzungen zu schaffen, die das Problem vereinfachen.

Utsumi et al. [UO99] verwenden beispielsweise mehrere Kameras, die die Hand von oben und von der Seite betrachten. Anhand der verschiedenen Perspektiven werden verschiedene Konturen der Hand extrahiert und auf die Form der Hand geschlossen (Abbildung 2.32). Weist die Hand eine zuvor definierte Form auf wird sie als Hand erkannt und entsprechend zur Steuerung eines Systems eingesetzt.

Die Form der Hand wird oft zur Bestätigung einer Handdetektion eingesetzt. Ghosh et al. [GZC+10] verwenden daher die Hautfarbe zur Detektion, schränken allerdings gefundene Bereiche anhand der detektierten Form ein. Hierzu wird zu-

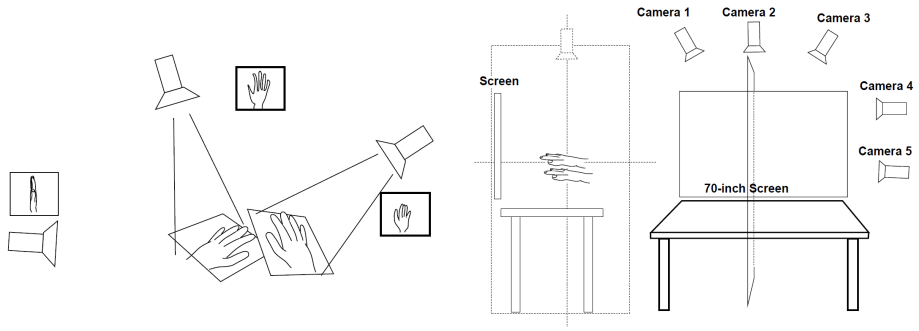


Abbildung 2.32: Sensoraufbau zur Handdetektion aus [UO99]

nächst die Kontur der hautfarbenen Region bestimmt (Abb. 2.33a). Mit Hilfe einer *Douglas-Peucker Annäherung* der unterschiedlichen hautfarbenen Regionen wird durch das Weglassen einzelner Punkte die Kontur stark vereinfacht (Abbildung 2.33b). Anhand dieser Kontur können Annahmen über die Position der Finger erstellt werden. In Abbildung 2.33c sind solche Annahmen über die Finger in Blau dargestellt.

Zu erkennen ist, dass der Holztisch im Hintergrund fälschlicherweise als hautfarbene Region aufgefasst wurde und als mögliche Hand detektiert wurde. Mit Hilfe der Form, genauer der Position der Finger, kann diese Fläche ausgeschlossen werden, da sie nicht der gewünschten Handform der ausgespreizten Finger entspricht. Die Form der Hand wird zu zwei Zwecken verwendet. Zum einen um Fehldetektionen zu vermeiden und zum anderen um Aussagen über die Form der Hand zu ermöglichen. Wird die Hand nach der Initialisierungsphase der ausgespreizten Finger weiter verfolgt und nimmt dabei andere Formen an, kann mit Hilfe des selben Verfahrens die Stellung der Finger und somit eine Bedeutung der Hand ermittelt werden.

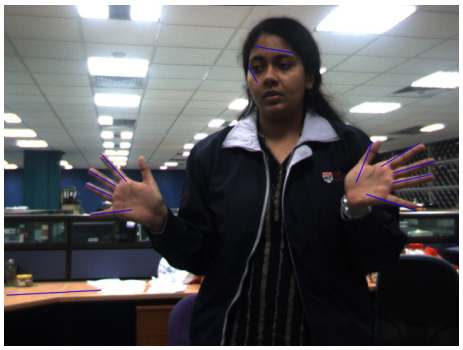
Die Form der Hand ist besonders dann aussagekräftig, wenn bestimmte Handformen vorgeben werden. Choi et al. [CSP09] verwenden die Handdetektion zur Steuerung eines Augmented Reality Systems, das die drei Handformen ausgebreitete Hand, zeigende Hand und greifende Hand erkennt (siehe Abbildung. 2.34).

Hautfarbene Regionen werden extrahiert und mit Hilfe des Kantenbildes voneinander separiert (Abbildung 2.35).

Weist das separierte hautfarbene Segment Ähnlichkeit zu einem der drei möglichen Handformen auf, wird die Hand als solche detektiert. Die Vorausetzung, dass der Nutzer eine fest vorgegebene Handform erzeugt, nutzen auch Ren et al. [RMYZ11]. In einem Anwendungsfall des Spiels “Stein Papier Schere” sind die

(a) Extraktion der Kontur einer Region mit Hautfarbe

(b) Douglas-Peucker Annäherung der Regionen



(c) Ergebnis der Fingerlinien Extraktion

Abbildung 2.33: Extraktion der Kontur und Finger einer Person [GZC+10].

(a) Ausgebreitete

(b) Zeigende

(c) Greifende

Abbildung 2.34: Gültige Handgesten aus [CSP09].

(a) Kantenbild **(b)** Regionen mit Hautfarbe **(c)** Handregion separiert unter Verwendung von (a)

Abbildung 2.35: Extraktion der Hand und dem Unterarm aus [CSP09].

(a) Stein **(b)** Papier **(c)** Schere

Abbildung 2.36: Fest definierte Formen einer Hand aus [RMYZ11].

Formen der Hand auf die genannten drei Formen eingeschränkt (Abb. 2.36). Auf diese Weise wird die Form der Hand zur Bestätigung der Detektion eingesetzt und zugleich die Bedeutung der Hand ausgewertet.

Zur Ermittlung der Form der Hand wird nach folgenden Schritten vorgegangen. Erstens wird die Hand vom Rest der Szene getrennt (Abb. 2.37a). Dies geschieht aufgrund von Farbinformationen [LC09] oder unter Verwendung der Tiefeninformation [RYZ11]. Im zweiten Schritt wird das Segment aufbereitet und gegebenenfalls mit morphologischen Operationen Lücken geschlossen (Abb. 2.37b und 2.37c). Im dritten Schritt wird eine Distanztransformation auf dem Handsegment durchgeführt. Abbildung 2.37d zeigt das Ergebnis dieser Distanztransformation, in der die Entfernung jedes Punktes zum Rand der Kontur bestimmt wird. Lee et al. nutzen diese Daten um die Fingerspitzen zu detektieren [LC09].

2.2.3 Bewegung

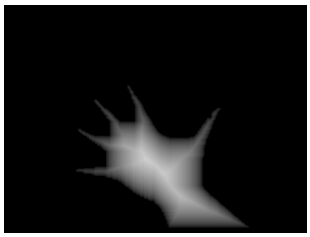
Die Hand einer Person kann als sich bewegendes Objekt aufgefasst werden. Vor allem in Anwendungsfällen in denen der Verlauf der Hand über die Zeit verfolgt wird, um eine Bewegung der Person als Geste aufzufassen, eignet sich die Detektion der Bewegung selbst um die Hand zu detektieren.

Die Szene in der sich die Hand befindet wird dabei über die Zeit, beispielsweise mit einer Videoaufnahme betrachtet. Eine Möglichkeit besteht darin die Differenz zweier aufeinanderfolgender Einzelbilder zu erzeugen. Ghosh et al. [GZC+10] verwenden dies um den unbewegten Hintergrund vom Vordergrund zu trennen.

(a) Eingabebild

(b) Segmentierte Hand unter Verwendung der Farbinformation

(c) Aufbereitetes Handsegment



(d) Distanztransformation des Handsegmentes

Abbildung 2.37: Extraktion der Form der Hand nach [LC09].

Zhang et al. [ZAA11] berechnen einen Bewegungswert für jedes Pixel des Kamerabildes. Die Schreibweise $I(x, y, i)$ steht für den Grauwert des Pixel an der Bildposition (x,y) des i-ten Einzelbildes der Videoaufnahme. Mit Hilfe der in Formel 2.11, 2.12 und 2.13 gezeigten Formeln wird durch den Vergleich der Werte an der Stelle $I(x, y, i - z)$ und $I(x, y, i + z)$ ein Bewegungswert $M(x, y, i)$ berechnet.

$$I_1(x, y, i) = |I(x, y, i) - I(x, y, i - z)| \tag{2.11}$$

$$I_2(x, y, i) = |I(x, y, i) - I(x, y, i + z)| \tag{2.12}$$

$$M(x, y, i) = min(I_1(x, y, i), I_2(x, y, i)) \tag{2.13}$$

Der Bewegungswert $M(x, y, i)$ ist in Abbildung 2.38c dargestellt.

Als weiteres Kriterium für die Detektion der Bewegung verwenden Zhang et al. den Bewegungsrest. Es wird die Eigenschaft genutzt, dass es sich bei der Hand um eine nicht rigide Form handelt, d.h., dass sich die Hand in Größe und Form verändert, während sie sich bewegt. Zur Berechnung des Bewegungsrestes wird das Bild in Blöcke eingeteilt. Durch den Vergleich der Blöcke mit den Blöcken des Einzelbildes des nächsten Einzelbildes kann festgestellt werden, welcher Block sich bewegt hat. Weist ein Block nur geringe Ähnlichkeit zu einem Block des nächsten Einzelbildes auf, ist dies ein Hinweis darauf, dass sich der Inhalt des Blockes verändert hat, da sich z.B. eine Hand bewegt hat. In Abbildung 2.38d ist

(a) Einzelbild

(b) Bewertung mit Hilfe der Hautfarbe

(c) Bewertung mit Hilfe der Bewegung

(d) Bewertung mit Hilfe des Bewegungsrestes

Abbildung 2.38: Handdetektion nach [ZAA11].

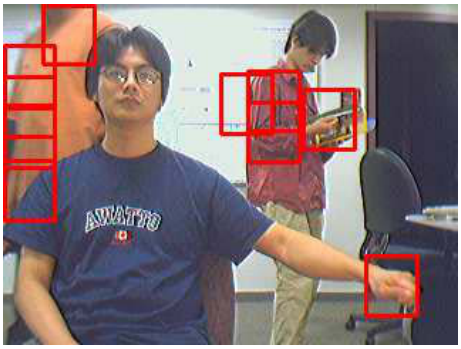


Abbildung 2.39: Darstellung der 10 besten Kandidaten für eine Hand. Ermittelt durch Kombination der Hautfarbe, der Bewegung und des Bewegungsrests nach [ZAA11]

der Bewegungsrest der Pixel visualisiert. Die Abbildung 2.38 zeigt die drei von Zhang et al. genutzten Eigenschaften im Überblick. Zur Detektion der Hand wird die Bewegungsinformation mit der Farbinformation (2.38b) kombiniert eingesetzt.

Der Nachteil bei dieser Vorgehensweise wird in Abbildung 2.39 ersichtlich. Aufgrund der Bewegungsdetektion führen bewegte Objekte im Hintergrund zu Fehlerkennungen.

Zur Behebung des Problems werden weitere Algorithmen angewendet und beispielsweise eine Verfolgung der einzeln bewerteten Regionen durchgeführt. Aus diesem Grund verwenden Wu et al. einen Partikelfilter um sich bewegende Regionen zu verfolgen [WZZ+10]. Zur Detektion und Verfolgung der Hände stellt sich eine Person vor den Sensor und bewegt ihre Hände. Die Objekte, welche sich in Bewegung befinden, werden als mögliche Hände detektiert und weiter verfolgt. Um Fehlerkennungen zu reduzieren werden zusätzlich die Farbe und die Form des verfolgten Bereiches analysiert und die Hand detektiert.

Doliotis et al. verwenden ebenfalls die Farbinformation in Kombination mit der Bewegungsdetektion [DSM+11]. Die Bewegung wird hierbei durch das Bilden der Differenz zweier aufeinanderfolgender Einzelbilder detektiert (Abbildung 2.40c). Durch Multiplikation des Bewegungswertes in Verbindung mit der Bewertung des Bildes anhand der Hautfarbendetektion werden die beiden Eigenschaften Bewegung und Farbe kombiniert (Abbildung 2.40d). Bildbereiche, welche sowohl in Hautfarbe als auch Bewegung auf eine Hand hindeuten, werden somit detektiert (Abbildung 2.40e). Durch dieses Vorgehen führen sich bewegende hautfarbende Bildbereiche im Hintergrund ebenfalls zur Fehldetektion (Abb. 2.40f). Um diesen Nachteil auszugleichen nutzen Doliotis et al. zusätzlich die Tiefeninformation des Sensors. Das erzeugte Tiefenbild (Abbildung 2.41b) wird anhand der Tiefeninformation in einzelne Segmente zerlegt (Abbildung 2.41d). Innerhalb eines Segmentes wird anhand der Bewegungsinformation aus der Differenz der RGB-Bilder, der Hautfarbenbewertung und der Tiefeninformation ein Bewertungswert gebildet

(a) Original Bild

(b) Hautfarbenerkennung

(c) Differenz aufeinanderfolgender Einzelbilder

(d) Multiplizierter Bewertungswert der Hautfarbenerkennung und der Differenz der Einzelbilder

(e) 15 beste Handkandidaten

(f) Bester Kandidat für eine Hand (Hier ist die Handdetektion fehlgeschlagen)

Abbildung 2.40: Verwendung des RGB-Bildes zur Handdetektion [DSM+11].

(Abbildung 2.41e). Auf diese Weise kann durch zusätzliche Verwendung der Tiefendaten die Hand detektiert werden (Abbildung 2.41f). Dies ist ein Beispiel dafür die Bewegungsinformation und gleichzeitig die Information über die Position der Hand zu nutzen. In diesem Falle wird die mögliche Position der Hand über das detektierte Segment innerhalb der Tiefendaten eingeschränkt.

2.2.4 Position

Auch die Position der Hand kann zur Detektion eingesetzt werden. Häufig wird eine Kombination der anderen Merkmale wie Hautfarbe, Form oder Bewegung in Verbindung mit der Position eingesetzt.

Coogan et al. [CAHS06] verwenden die Hautfarbe zur Detektion der Hand. Unter Verwendung von Bewegungsdetektion wird der bewegte Vordergrund vom unbewegten Hintergrund getrennt [CAHS06]. Um den Suchraum der Hand weiter zu reduzieren wird ein Kalman Filter verwendet mit dem die Position der Hand verfolgt und vorhergesagt wird. Durch die Verfolgung der Hand kann dem Problem der Verdeckung entgegengewirkt werden. Bewegt sich beispielsweise die hautfarbene Handregion über die hautfarbene Gesichtsregion, behandelt das System von Coogan et al. diese Situation explizit und vermeidet so, dass versehentlich eine Gesichtsregion als Hand detektiert wird.

Cerlinca et al. schränken die möglichen Positionen der Hand ein. Mit Hilfe der Gesichtsdetektion wird die Position und Größe des Kopfes bestimmt. Basierend auf den Studien von Leonardo da Vinci zum Körperaufbau kann nach einer Detektion des Kopfes auf mögliche Positionen der Hände zurückgeschlossen werden. Zur Detektion selbst wird die Hautfarbe der Hand eingesetzt. Befindet sich die mögliche Hand innerhalb einer gültigen Position wird sie detektiert und je nach Position wird ihr eine Bedeutung zugewiesen (Abbildung 2.42).

Besonders durch die Verwendung von Tiefeninformation ist die Eigenschaft der Position der Hand im dreidimensionalen Raum aussagekräftig. Genutzt wird die Eigenschaft, dass die Hand im Gegensatz zum Hintergrund eine geringe Entfernung zum Sensor aufweist [BB10, FXZ+12, RYZ11, LF04]. Bernard et al. verwenden eine Stereokamera zur Erzeugung der Tiefeninformation. Auf dieser 3D-Tiefenkarte wird das Verfahren von Otsu [Ots79] angewendet, welches die Schwellwerte bestimmt um die Tiefenwerte in mehrere Ebenen zu trennen. In der Ebene, die am nächsten zur Kamera ist, befindet sich die Hand, in der folgenden Ebene der Körper und auf der nächsten Ebene der Hintergrund. Abbildung 2.43 zeigt wie aus der Tiefenkarte des Stereobildes 2.43a die erste Ebene bestimmt wird. Anhand des Binärbildes, welches die erste Ebene repräsentiert (Abbildung 2.43a) wird die Hand detektiert. Das Rauschen der Handebene wird durch mehrfache Anwendung von Erosion und Dilatation reduziert. Der Schwerpunkt dieser binären Maske stellt die Position der Hand dar.

(a) Original Bild

(b) Tiefendaten

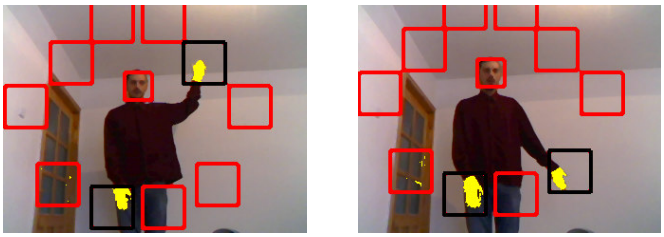
(c) Segmentierung mittels Tiefendaten

(d) Segment einer Person

(e) Darstellung des Bewertungswertes berechnet aus dem Bewertungswert für die Differenz der RGB-Einzelbilder, der Hautfarbenbewertung und der Tiefeninformation

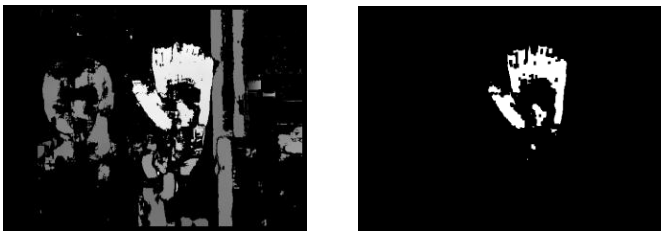
(f) Erfolgreiche Detektion der Hand

Abbildung 2.41: Verwendung der Tiefendaten zur Handdetektion [DSM+11].



(a) **(b)**

Abbildung 2.42: Einteilung möglicher Handpositionen relativ zum Kopf der Person aus [CPVC07]. Je nach Position der Hand ergeben sich unterschiedliche Befehle.



(a) Tiefendaten einer Stereokamera

(b) Binäre Maske der Hand nach Anwendung des Schwellwertes

Abbildung 2.43: Handdetektion mit Hilfe eines Otsu Schwellwertes nach [BB10].

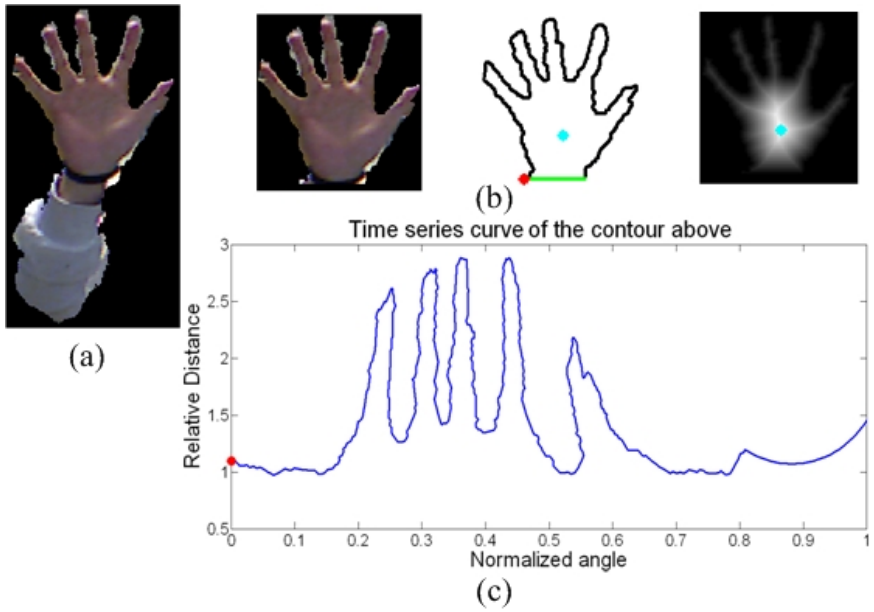


Abbildung 2.44: Detektion der Hand nach [RYZ11]. (a) Segment der Hand extrahiert durch Tiefenschwellwert (b) präziser segmentierte Hand mit schwarzem Band (c) Signatur der Hand

Wenn Voraussetzungen an die Nutzung gestellt werden, eignen sich Tiefendaten auch für die Detektion der Form der Hand. Ren et al. setzen voraus, dass die Hand das nächstgelegene Objekt im Sichtbereich des Sensors ist [RYZ11]. Der Nutzer soll zudem ein schwarzes Band um das Handgelenk tragen. Die Hand stellt den 3D-Punkt dar, der einen Minimalabstand zum Sensor aufweist. In dem Bereich zwischen diesem Minimalabstand und einigen Zentimetern hinter diesem Minimalabstand befindet sich die Hand. Abbildung 2.44 zeigt die Extraktion der Hand anhand ihrer Tiefendaten. Mit Hilfe der Farbinformation des schwarzen Bandes wird die Form der Hand präziser bestimmt. Analog zur Vorgehensweise der Nutzung der Silhouette einer Person von Hordern et al. [HK10] wird die Form der Hand bestimmt. Die Signatur wird durch die Bestimmung der euklidischen Distanz zwischen Schwerpunkt und Kontur ermittelt (Abbildung 2.44).

Neben dieser einfachen Form der Handdetektion als Objekt, welches global am nächsten zum Sensor steht, gibt es die Möglichkeit die Hand als Objekt aufzufassen, was in Relation zur Person am nächsten zum Sensor liegt.

Liu et al. verwenden hierzu eine TOF-Kamera in Verbindung mit der Annahme, dass sich nur eine Person im Kamerabild befindet [LF04]. Die Person wird als einziges Objekt innerhalb eines fest definierten Abstands zur Kamera detektiert. Zusätzlich bestimmen Liu et al. den Kopf der Person. Die Tiefenpunkte der Person, welche sich in einem bestimmten Entfernungsverhältnis im Bezug zum Kopf

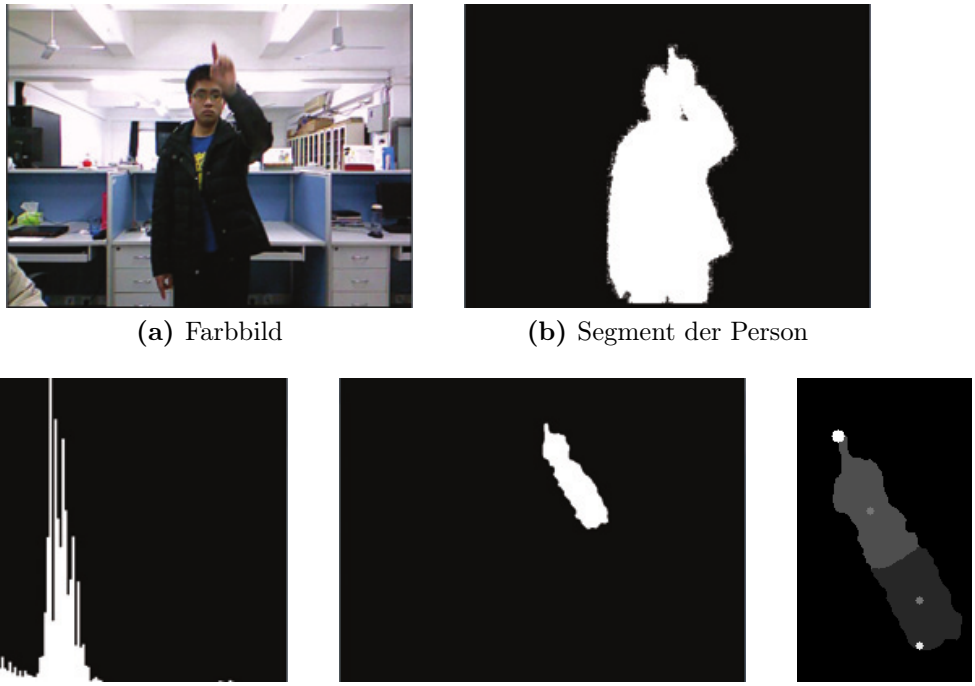
(a) Mensch als Objekt im Vordergrund

(b) Separieren der Hand mit Hilfe der Entfernung zur Kamera

Abbildung 2.45: Separierung der Hand zur Bestimmung der Form [LF04].

befinden, werden als Hand detektiert. Befinden sich die Tiefendaten der Person im Entfernungsbereich in dem sich der Kopf befindet wird keine Hand detektiert. Streckt die Person ihre Hand nach vorne aus befinden sich die Tiefenbereiche in geringerer Entfernung als der Kopf zur Kamera. Die detektierte und segmentierte Hand, zu sehen in Abbildung 2.45, wird anschließend eingesetzt um die Form der Hand genauer zu bestimmen.

Feng et al. [FXZ+12] erweitern dieses Vorgehen um mehrere Hände verschiedener Personen zu detektieren. Es wird eine Personendetektion unter Verwendung der Tiefendaten eines Kinect Sensors eingesetzt um die Segmente der Personen zu ermitteln. Innerhalb eines solchen Segmentes, zu sehen in Abbildung 2.46b, wird ein Histogramm der Entfernungswerte erstellt. Dieses Histogramm, dargestellt in Abbildung 2.46c, wird zur Detektion der Hand eingesetzt. Die stärkste Anhäufung der Histogrammwerte stellt den Körper der Person dar. Der Bereich, der sich vor dem Körper befindet, stellt die Hand dar. Abbildung 2.46d stellt die Tiefenwerte dar, welche sich vor dem Körper befinden. Zur Trennung von Armbereich und Hand verwenden Feng et al. anschließend ein K-Means Clustering. Die Region, welche sich gegenüber dem Armbereich befindet, wird als Fingerspitze interpretiert.



(a) Farbbild **(b)** Segment der Person

(c) Tiefenhistogramm der Person **(d)** Ergebnissegment **(e)** Extraktion der Fingerspitze

Abbildung 2.46: Handdetektion mit Hilfe eines Tiefenhistogramms der Person nach [FXZ+12].

Kapitel 3

Personendetektion

In diesem Kapitel wird das entwickelte System für die Detektion von Personen beschrieben. Im ersten Abschnitt wird hierzu eine Systemübersicht gegeben indem der Ablauf des Systems beschrieben wird. In den folgenden Abschnitten werden einzelne Bestandteile dieses Ablaufs weiter erläutert.

3.1 Systemübersicht

Der Ablauf des Gesamtsystems ist in Abbildung 3.1 skizziert. Die Punktewolke der Kinect entspricht den Eingabedaten. Jeder Punkt innerhalb einer Punktewolke könnte zu einer Person gehören. Aus diesem Grund wird im ersten Schritt mit Hilfe einer Kandidatensuche der Suchraum eingeschränkt. Durch Verwendung eines einfachen Personenmodells werden die Punkte ermittelt, welche den höchsten Punkt einer Person darstellen könnten.

Auf Basis der ermittelten Kandidaten werden Merkmale für jeden Kandidat bestimmt. Zu diesem Zweck wird das Relief- und Breitenmerkmal vorgestellt, welches das Relief bzw. die Breite des sichtbaren Oberkörpers der Person erfasst.

Mit Hilfe eines Annotationswerkzeugs (Anhang A) wurden Trainingsdaten erstellt, indem die Position von Personen annotiert wurde. Auf dem Trainingsdatensatz wird die Kandidatensuche ausgeführt. Anhand der Annotation ist für jeden Kandidaten bekannt, ob es sich um eine Person oder ein Objekt handelt. Für jeden Kandidaten werden die Merkmale erzeugt und zum Training einer SVM verwendet.

Ein zu klassifizierender Kandidat aus der Kandidatensuche kann unter Verwendung der SVM klassifiziert werden. Wurde ein Kandidat als Person klassifiziert kann im Anschluss eine Handdetektion durchgeführt werden.

In dem folgenden Abschnitt 3.2 wird die Kandidatensuche im Detail beschrieben. Abschnitt 3.3 erläutert die Erstellung der neu entwickelten Merkmale. Die

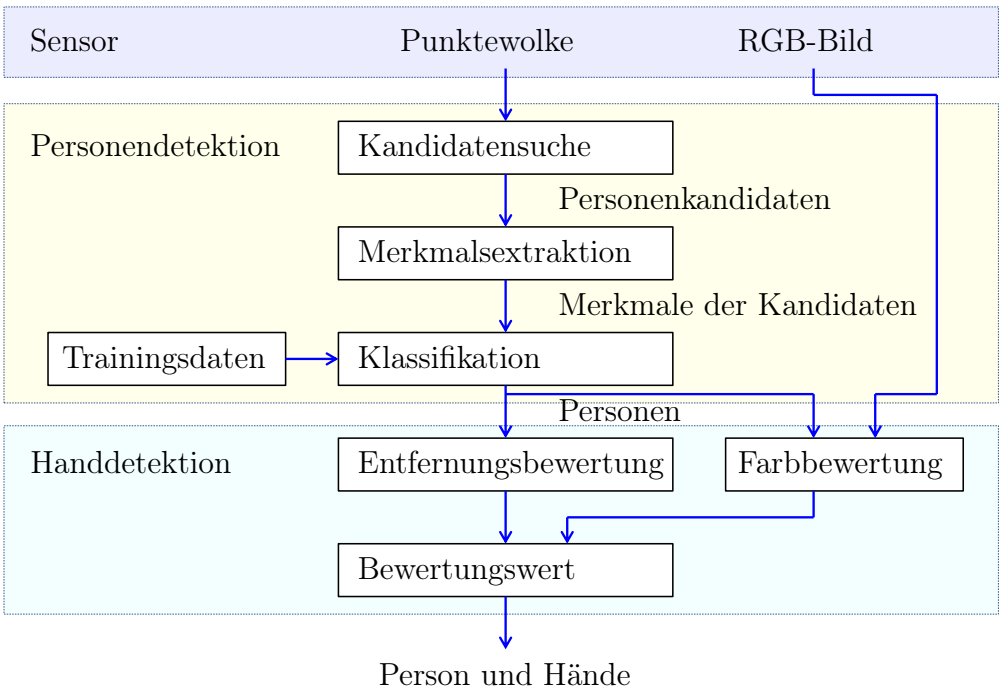
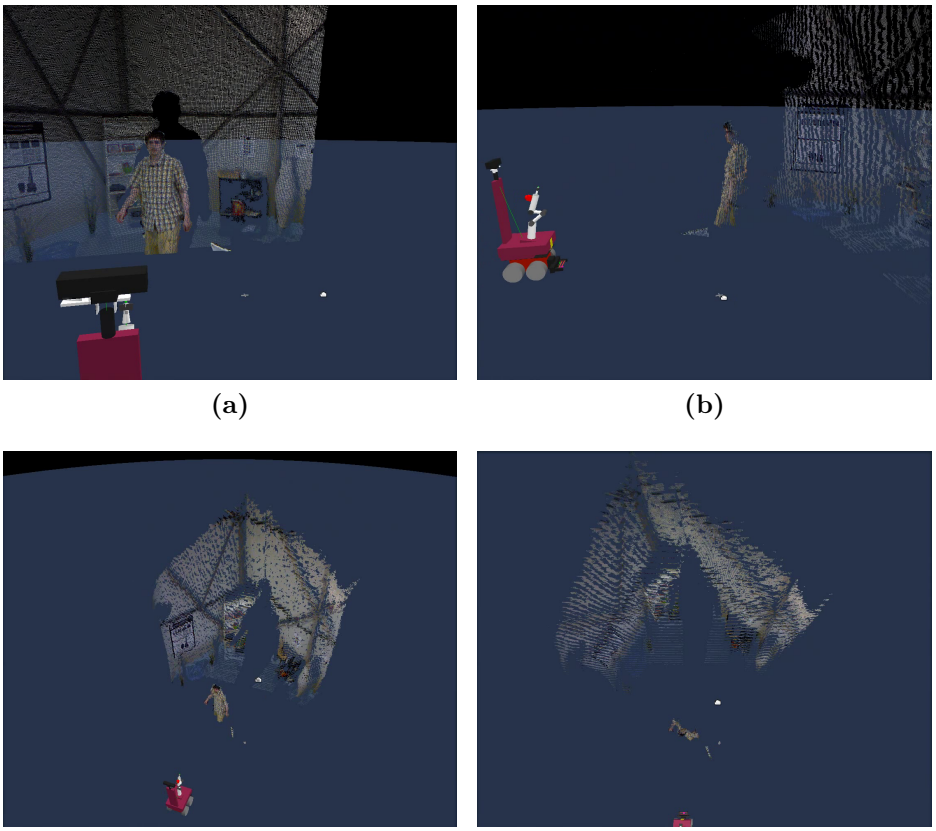


Abbildung 3.1: Systemüberblick

Vorgehensweise zur Klassifikation dieser Merkmale wird in Abschnitt 3.4 beschrieben.



(a) **(b)**

(c) **(d)**

Abbildung 3.2: Versuchsaufbau zur Personendetektion.

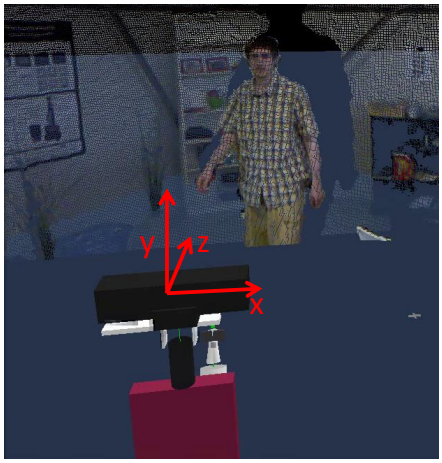
3.2 Kandidatensuche

3.2.1 Versuchsaufbau

Zur Detektion einer Person wird der Kinect Sensor verwendet. Dieser ist beispielsweise auf einem Roboter angebracht und beobachtet eine Szene. Abbildung 3.2 stellt den Aufbau und zugleich die visualisierten Sensordaten dar. Die 3D-Punktewolke der Tiefendaten in Verbindung mit den RGB-Daten der Farbkamera der Kinect bilden eine RGBD-Punktewolke.

3.2.2 Einteilung der Punktewolke

Im ersten Schritt werden zur Personendetektion Bereiche in den Tiefeninformationen ermittelt, welche Personen darstellen können. Hierzu wird die Punktewolke in



(a)

Abbildung 3.3: Aufbau des Koordinatensystems.

einzelne Schichten zerlegt und auf den einzelnen Schichten Kandidaten für mögliche Personen ermittelt.

Ein 3D-Punkt der Punktewolke besteht aus den Koordinaten x,y und z. Der Ursprung der Punktewolke liegt bei der Koordinate $(0, 0, 0)$ im Mittelpunkt des Sensors. Der Aufbau des Koordinatensystems ist in Abbildung 3.3 skizziert. Die $x-$Koordinate stellt die horizontale Entfernung nach rechts bzw. links dar. Die $y-$Koordinate drückt die Höhe des Punktes aus. Je größer die $y-$Koordinate, umso höher ist der Punkt über dem Boden. Die $z-$Koordinate steht für die horizontale Entfernung zum Sensor. Je größer die $z-$Koordinate umso weiter ist der Punkt vom Sensor entfernt.

Um die Punktewolke weiter verarbeiten zu können, werden Schwellwerte der $z-$Koordinate ermittelt, um die Punktewolke in Tiefenbereiche zu separieren. Es wird über alle Punkte der Punktewolke iteriert und anhand der $z-$Koordinate der Punktewolke ein Histogramm gefüllt. Die Vorgehensweise zur Histogrammerstellung ist im Pseudocode (siehe unten, Algorithmus 1) skizziert. Es wird über die Punktewolke iteriert und die $z-$Koordinate des Punktes verarbeitet. Anhand der $z-$Koordinate des Punktes wird die Histogrammklasse ermittelt. Die ersten 10 Meter werden in insgesamt $N = 100$ Histogrammklassen unterteilt. Eine Histogramklasse deckt dabei einen Tiefenbereich von $a = 0.1$ Metern ab. Indem die $z-$Koordinate durch die Größe des Tiefenbereichs a dividiert wird, wird der Index der Histogrammklasse bestimmt. Der Histogrammwert an der Stelle des Index $Histo[i]$ und an den benachbarten Histogrammstellen $Histo[i-1]$ und $Histo[i+1]$ wird erhöht. Auf diese Weise werden Diskretisierungsfehler vermieden und zugleich das Histogramm geglättet.

Algorithmus 1 Erstellung des Tiefenhistogramms

```
for jeden Punkt p in der Punktewolke do
    z ←z-Koordiante aus Punkt p
    i ← z/a
    if i > 1 then
        Histo[i − 1] ← Histo[i − 1] + 1
    end if
    if i + 1 < N then
        Histo[i + 1] ← Histo[i + 1] + 1
    end if
    if i > 0 AND i < N then
        Histo[i] ← Histo[i] + 1
    end if
end for
```

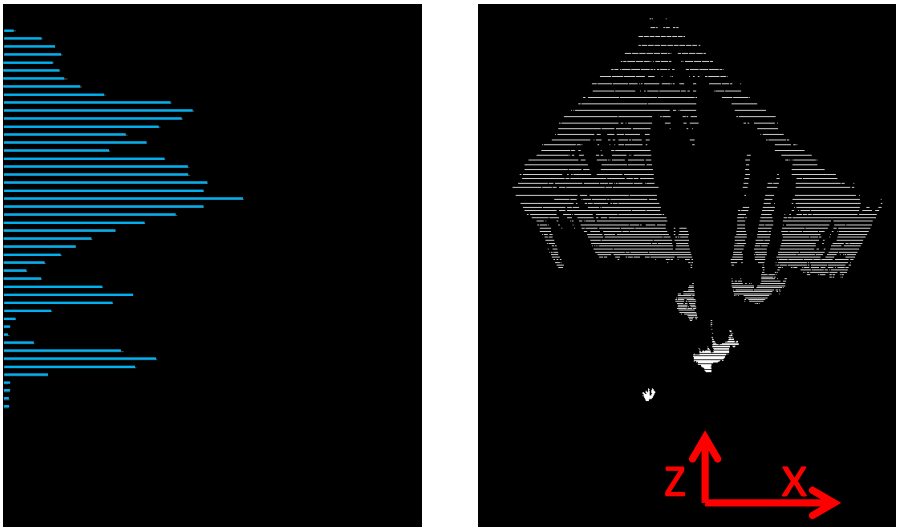
Abbildung 3.4a visualisiert das entstandene Histogramm einer ähnlichen Szene, wie in Abbildung 3.2 zu sehen. Zur Verdeutlichung der Bedeutung des Histogramms ist neben dem Histogramm in Abbildung 3.4b die Draufsicht auf die Szene abgebildet. Die Draufsicht stellt die Projektion der Szene auf die horizontale Ebene der $z-$ und $x-$Achse der Punktewolke dar. Personen und größere Objekte bilden Anhäufungen in der Punktewolke, der Draufsicht und dem Histogramm.

Zur Trennung der Punktewolke in mehrere Tiefenschichten werden daher die lokalen Minima im Histogramm bestimmt. Die Personen, welche oft ein Maximum bilden, befinden sich zwischen den lokalen Minima.

Abbildung 3.5 stellt die Trennung der Tiefenwerte anhand der lokalen Minima dar. In Gelb sind die $z-$Koordinaten der lokalen Minima des Histogramms dargestellt. Diese bilden die Trennungskanten zwischen den einzelnen Abschnitten. Das Tiefenbild der Kinect lässt sich auf diese Weise in mehrere Abschnitte unterteilen. Für jeden Tiefenabschnitt wird ein Binärbild erstellt, indem nur die Punkte weiß eingezeichnet werden, welche sich innerhalb des entsprechenden Tiefenbereiches befinden. Die Perspektive der Binärbilder entspricht dabei der normalen Perspektive mit der die Kinect die Szene beobachtet. Diese Binärbilder sind in Abbildung 3.6 für die ersten Abschnitte dargestellt. Die Binärbilder entsprechen jeweils den in Abbildung 3.6 vermerkten Tiefenbereichen von (a) bis (g).

Auf den Binärbildern 3.6b und 3.6c sind die Silhouetten der beiden Personen, welche sich in der Szene befinden, zu erkennen.

Zur Detektion dieser Personen wird im Folgenden eine einfache modellbasierte Suche verwendet.



(a) 1D-Histogramm **(b)** Ansicht der Punktewolke von oben

Abbildung 3.4: 1D-Histogramm der Tiefenwerte.

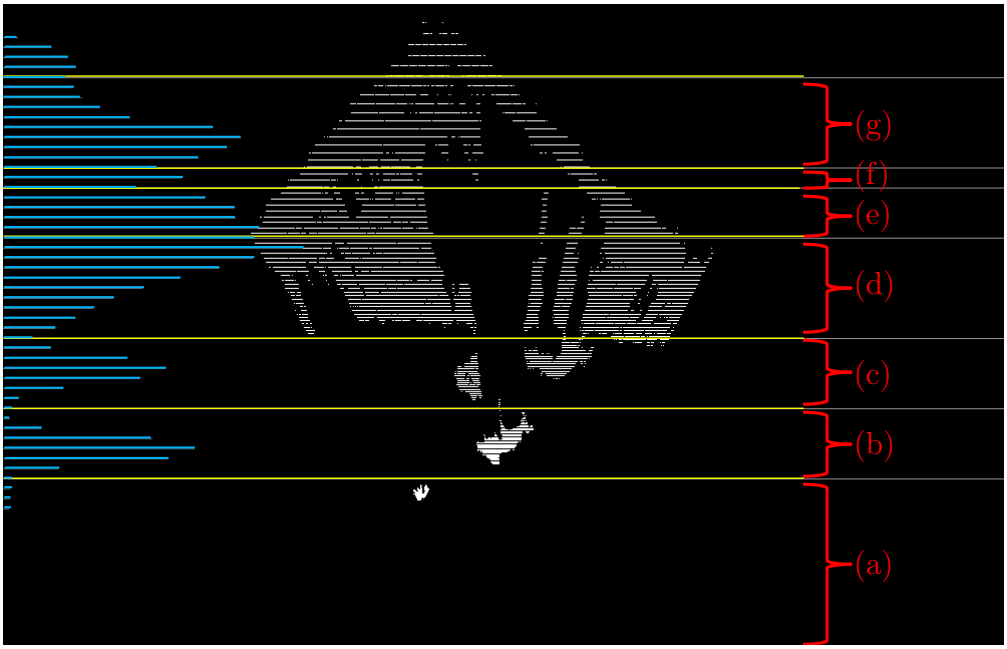


Abbildung 3.5: Einteilung der Punktewolke in Tiefenbereiche mit Hilfe des 1D-Histogramms.

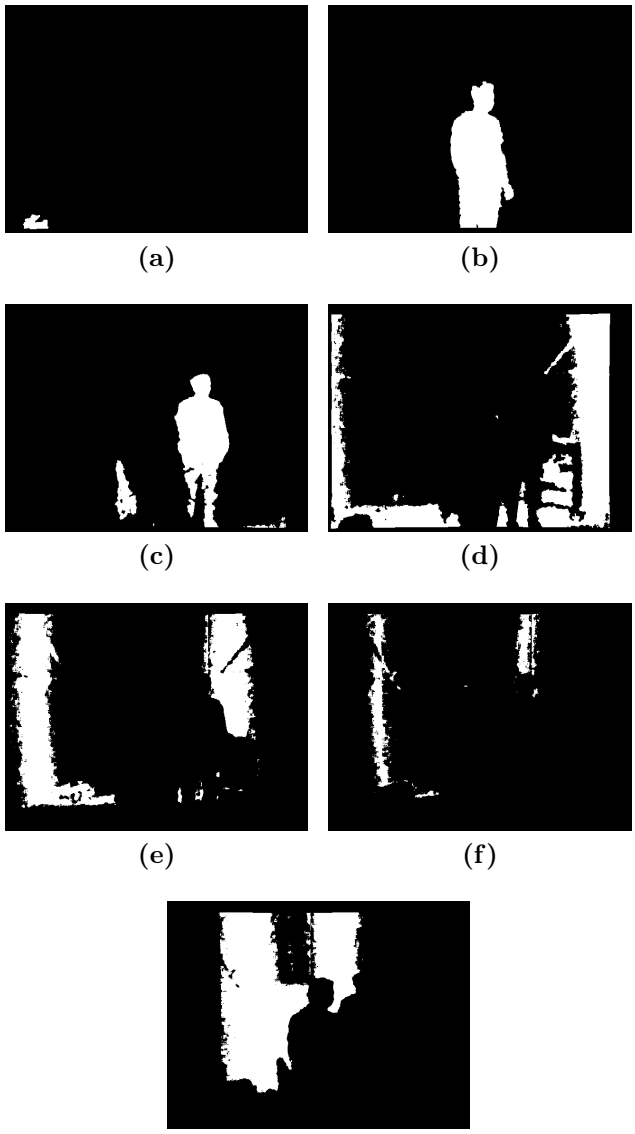


Abbildung 3.6: Abschnitte des Tiefenbildes a bis g aus Abbildung 3.5

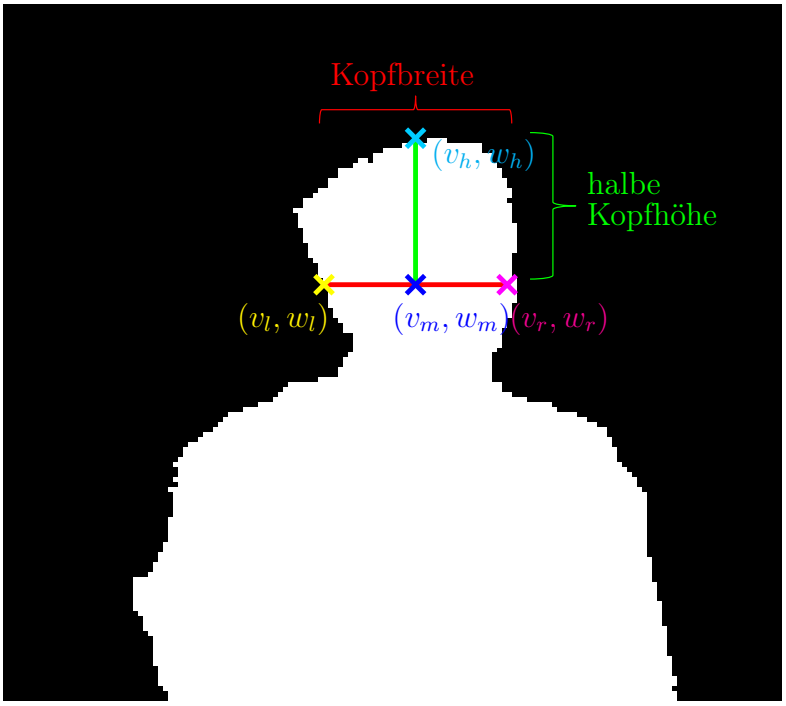


Abbildung 3.7: Einfaches Modell zur Bestimmung geeigneter Personenkandidaten

3.2.3 Modellbasierte Suche

Bei der modellbasierten Suche wird in jeder Tiefenschicht, repräsentiert durch das Binärbild, nach einem Objekt gesucht, welches der Breite eines Kopfes entspricht.

Es wird ausgenutzt, dass zu jedem weißen Pixel auf dem Binärbild ebenfalls die 3D-Koordinate des Punktes bekannt ist. Für eine Pixelkoordinate v und w des 2D-Bildes sind die 3D-Koordinaten der Punktewolke $(x, y, z)^T$ bekannt (siehe Formel 3.1).

$$\mathcal{I}(v, w) = (x, y, z)^T \tag{3.1}$$

v gibt die horizontale Entfernung vom linken Bildrand an (Spalte) und w die vertikale Entfernung in Pixel vom oberen Bildrand (Zeile).

Nach dem in Abbildung 3.7 dargestellten Modell wird in jedem Binärbild gesucht. Gesucht wird der höchste Punkt $\mathcal{I}(v_h, w_h)$ eines möglichen Kopfes. Hierzu wird für jeden weißen Punkt im Binärbild die Hypothese aufgestellt, dass der Punkt der höchste Punkt eines Kopfes sei. Anschließend wird geprüft, ob die Anforderungen an diesen Punkt erfüllt werden. Die Anforderungen sind, dass sich unterhalb des höchsten Punktes $\mathcal{I}(v_h, w_h)$ ein zweiter Punkt $\mathcal{I}(v_m, w_m)$ auf halber Kopfhöhe $\theta_{halbeKopfhoehe}$ befindet. $\mathcal{I}(v_m, w_m)$ ist unterhalb von $\mathcal{I}(v_h, w_h)$ wenn beide Punkte im Binärbild innerhalb der selben Spalte liegen (Formel 3.2).

$$v_m = v_h \tag{3.2}$$

Die Entfernung zwischen beiden Punkten wird über die euklidische Distanz der Punkte innerhalb der 3D-Punktewolke bestimmt (Formel 3.3).

$$\theta_{halbeKopfhoehe} = \|\mathcal{I}(v_m, w_m) - \mathcal{I}(v_h, w_h)\| \tag{3.3}$$

Für den linken Punkt des Kopfes $\mathcal{I}(v_l, w_l)$ und den rechten Punkt des Kopfes $\mathcal{I}(v_r, w_r)$ muss gelten, dass diese innerhalb des Binärbildes in der selben Zeile liegen wie der Punkt auf halber Kopfhöhe $\mathcal{I}(v_m, w_m)$ (Formel 3.4).

$$w_l = w_m = w_r \tag{3.4}$$

Weiterhin muss gelten, dass die Entfernung in 3D zwischen dem linken und rechten Punkt größer als eine vorgegebene minimale Kopfbreite und kleiner als eine vorgegebene maximale Kopfbreite ist (Formel 3.5).

$$\theta_{minKopfbreite} < \|\mathcal{I}(v_r, w_r) - \mathcal{I}(v_l, w_l)\| < \theta_{maxKopfbreite} \tag{3.5}$$

Das Ergebnis der in Formel 3.2 bis 3.5 dargestellten Anforderungen sind Objekte, welche in Kopfhöhe eine Breite aufweisen, die ähnlich der eines Menschen sind. Um weitere Objekte herauszufiltern, bei denen es sich nicht um Personen handelt, wird der Füllungsgrad der in Abbildung 3.8 dargestellten Flächen bestimmt. Insgesamt wird der Füllungsgrad von vier Flächen im Binärbild überprüft. Die erste Fläche, in Abbildung 3.8 blau dargestellt, befindet sich oberhalb des Kopfes. Die Abmessungen der Fläche entsprechen der Kopfbreite und der halben Kopfhöhe in Pixel. Die zweite und dritte Fläche befinden sich rechts und links neben dem Kopf. Ihre Abmessungen entsprechen jeweils der halben Kopfbreite und der halben Kopfhöhe. Die vierte Fläche befindet sich unterhalb des Kopfes. Ihre Breite entspricht der Kopfbreite in Pixel und verläuft bis zum unteren Bildrand des Binärbildes.

Für die vier Bereiche wird jeweils der Füllungsgrad in Prozent innerhalb des Binärbildes errechnet. Liegt der Füllungsgrad der Bereiche innerhalb vorgegebener Schwellwerte ($\theta_{BoxLinks}$, $\theta_{BoxRechts}$, $\theta_{BoxOben}$, $\theta_{BoxUnten}$) wird das Objekt als möglicher Personenkandidat aufgefasst. Die Idee ist, dass die Bereiche oberhalb und neben dem Kopf möglichst nicht gefüllt sein sollten. Der Raum oberhalb und neben einem Kopf sollte frei sein. Der Bereich unterhalb des Kopfes sollte möglichst gefüllt sein, da sich dort der Körper der Person befindet. Der untere Bereich einer Person könnte durch Objekte, welche vor der Person stehen, verdeckt werden. Beispielsweise könnte eine Person hinter einem Möbelstück oder einer anderen Person stehen. Dadurch existieren Tiefenschichten näher an der Kamera. Innerhalb der Tiefenschicht der Person, dargestellt durch das Binärbild, wäre der Füllungsgrad des unteren Abschnittes sehr gering. Aus diesem Grund zählt bei Berechnung des

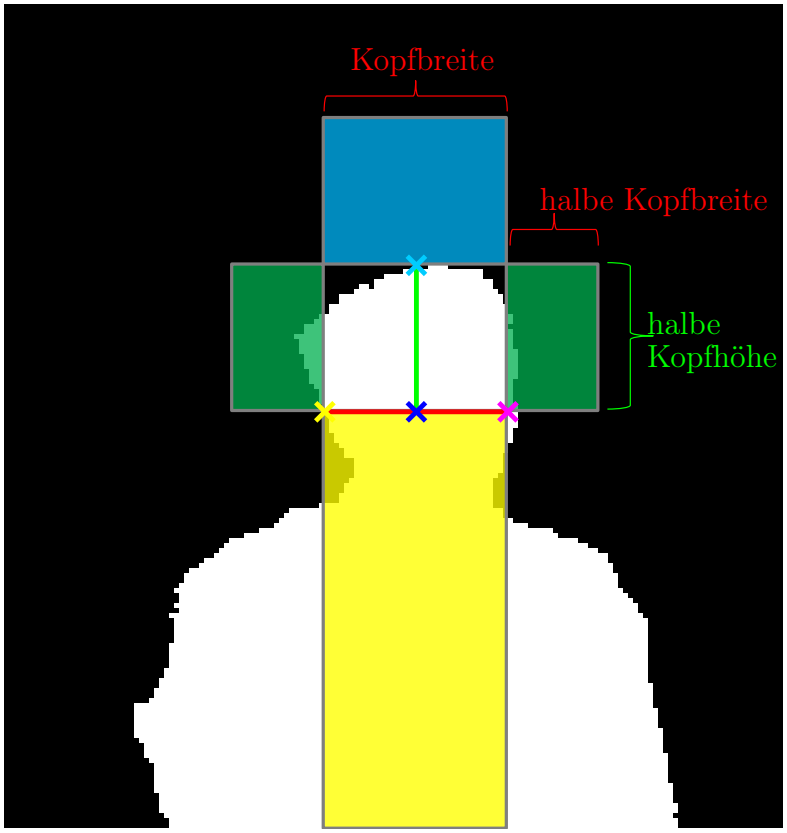


Abbildung 3.8: Flächen zur Bestimmung des Füllungsgrades der Bildbereiche

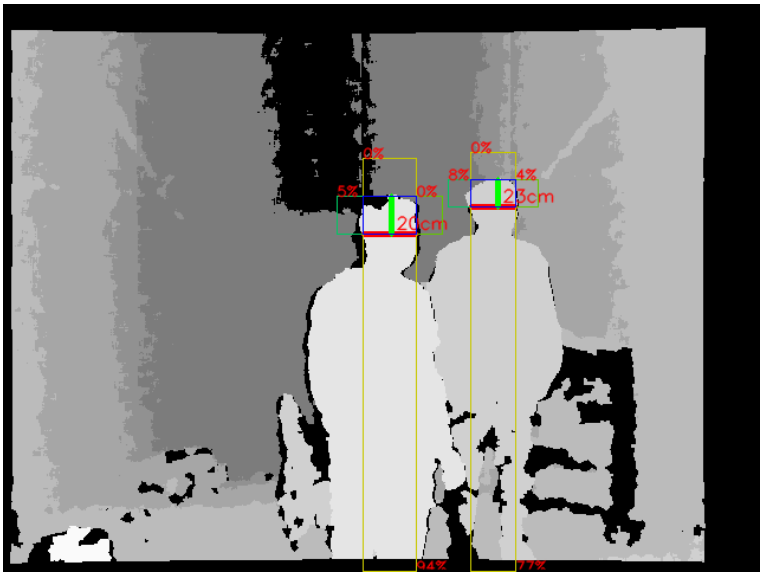


Abbildung 3.9: Ergebnis der Kandidatensuche

Füllungsgrades des unteren Abschnittes auch ein Pixel als gefüllt, wenn es in einem der davorliegenden Tiefenschichten gefüllt ist.

Insgesamt wird durch die Anwendung des einfachen Personenmodells, durch Kopfbreite und Füllungsbereiche, der Suchraum der gesamten Punktewolke auf wenige Kandidaten reduziert. Das Ergebnis dieser Kandidatensuche ist in Abbildung 3.9 zu sehen. Die einzelnen Tiefenschichten sind jeweils durch unterschiedliche Grauwerte dargestellt. Bei den gefundenen Personenkandidaten sind Kopfbreite und der Füllungsgrad der Flächen in Prozent angegeben. Objekte welche die falsche Breite aufweisen, beispielsweise Wände oder große Möbelstücke wie Sessel, werden aussortiert.

Auf Grund der Einfachheit des Modells ist es möglich die Daten auf wenige Personenkandidaten zu reduzieren. Unter den Personenkandidaten gibt es Kandidaten, die das Modell erfüllen, aber keine Personen sind. Abbildung 3.10 zeigt ein solches Beispiel in dem eine Stuhllehne das Modell erfüllt. Die Personenkandidaten werden deshalb im nächsten Schritt genauer klassifiziert.

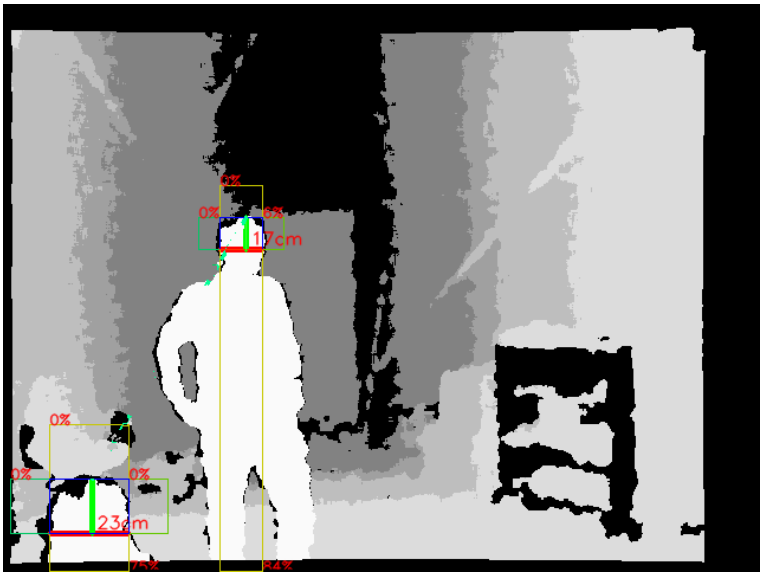
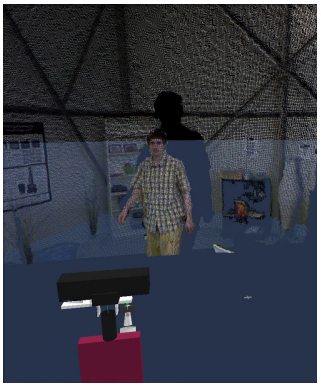
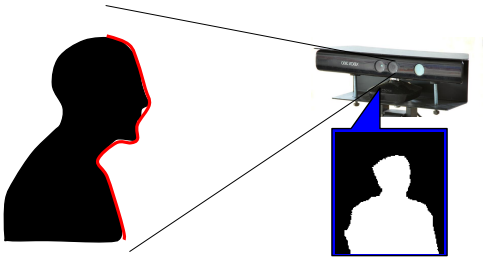


Abbildung 3.10: Beispiel eines Objektes, welches ebenfalls die Anforderungen eines Personenkandidaten erfüllt.



(a) Person blickt frontal auf den Sensor



(b) Seitliche Ansicht auf die Szene. Bildbereich der Person in Blau umrandet, Relief in Rot dargestellt

Abbildung 3.11: Skizze des Relief einer frontal vor dem Sensor stehenden Person.

3.3 Merkmale

Zur genaueren Klassifizierung der Personenkandidaten werden für jeden Kandidat zwei Merkmale bestimmt. Hierzu gehört das Reliefmerkmal und das Breitenmerkmal, die in den folgenden zwei Abschnitten vorgestellt werden.

3.3.1 Reliefmerkmal

Das Wort Relief wird beispielsweise im Bereich der Geographie verwendet. Eine Reliefkarte wird zur Darstellung der Geländeform genutzt. Bereiche der Erdoberfläche, welche Koordinaten in 3D besitzen, werden auf einer 2D-Karte dargestellt. Anhand der Reliefkarte ist es möglich die Höhe über dem Meeresspiegel abzulesen. Auf diese Weise können beispielsweise Berge und Täler identifiziert werden ohne die Punkte direkt im dreidimensionalen Raum darstellen zu müssen.

Die Grundidee des Reliefmerkmales ist es das Relief des Oberkörpers der Person zu nutzen. In Abbildung 3.11a ist eine Beispielszene dargestellt, in der eine Person frontal auf einen Sensor blickt, der im Vordergrund zu sehen ist. Abbildung 3.11b skizziert die selbe Szene mit dem Unterschied, dass die Szene von der Seite betrachtet wird. Auf dem Sensorbild ist eine frontal stehende Person zu erkennen. Diese ist blau umrandet dargestellt. Das zu extrahierende Relief der frontal stehenden Person ist mit einer roten Linie eingezeichnet.

Für das Reliefmerkmal wird der Bildbereich des Kopfes und Teile des Oberkörpers betrachtet. Mit Hilfe der gegebenen Daten aus der Kandidatensuche ist die Kopfposition und die Breite, sowie die Höhe des Kopfes, bekannt. Zur Erstellung des Merkmals wird nur der Bereich, der in Abbildung 3.12a mit einem roten

Rechteck markiert ist, verwendet. Die Breite des Bereiches entspricht der ermittelten Kopfbreite. Die Höhe des Bereiches ist durch einen Schwellwert festgelegt. In der Praxis hat sich ein Schwellwert von 50cm bewährt. Ausgehend vom höchsten Punkt des Kopfes befindet sich innerhalb der ersten ca. 25 bis 30cm der Kopf bis zum Kinn, gefolgt vom Hals und Teilen des oberen Bereiches der Brust.

Abbildung 3.12b stellt das Relief des Bereiches dar. Zur Visualisierung werden die $z-$Koordinaten, welche die Entfernung zwischen Objekt und Sensor darstellen, als Grauwerte abgebildet. Je heller ein Pixel, desto näher befindet sich der Punkt am Sensor. In diesem Beispiel blickt der Sensor frontal auf die Person. Der Bereich des Gesichtes ist näher als der Bereich des Halses.

Die Grundidee ist es, die Information des 2D-Reliefs aus der Abbildung in einem 1D-Relief-Merkmal zu erfassen. Eine einfache Möglichkeit, ein Merkmal aus diesem Relief zu erstellen, ist die Entfernungswerte in der mittleren Pixelspalte des Bereiches von oben nach unten aufzureihen. Das Problem ist, dass sich das Merkmal stark verändert, wenn die Person ihren Kopf um wenige Zentimeter nach links oder rechts bewegt. In einem Fall würde die mittlere Pixelspalte z.B. über die Nase verlaufen in einem anderen Fall nicht. Um das Merkmal gegenüber diesen kleinen Änderungen robust zu machen wird für jede Zeile des 2D-Reliefes der mittlere Entfernungswert bestimmt. Durch die Bildung des Mittelwertes jeder Zeile wird das Merkmal zudem robuster gegenüber dem Sensorrauschen.

Sei f_i eine Funktion, die das i-te Element an der Stelle eines Vektors zurückgibt (Formel 3.7).

$$f_i \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = x_i \tag{3.6}$$

Der Grauwert eines Pixels kann im Reliefbild über die $z-$Koordinate der Punktewolke ermittelt werden:

$$z = f_3(\mathcal{I}(v, w)) \tag{3.7}$$

Das 2D-Reliefbild enthält nur den Teilbereich des Kopfausschnittes. Innerhalb des Reliefbildes werden nur die Pixel, die zum Objekt gehören einem Grauwert zugeordnet. Pixel, die aus einer anderen Tiefenschicht als das gerade betrachtete Objekt stammen, werden schwarz dargestellt. Ihr Entfernungswert wird auf 0 festgelegt.

Für jede Zeile w des Reliefbildes wird der mittlere Entfernungswert der Punktewolke bestimmt. Hierzu werden die Entfernungswerte einer Zeile aufsummiert und durch die Anzahl der Pixel N_w der Zeile w, welche zum Objekt gehören, ge-

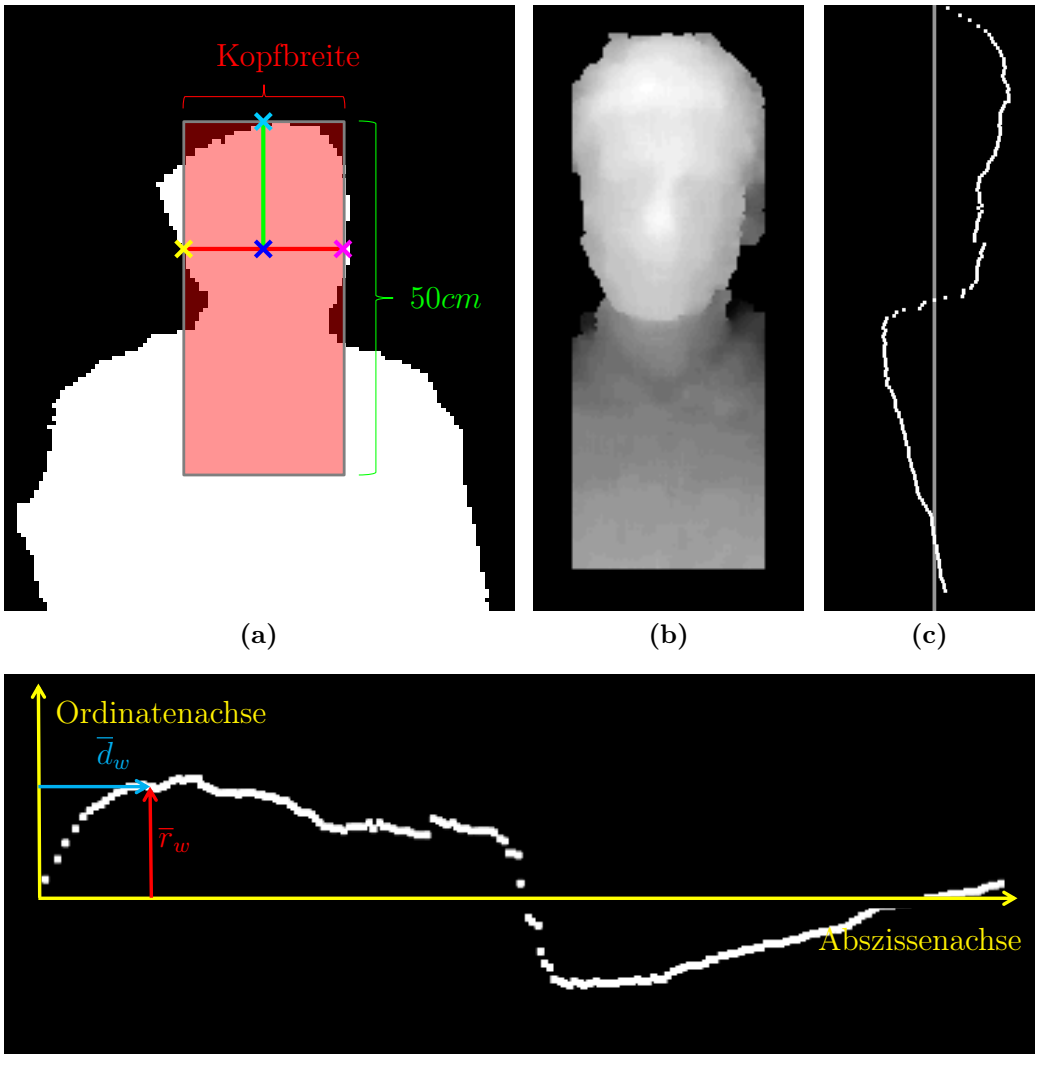


Abbildung 3.12: Extraktion des Reliefmerkmals: (a) verwendete Region des Kopfes und Brustbereiches, (b) Darstellung der Entfernungswerte des Kopfbereiches, (c) Extrahiertes Merkmal, (d) Extrahiertes Merkmal zur Darstellung als Funktion um 90 Grad gedreht

teilt (siehe Formel 3.8). N_w gibt die Anzahl der Pixel der Zeile w an für die gilt $f_3(\mathcal{I}(v,w)) > 0$. V gibt die Gesamtanzahl der Pixel der Spalten an.

$$\overline{r}_w = \frac{\sum_{v=0}^{V} f_3(v,w)}{N_w} \tag{3.8}$$

Zu jedem mittleren Entfernungswert $\overline{r}_w$ wird die Entfernung $\overline{d}_w$ zum höchsten Punkt des Kopfes innerhalb der 3D-Punktewolke bestimmt. Hierzu wird der Mittelwert der $x-$Koordinate jeder Zeile gebildet.

Um die Entfernung innerhalb der Punktewolke in vertikaler Richtung zu bilden wird die $x-$Koordinate der 3D-Punktewolke des obersten Kopfpunktes $\mathcal{I}(v_h, w_h)$ subtrahiert:

$$\overline{d}_w = \frac{\sum_{v=0}^{V} f_1(v,w)}{N_w} - f_1(\mathcal{I}(v_h, w_h)) \tag{3.9}$$

Als Ergebnis dieser Berechnung liegt für jede Zeile w des Reliefbildes ein mittlerer Reliefwert $\overline{r}_w$ und der Abstand zum höchsten Kopfpunkt dieser Zeile $\overline{d}_w$ vor. $\overline{r}_w$ und $\overline{d}_w$ sind jeweils in Metern angegeben, da zur Berechnung der Werte nur die Koordinaten der 3D-Punktewolke verwendet wurden. Die 2D-Koordinaten des Reliefbildes wurden zur Auswahl der 3D-Punkte verwendet.

Die Werte $\overline{r}_w$ und $\overline{d}_w$ lassen sich in einem Koordinatensystem darstellen. Zur Verdeutlichung der Bedeutung dieser Werte zeigt Abbildung 3.12c die Punkte passend zum Reliefbild 3.12b gedreht. Im Bereich des Reliefs des Gesichtes weisen die Zeilen eine niedrigere durchschnittliche Entfernung zur Kamera auf als beispielsweise am Hals. Um die Werte als Funktion darstellen zu können wird $\overline{r}_w$ auf der Abszissenachse und $\overline{d}_w$ auf der Ordinatenachse eines Koordinatensystems aufgetragen. Diese Darstellung in Abbildung 3.12d entspricht der um 90° gedrehten Abbildung 3.12c. Für eine bessere Darstellung der Werte wurden die Werte normalisiert, so dass der Mittelwert aller $\overline{r}_w$ der Abszissenachse entspricht.

Je nachdem in welcher Ausrichtung eine Person vor dem Sensor steht, ergeben sich unterschiedliche Reliefverläufe. Steht eine Person frontal zur Kamera ergibt sich ein anderes Relief, als wenn sie seitlich oder abgewandt steht (siehe Abbildung 3.13).

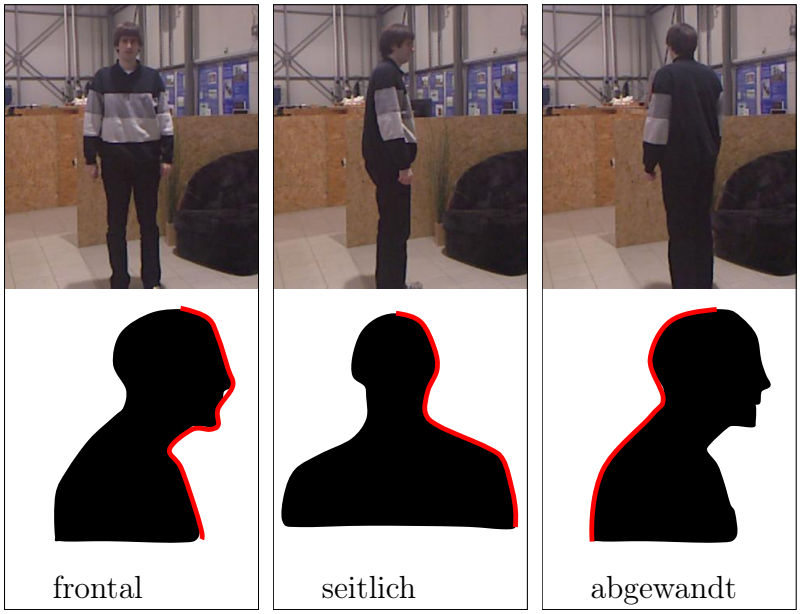
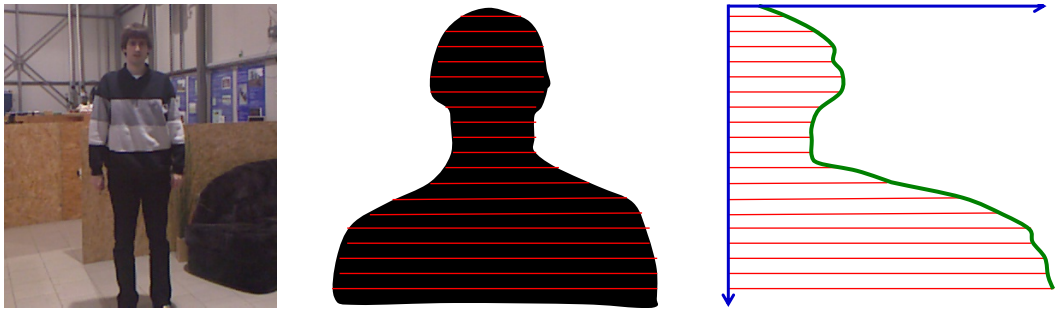


Abbildung 3.13: Je nach Ausrichtung der Person ergibt sich eine andere Reliefart: (a) Frontal (b) Seitlich (c) Abgewandt



(a) Situation **(b)** Skizze der Breiten der Person **(c)** Breitenmerkmal

Abbildung 3.14: Breitenmerkmal einer frontal zum Sensor stehenden Person.

3.3.2 Breitenmerkmal

Das Breitenmerkmal nutzt die Breite eine Person. Steht eine Person frontal zum Sensor ist die Person auf Schulterhöhe breiter als auf Kopfhöhe (siehe Abbildung 3.14a und 3.14b).

Zur Extraktion der Breite einer Person wird wie beim Reliefmerkmal der Bereich der obersten 50cm der Person verwendet. Im Gegensatz zum Reliefmerkmal wird nun nicht das Relief der Person d.h. nicht die $z-$Koordinate der Punktewolke verwendet, sondern die $x-$Koordinate der Punktewolke. Aus der Kandidatensuche wird das Binärbild der möglichen Person betrachtet. Für jede Zeile i des Binärbildes wird jeweils der äußerste linke (v_l^i, w_l^i) und äußerste rechte (v_r^i, w_r^i) Punkt bestimmt, der zum Objekt gehört.

Über die $x-$Koordinate der 3D-Punktewolke beider Punkte wird die horizontale Entfernung zwischen beiden Punkten bestimmt (siehe Formel 3.10). b^i steht somit für die Breite des Objektes in der Zeile i.

$$b^i = |f_1(\mathcal{I}(v_l^i, w_l^i)) - f_1(\mathcal{I}(v_r^i, w_r^i))| \tag{3.10}$$

Mit Hilfe der $y-$Koordinate wird die vertikale Entfernung d^i zwischen dem höchsten Punkt des Kopfes (v_h, w_h) und der jeweiligen Zeile bestimmt (siehe Formel 3.11).

$$d^i = |f_2(\mathcal{I}(v_l^i, w_l^i)) - f_2(\mathcal{I}(v_h, w_h))| \tag{3.11}$$

Abbildung 3.15 skizziert die jeweiligen Punkte für eine Zeile i. Sowohl der Wert d^i als auch der Wert b_i bezeichnen Distanzen in Meter, da sie mit den Koordinaten aus der 3D-Punktewolke berechnet wurden. Die Werte d^i und b_i lassen sich auf ein Koordinatensystem auftragen (siehe Abbildung 3.17). Abbildung 3.16 zeigt das um 90° gedrehte Merkmal passend zum daneben zu sehenden Binärbild der Person. Die Abschnitte der Person, wie Gesichts-, Hals- oder Schulterbereich lassen sich im Breitenmerkmal wiederfinden.

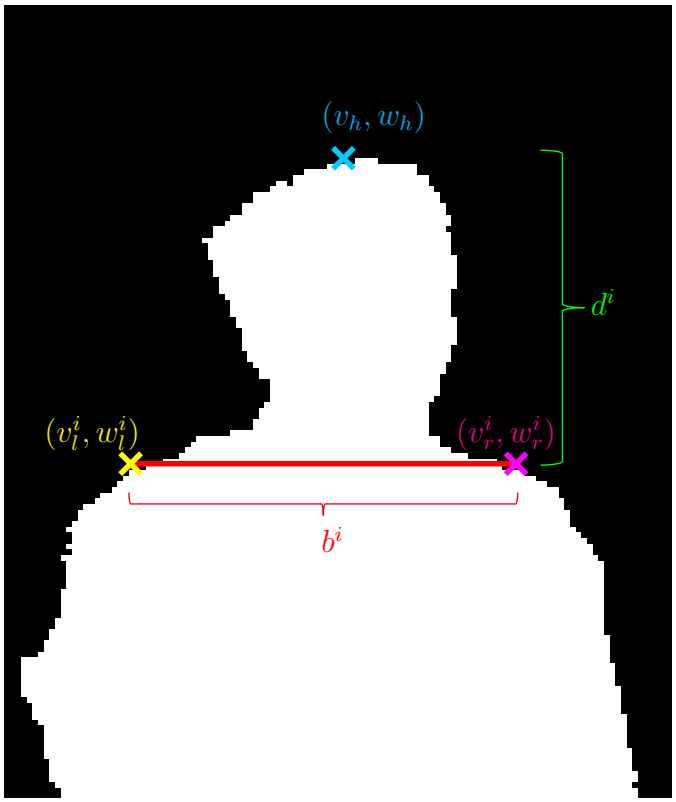


Abbildung 3.15: Skizze der Punkte für eine Zeile i

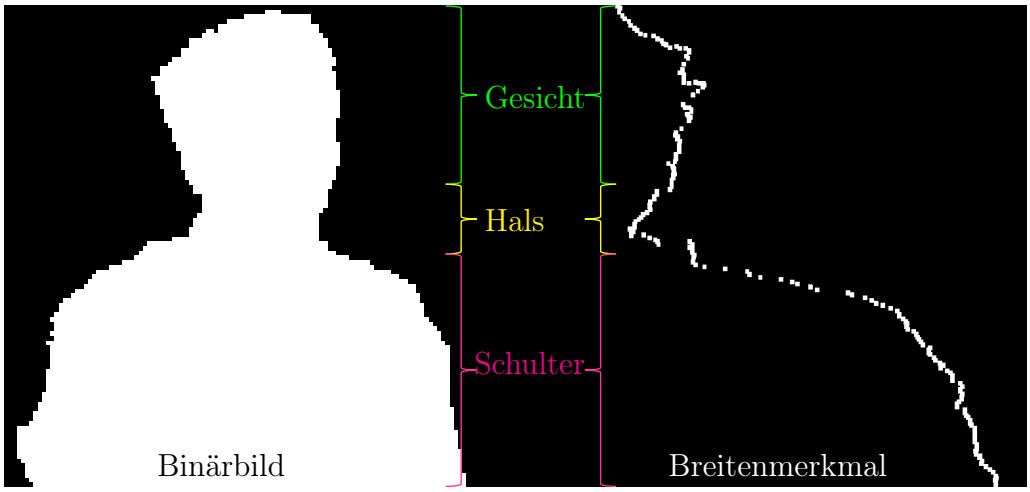


Abbildung 3.16: Breitenmerkmal einer Person

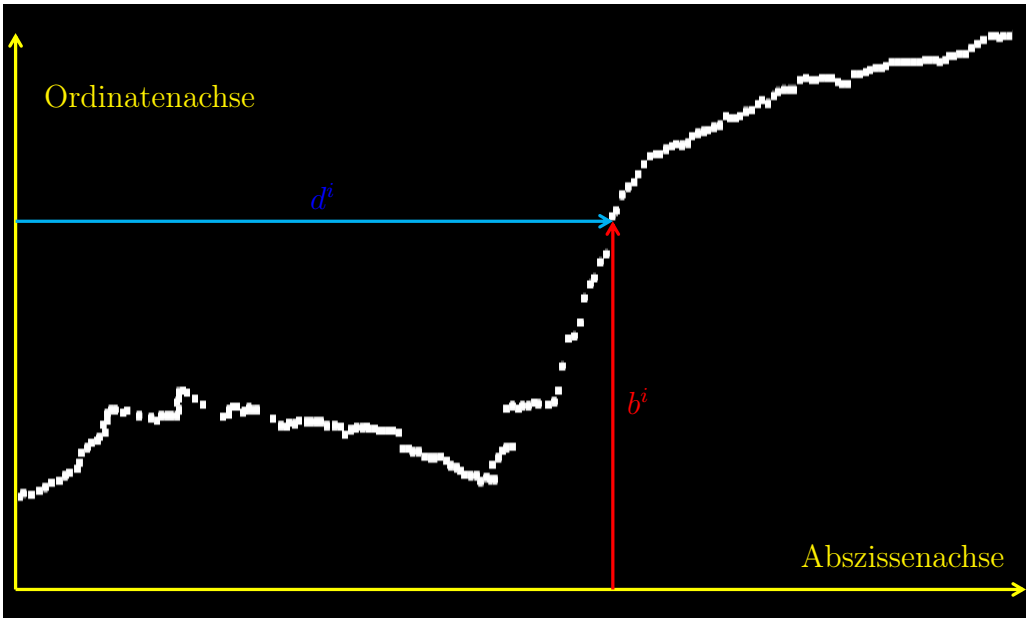


Abbildung 3.17: Breitenmerkmal um 90° gedreht.

Je nach Ausrichtung der Person verändert sich, wie beim Reliefmerkmal auch, das Breitenmerkmal. Im Gegensatz zum Reliefmerkmal ist das Breitenmerkmal einer abgewandten Person sehr ähnlich zu dem einer zugewandten Person (siehe Abbildung 3.18).

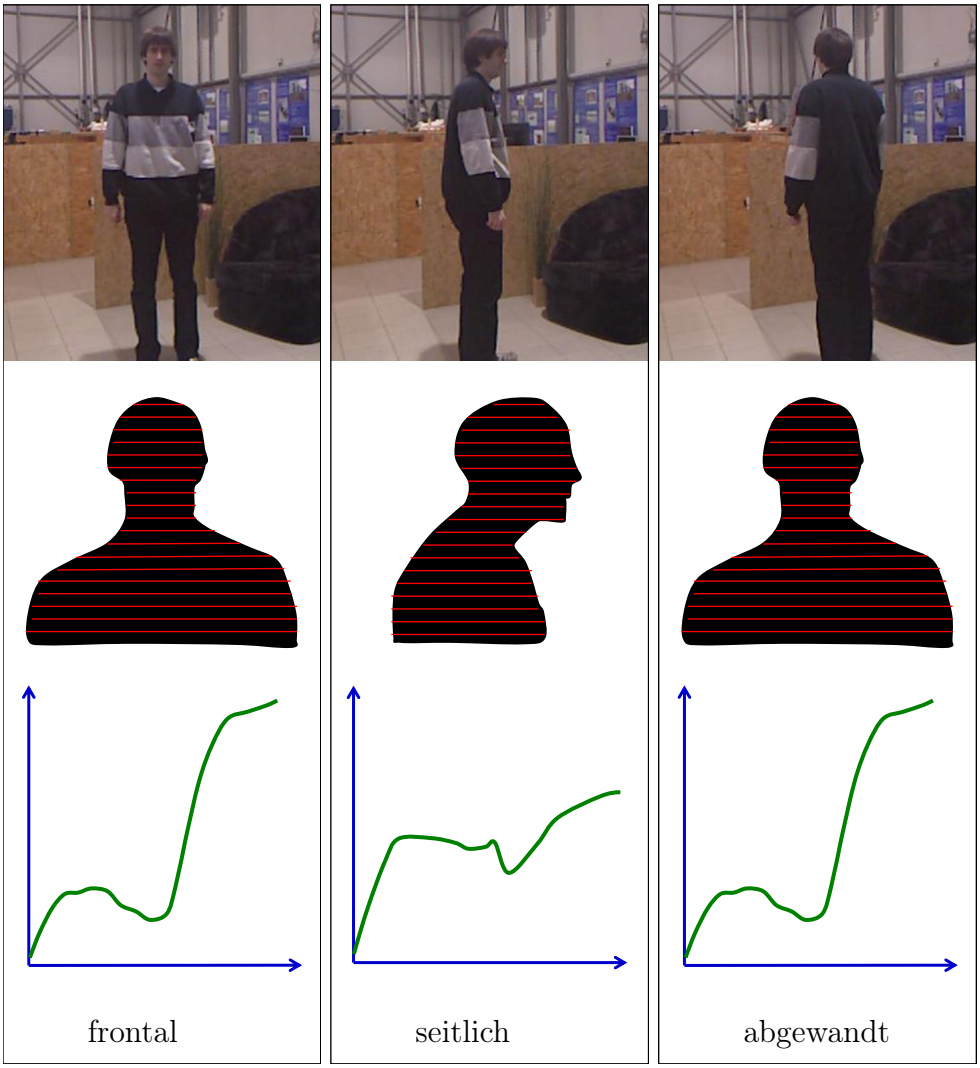


Abbildung 3.18: Das Breitenmerkmal bei unterschiedlicher Ausrichtung.

3.4 Klassifikation

Mit Hilfe der Kandidatensuche aus Kapitel 3.2 wurden Kandidaten ermittelt, welche mögliche Personen darstellen. In Kapitel 3.3 wurden anhand dieser Kandidaten das Relief- und Breitenmerkmal erstellt. In diesem Kapitel wird aufgezeigt, wie diese Merkmale für eine Klassifikation vorverarbeitet und anschließend verwendet werden.

3.4.1 Vorbereitung der Merkmale

Anhand der Kandidaten wird das Relief- und Breitenmerkmal erzeugt. Die Merkmale lassen sich als Funktion mit diskreten Werten darstellen. Diese Merkmale werden mit Hilfe einer Fourier-Transformation aufbereitet um sie anschließend klassifizieren zu können.

Die Vorgehensweise, die Fourier-Transformation zu nutzen, ist beispielsweise aus dem Vergleich von 2D-Formen bekannt. Hordern et al. [HK10] transformiert die Signatur einer 2D-Form um einen Merkmalsvektor zu erzeugen (Abschnitt 2.1.2 Formel 2.8).

Zur Verdeutlichung des Nutzen der Fourier-Transformation wird im Folgenden das Reliefmerkmal einer frontal zum Sensor stehenden Person transformiert. Das Reliefmerkmal einer frontal zum Sensor stehenden Person ist in Abbildung 3.19 zu sehen. Die um 90° gedrehte Ansicht in Abbildung 3.20a entspricht der Darstellung der Werte als Funktion. Auf der Abszissenachse ist der Abstand $\overline{d}_w$ des Punktes vom obersten Kopfpunkt aufgetragen. Insgesamt werden die ersten 500mm (50cm) betrachtet. Die Ordinatenachse enthält die Reliefwerte $\overline{r}_w$, die ebenfalls in Millimetern angegeben sind.

Die Fourier-Transformation bildet ein Signal in den Frequenzraum ab. Da es sich bei den Werten des Reliefmerkmals um diskrete Werte handelt, wird zur Anwendung der Fourier-Transformation die Diskrete Fourier-Transformation (DFT) eingesetzt (Formel 3.12) [TS09, HK10] .

$$s_k = \frac{1}{N} \sum_{t=0}^{N-1} r(t) e^{-j2\pi nk/N}, k = 0, 1, ..., N-1 \tag{3.12}$$

Als Eingabewerte wird eine Reihe von komplexen Zahlen $r_0, ..., r_{N-1}$ benötigt. Bei diesen Werten handelt es sich um die reellen Werte des Merkmals aus Abbildung 3.20a. Der Imaginärteil der komplexen Zahl wird für alle Eingabewerte auf 0 gesetzt. Die Ausgabewerte der Fourier-Transformation $s_0, ..., s_{N-1}$ sind ebenfalls komplexe Zahlen.

Jede komplexe Zahl, bestehend aus einem Realteil $Re(s_n)$ und einem Imaginärteil $Im(s_n)$, repräsentiert einen Sinusoid. Ein Sinusoid stellt eine Sinusfunkti-

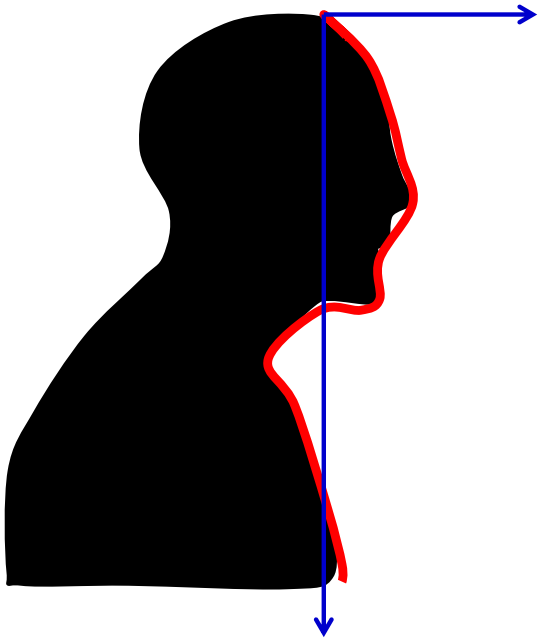
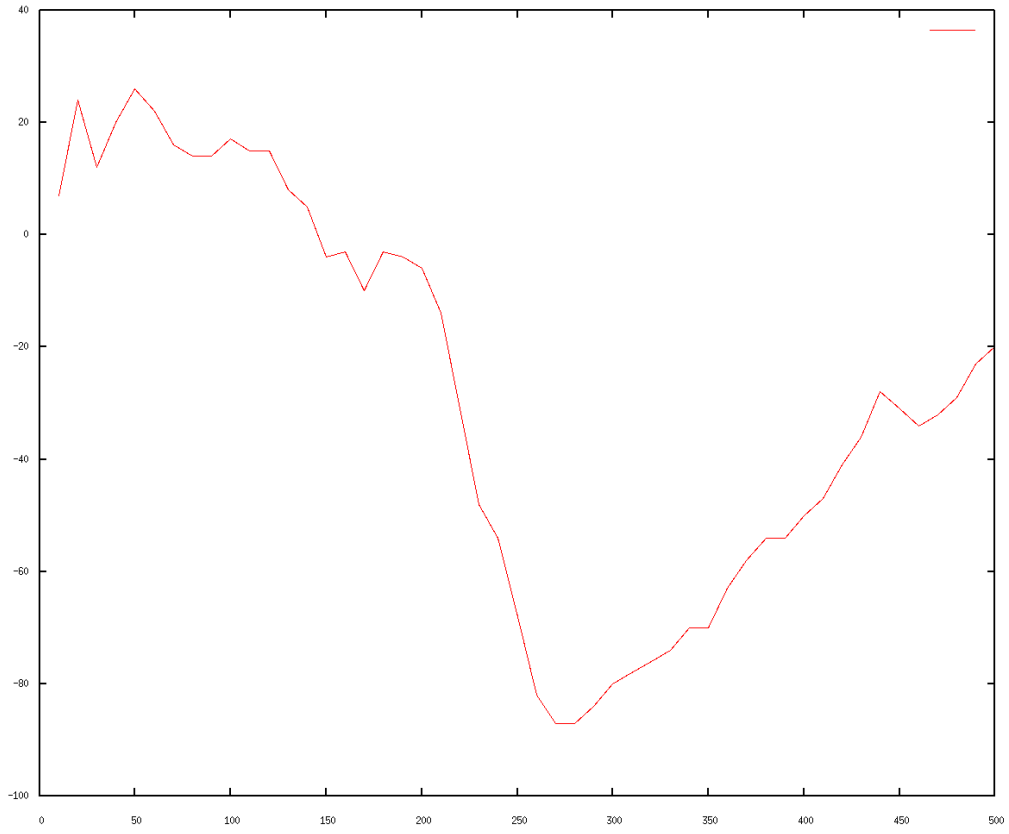


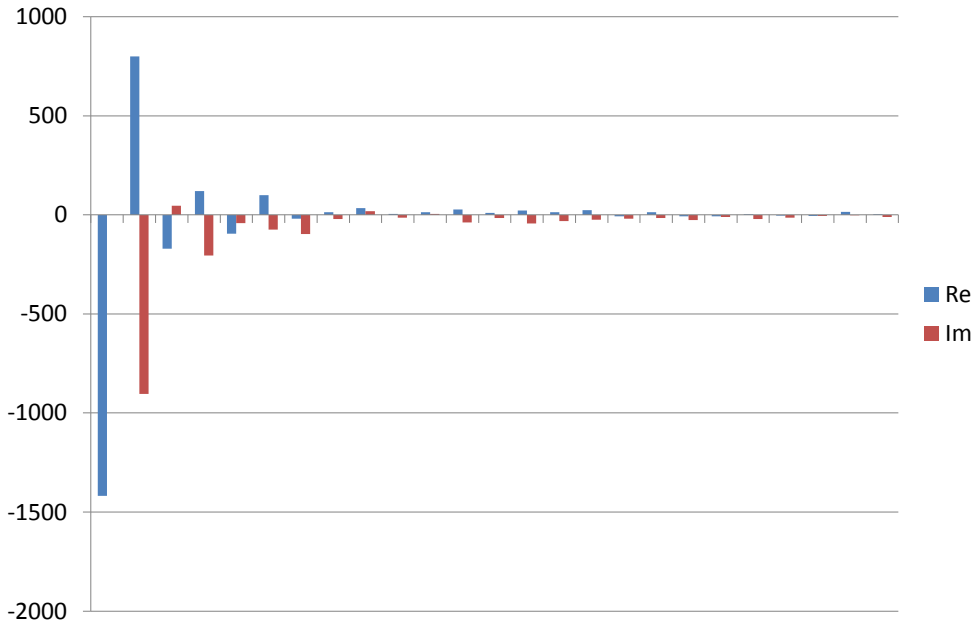
Abbildung 3.19: Skizze des Reliefmerkmals einer frontal zum Sensor stehenden Person.

on dar, welche in Amplitude, Frequenz und Phasenverschiebung skaliert werden kann. Der Realteil und der Imaginärteil der transformierten Eingabewerte aus Abbildung 3.20a ist in Abbildung 3.20b dargestellt. Mit Hilfe der Inversen Diskreten Fourier-Transformation (IDFT) ist es möglich durch Nutzung der transformierten Werte im Frequenzraum die Ausgangswerte wieder herzustellen. Indem vor der Rücktransformation Werte im Frequenzraum auf 0 gesetzt werden, ist es möglich, die Bedeutung einzelner Werte im Frequenzraum aufzuzeigen. Wird beispielsweise im Frequenzraum nur die komplexe Zahl s_0 behalten und alle anderen Werte $s_1, ..., s_{N-1}$ auf 0 gesetzt, kann der Sinusoid passend zu s_0 dargestellt werden. Abbildung 3.22a zeigt die ersten vier Sinusoiden, indem die IDFT jeweils einzeln auf die jeweiligen Elemente s_0, s_1, s_2, s_3 und s_4 ausgeführt wurden. Der erste Wert im Frequenzraum s_0 steht für den Gleichanteil der Funktion und bildet daher eine Funktion parallel zur Abszissenachse. Durch Überlagerung aller Sinusoide ist es möglich die Ausgangswerte wieder herzustellen. Die komplexen Werte innerhalb des Frequenzraumes sind nach ihrer Frequenz geordnet. Die Frequenzen, die sich im vorderen Bereich des Frequenzraumes befinden, stellen die wichtigsten Frequenzen dar.

Der zweite Wert s_1 stellt den wichtigsten Sinusoiden des Signals dar. Der Sinusoid zu s_1 weist eine hohe Ähnlichkeit zu den Ausgangswerten des Reliefmerkmals auf. Zur Verdeutlichung wurde in Abbildung 3.23b nur der Wert der ersten bei-



(a)



(b)

Abbildung 3.20: Fouriertransformiertes Reliefmerkmal: (a) Darstellung als Funktion (b) Frequenzspektrum

den Sinusoiden s_0 und s_1 verwendet um das Ausgangssignal wiederherzustellen. Abbildung 3.23 demonstriert wie durch Hinzunahmen weiterer Sinusoide das Ausgangssignal angenähert wird.

Die Transformation des Merkmals in den Frequenzraum bietet Vorteile. Beispielsweise zeigt Abbildung 3.23f wie durch das Entfernen des ersten Wertes s_0 der Gleichanteil aus dem Merkmal entfernt werden kann. Diese Eigenschaft ist nützlich, da der oberste Punkt des Kopfes, der als Referenzpunkt bei Erstellung des Merkmals dient, nicht immer an der selben Stelle des Kopfes liegen muss. Abbildung 3.24 skizziert einen solchen Fall, indem eine Verschiebung zwischen zwei ähnlichen Reliefmerkmalen besteht. Durch die Fourier-Transformation des Merkmals stellt diese Verschiebung kein Problem dar, da durch Entfernen des Gleichanteils eine Normalisierung durchgeführt wird.

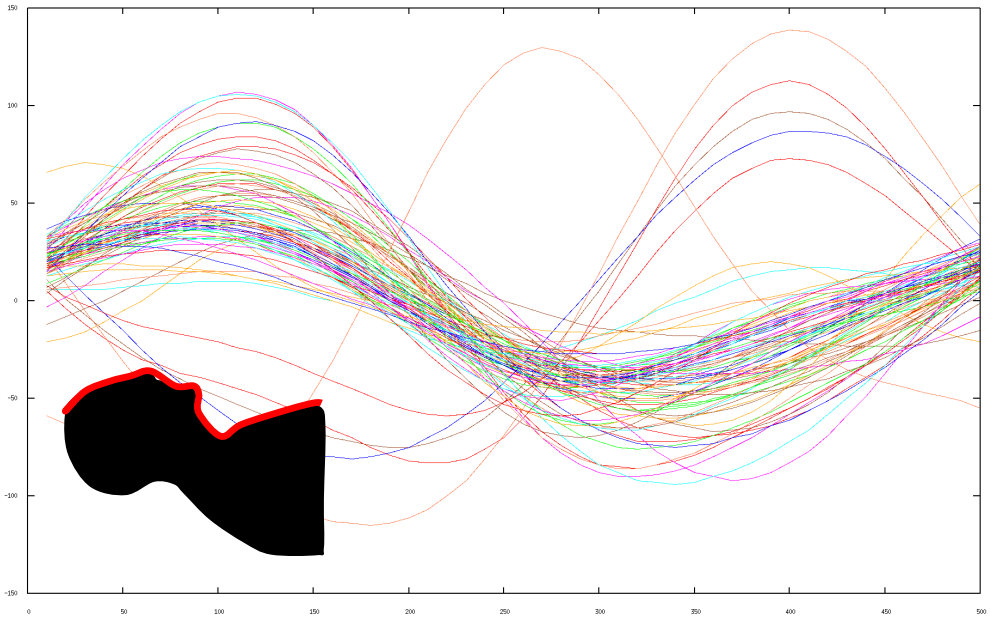
Ein weiterer Vorteil liegt in der Länge des Merkmals. Die Länge des Merkmales im Frequenzraum kann reduziert werden, um beispielsweise die Klassifikation zu beschleunigen.

Der größte Vorteil ist, dass im Frequenzraum die Werte geordnet sind. Bereits ein Sinusoid s_1 kann eine hohe Aussagekraft über das Merkmal haben. Dies wird in Abbildung 3.21a deutlich, in der die Sinusoide s_1 mehrerer Reliefmerkmale abgebildet sind. Die meisten Sinusoiden weisen eine sehr ähnliche Form auf. Auf der Abbildung sind ebenfalls Ausreißer zu erkennen die beispielsweise darauf zurückzuführen sind, dass eine Person am Bildrand des Sensors stand und das Reliefmerkmal nicht vollständig erfasst wurde. Zum Vergleich dazu zeigt Abbildung 3.21b das Reliefmerkmal verschiedener Objekte. Diese weisen viele unterschiedliche Sinusoide auf.

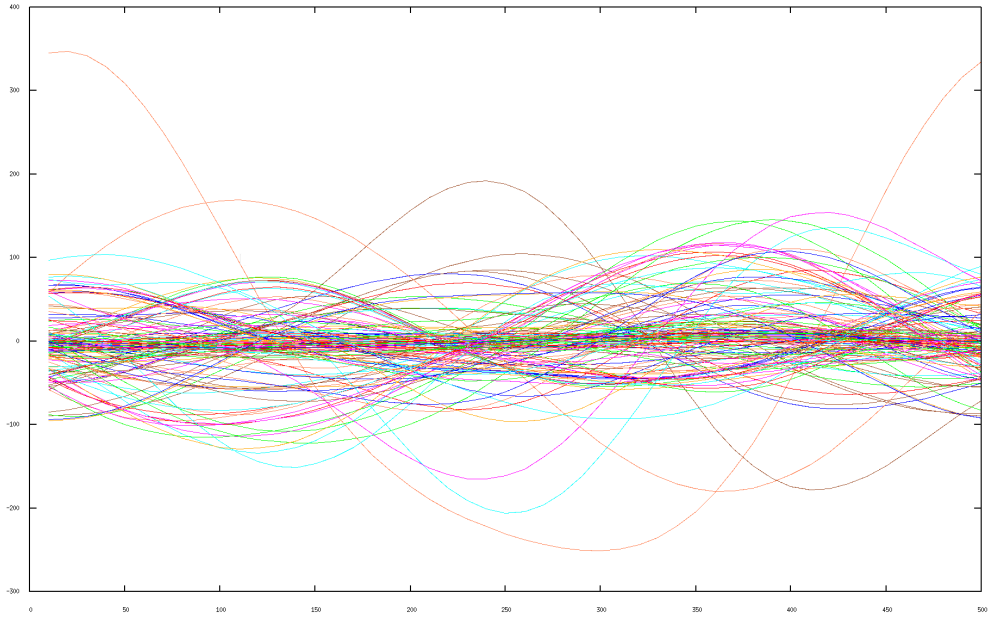
Mit Hilfe der Fourier-Transformation kann ein Merkmalsvektor $\mathbf{f}$ erzeugt werden, der die Elemente des Frequenzraums enthält (Formel 3.13). Da die Eingabewerte der Fourier-Transformation nur aus reellen Zahlen bestehen, gibt es nur $N/2$ unterschiedliche Frequenzen. Der erste Wert s_0 des Gleichanteils wird entfernt um eine Normalisierung durch eine gegebene Verschiebung zwischen beiden Merkmalen zu verhindern.

$$f_{k-1} = s_k, k = 1, ..., N/2 \tag{3.13}$$

Im Gegensatz zur Vorgehensweise von Hordern et al. [HK10] wird der Gleichanteil s_0 entfernt und nicht der gesamte Merkmalsvektor durch den Gleichanteil geteilt. Indem durch den Gleichanteil dividiert wird, wird das von Hordern et al. genutzte Merkmal skalierungsunabhängig. Dieser Effekt wird benötigt, da eine Person, die weiter entfernt steht, eine kleinere 2D-Form darstellt als eine näher stehende Person. Da das in dieser Arbeit entwickelte Relief- und Breitenmerkmal mit maßstabsgerechten Werten in Metern (bzw. Millimetern) arbeitet, stellt die Skalierung des Merkmals eine wichtige Information dar. Dies wird beispielsweise

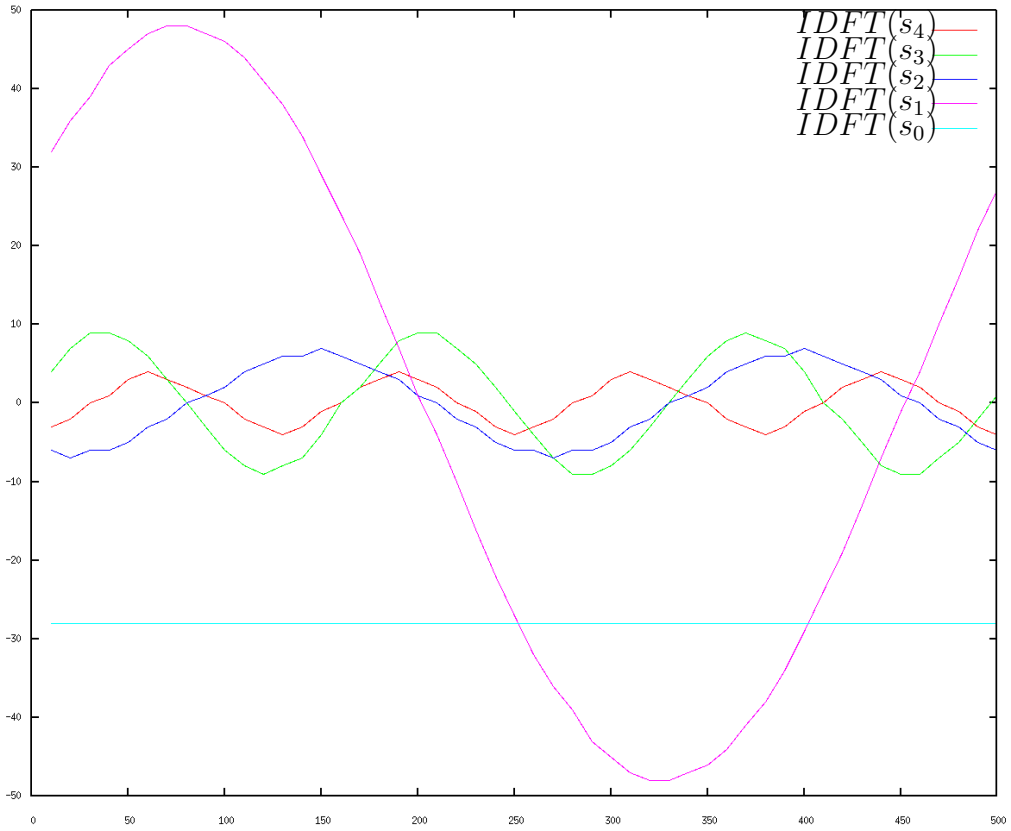


(a) Erster Sinusoid des Reliefs frontal stehender Personen



(b) Erster Sinusoid des Reliefs verschiedener Objekte

Abbildung 3.21: Darstellung des ersten Sinusoiden.



(a)

Abbildung 3.22: Darstellung der Sinusoide durch IDFT

durch die Darstellung des Sinusoiden aus Abbildung 3.23 deutlich, welche sich im selben Wertebereich wie das Ausgangssignal darstellen lässt. Dies bedeutet, dass das fouriertransformierte Reliefmerkmal indirekt geometrische Abmessungen einer Person enthält. Das Reliefmerkmal einer kleinen menschenähnlichen Spielzeugpuppe würde sich vom Reliefmerkmal einer lebensgroßen Person unterscheiden.

3.4.2 Training der Merkmale

Um zu ermitteln, ob es sich bei einem Kandidaten um ein Objekt oder eine Person handelt, muss das Merkmal entsprechend klassifiziert werden.

Da das Merkmal $\mathbf{f}$ einen Vektor darstellt, der mehrere Elemente enthält, ist eine direkte Visualisierung des Merkmalsraums nicht möglich. Es ist aber möglich ein einzelnes Element des Merkmals darzustellen. Das wichtigste Element des Merkmalsvektors ist das Element f_0, welches die komplexe Zahl s_1 im Frequenzraum darstellt. s_1 besteht aus dem Realteil $Re(s_1)$ und dem Imaginärteil $Im(s_1)$.

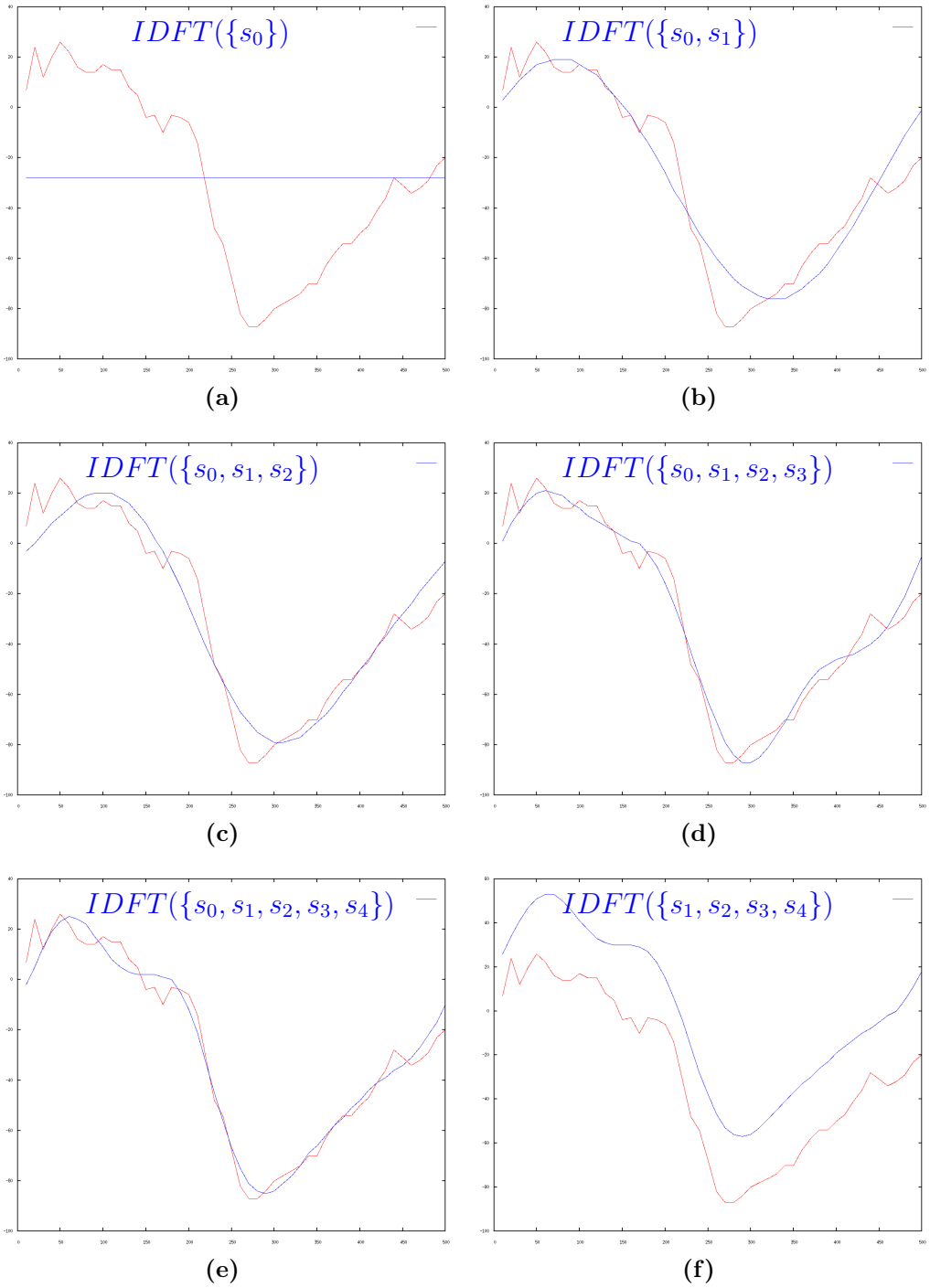


Abbildung 3.23: In Blau dargestellt: IDFT einzelne komplexe Werte des Frequenzspektrums. In Rot dargestellt: Ausgangswerte des Reliefmerkmals.

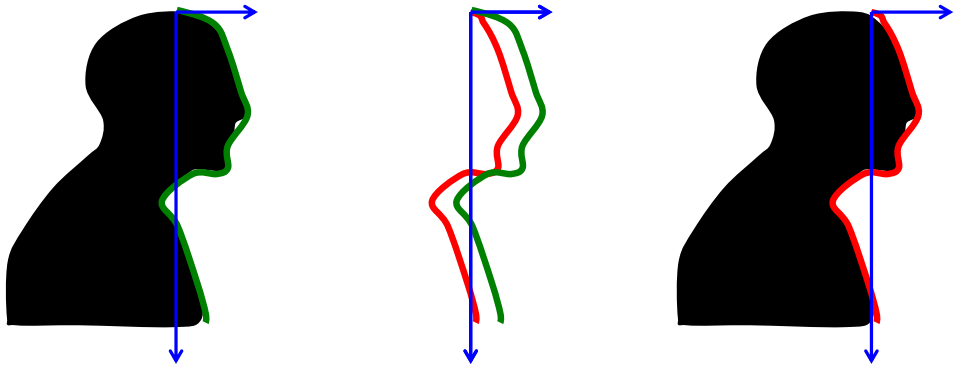


Abbildung 3.24: Normalisierung eines Merkmals mit Hilfe des Gleichanteils s_0. Durch Entfernen des Gleichanteils wird die horizontale Verschiebung zwischen beiden Reliefs behoben und die Merkmale werden etwa deckungsgleich.

Mit Hilfe von Trainingsdaten, bei denen bekannt ist ob es sich bei dem Kandidaten um ein Objekt oder eine Person handelt, lassen sich die Merkmale in ein 2D-Koordinatensystem auftragen. Abbildung 3.25a visualisiert einen solchen Merkmalsraum, indem sich auf der Abszissenachse $Re(s_1)$ und auf der Ordinatenachse $Im(s_1)$ befinden. In Grün dargestellt sind Merkmale von Personen, die dem Sensor zugewandt sind (frontal). Personen, welche vom Sensor abgewandt bzw. seitlich stehen, sind in rot eingezeichnet. Die übrigen Punkte in Blau stellen die Kandidaten dar, bei denen es sich nicht um Personen handelt.

Die Abbildung 3.25a demonstriert, dass es sich bei dem fouriertransformierten Merkmal um ein aussagekräftiges, starkes Merkmal handelt. Zu sehen ist, dass die Merkmale zugewandter bzw. abgewandter Personen deutliche Häufungspunkte im Merkmalsraum bilden. Zur Klassifikation ist deshalb ein starker Klassifikator beispielsweise eine SVM am geeignetsten. Eine SVM trennt den Merkmalsraum mit Hilfe einer Hyperebene in zwei Bereiche. Bei der Erstellung der Hyperebene wird versucht einen möglichst breiten Rand zwischen den abzutrennenden Klassen zu erhalten. Im genannten 2D-Fall aus Abbildung 3.25a entspricht diese Hyperebene einer Geraden. Eine Gerade g_1 trennt den Raum in zwei Teile (Abbildung 3.25b). Im linken Teil befinden sich die Objekte im rechten Teil die abgewandten und zugewandten Personen. Wie in der Abbildung zu erkennen, weist eine Trennung durch eine Gerade viele Fehlklassifikationen auf. Zudem ist der Abstand zwischen der Gerade und den Klassen sehr gering.

Die SVM bietet die Möglichkeit mehr als 2 Klassen voneinander zu trennen. Aufgrund der gestiegenen Anzahl der Klassen werden mehrere Hyperebenen zur Separierung der Klassen benötigt.

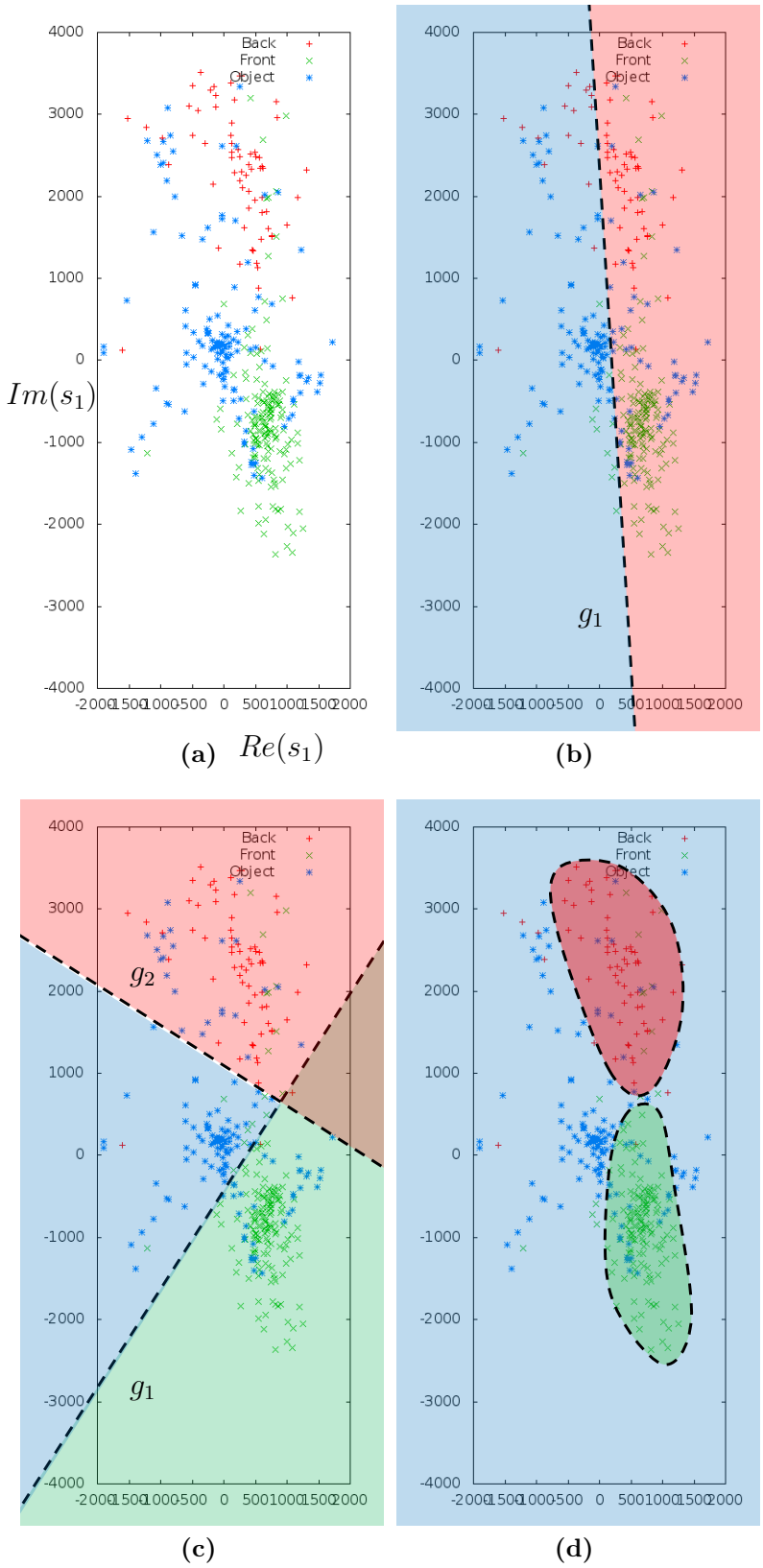


Abbildung 3.25: Skizze der Möglichkeiten einer Klassifikation: (a) Merkmalsraum, (b) lineare SVM mit 2 Klassen, (c) lineare SVM mit 3 Klassen, (d) nicht lineare SVM

Zur besseren Separierung der Klassen werden hierzu die Klassen der abgewandten und zugewandten Personen voneinander getrennt. Eine Gerade g_1, welche die Klasse der Objekte von der Klasse der zugewandten Personen abtrennt ist in Abbildung 3.25c zu sehen. Eine zweite Gerade g_2 kann die Klasse der Objekte von der Klasse der abgewandten Personen unterscheiden.

Im Vergleich zur Trennung mit nur einer Geraden wird durch das Aufteilen in mehrere Klassen der Rand zwischen den Klassen breiter. Eine Klassifikation ist somit besser und weniger anfällig für Überanpassung (engl. overfitting). Es gibt eine Reihe von Merkmalen die falsch klassifiziert werden, da eine Klassifikation nicht auf linearem Wege möglich ist.

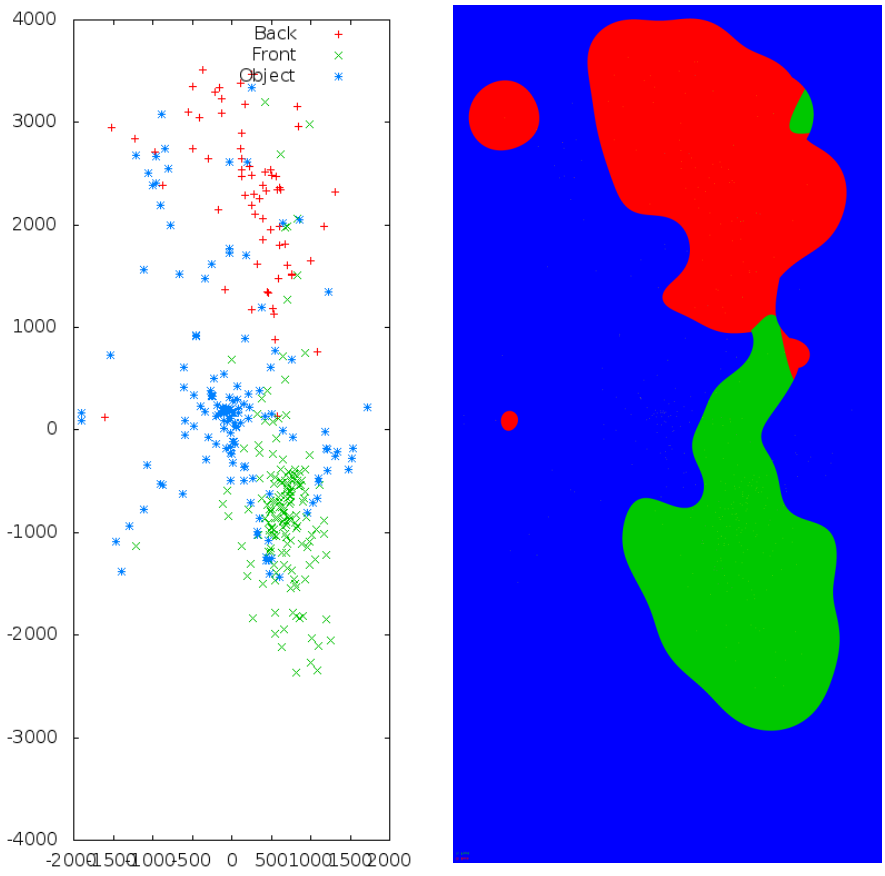
Idealerweise würde eine Klassifikation Inselbereiche rund um die Merkmale einer abgewandten und zugewandten Person erzeugen. Abbildung 3.25d skizziert solche Inselbereiche. Ein Punkt innerhalb der Darstellung steht für ein bestimmtes Sinusoid eines Merkmals. Eine zugewandte Person erzeugt ein typisches Sinusoid. Durch die Bildung eines Inselbereiches wird nun die Menge der möglichen typischen Sinusoide zum Rest abgegrenzt. Befindet sich das Merkmal innerhalb dieser Insel, gehört es zu einer Person, befindet es sich außerhalb, handelt es sich um ein Objekt, das keine Person ist.

Die Bildung dieser Inselbereiche ist nicht durch die Nutzung einer linearen SVM möglich. Aus diesem Grund wird eine nicht lineare SVM verwendet. Die nicht lineare SVM verwendet einen Kernel-Trick um die nicht linear trennbaren Klassen zu trennen. Hierzu wird der Merkmalsvektor auf einen höherdimensionalen Raum abgebildet. Durch diesen Trick werden die Merkmale im höherdimensionalen Raum durch eine Hyperebene trennbar. Die Visualisierung in diesem hörerdimensionalen Raum ist nicht mehr möglich, dennoch kann im zweidimensionalen Fall visualisiert werden, welcher Punkt innerhalb der 2D-Koordinaten zu welcher Klasse zugeordnet wird. Durch Verwendung eines RBF-Kernels, welcher eine radiale Funktion nutzt, wird der Inseleffekt erzielt. Die SVM von OpenCV stellt diesen Kernel-Trick implementiert zur Verfügung.

Abbildung 3.26b visualisiert das Ergebnis der trainierten SVM. Nach der Erstellung der SVM mit Hilfe der Trainingsdaten wurde jeder Punkt im Merkmalsraum mit Hilfe der SVM klassifiziert und entsprechend der ermittelten Klasse mit der passenden Farbe versehen.

Das Ergebnis entspricht einer trainierten SVM bei der nur der erste Wert des Merkmals verwendet wurde. Es ist nicht notwendig den gesamten Merkmalsvektor einzusetzen. Vor der Klassifikation des Merkmals und vor dem trainieren der SVM wird daher der Merkmalsvektor gekürzt. Der Wert $\theta_{Merkmalslaenge}$ gibt die Länge des genutzten Merkmalsvektors an.

Unter Verwendung der trainierten SVM wird ein zu ermittelndes Merkmal klassifiziert und die Person detektiert.



(a) **(b)**

Abbildung 3.26: Ergebnis der Klassifikation des Reliefmerkmals mit einer nichtlinearen SVM

Kapitel 4

Handdetektion

Im Anschluss an die Detektion der Person können die Hände der Person bestimmt werden.

4.1 Freistellen der Person

Für die Detektion der Hände ist es wichtig die Person möglichst exakt vom Hintergrund freizustellen. Aus der Kandidatensuche aus Kapitel 3.2 ist der höchste Punkt $\mathcal{I}(v_h, w_h)$ am Kopf der Person bekannt. Ausgehend von diesem Punkt wird ein *Region Growing* durchgeführt. Der höchste Punkt des Kopfes $\mathcal{I}(v_h, w_h)$ bildet die initiale Region. Zu dieser Region werden die benachbarten Punkte hinzugefügt. Die Punkte gelten als benachbart, wenn diese auf dem 2D-Bild benachbart sind. Die Nachbarn eines Punktes $\mathcal{I}(v, w)$ sind $\mathcal{I}(v+1, w)$, $\mathcal{I}(v-1, w)$, $\mathcal{I}(v, w+1)$ und $\mathcal{I}(v, w-1)$.

Es sei $\mathcal{I}(v_a, w_a)$ der Nachbar von $I(v_b, w_b)$. Befindet sich $\mathcal{I}(v_a, w_a)$ in der Region müssen folgende Bedingungen gelten damit auch $\mathcal{I}(v_b, w_b)$ in die Region aufgenommen werden kann. Die Distanz zum Nachbarn in 3D-Koordinaten des Punktes muss innerhalb eines Schwellwertes $\theta_{disNachbar}$ liegen (Formel 4.1).

$$\|\mathcal{I}(v_a, w_a) - \mathcal{I}(v_b, w_b)\| < \theta_{disNachbar} \tag{4.1}$$

Die Entfernung zum höchsten Punkt des Kopfes muss ebenfalls innerhalb eines Schwellwertes $\theta_{disKopf}$ liegen (Formel 4.2).

$$\|\mathcal{I}(v_h, w_h) - \mathcal{I}(v_b, w_b)\| < \theta_{disKopf} \tag{4.2}$$

Abbildung 4.1 skizziert einzelne grobe Zwischenschritte des Regionenwachstums. Ist das Wachstum der Region abgeschlossen sind die Pixel und somit auch die 3D-Punkte, die zur Person gehören genau abgegrenzt. Der folgende Teil der

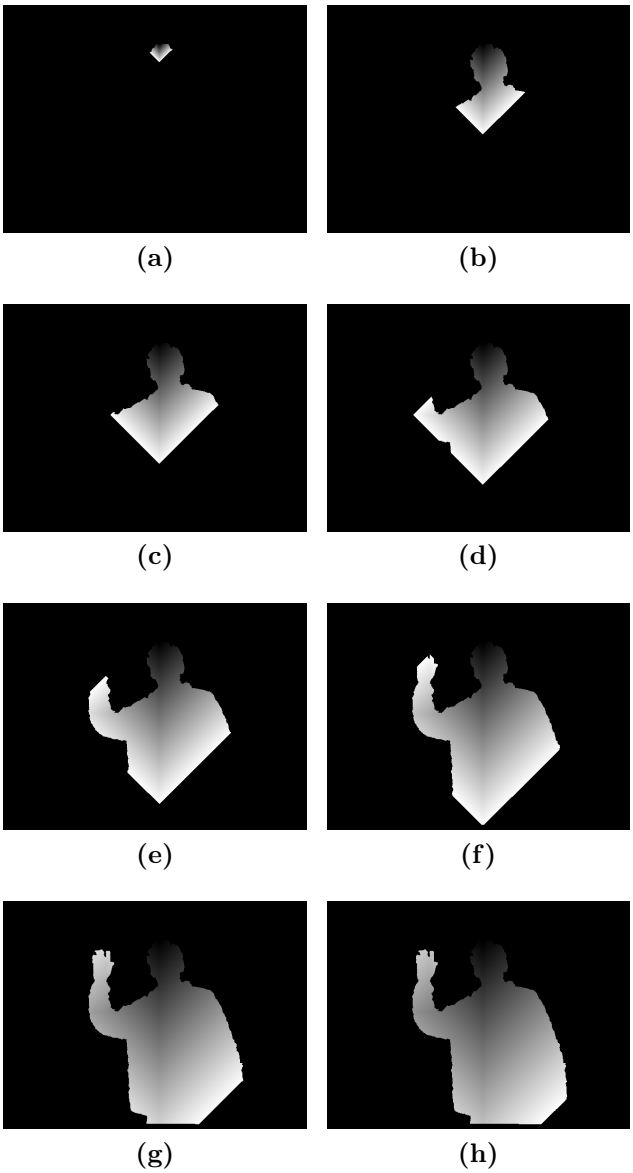


Abbildung 4.1: Freistellen der Region der Person ausgehend vom höchsten Punkt des Kopfes.

Handdetektion arbeitet ausschließlich auf der so bestimmten Region der Person. Die Region wird in einer Matrix M_{Person} gespeichert, welche die selbe Spaltenzahl V und Zeilenzahl W wie das Tiefenbild aufweist. Befindet sich ein Punkt an der Pixelstelle (v, w) innerhalb der Region der Person wird der Matrix an der Stelle $M_{Person}(v, w)$ der Wert 1 zugewiesen. Gehört das Pixel nicht zur Region ist der Wert der Matrix 0. Die Matrix kann dabei als Bild aufgefasst werden. In diesem Falle ein Binärbild.

4.2 Handentfernung

Nachdem die Person vom Rest des Bildes mit Hilfe des Regionenwachstums freigestellt ist, werden Eigenschaften dieser Region bestimmt. Eine nützliche Eigenschaft ist die Entfernung der Hand vom Körper einer Person. Streckt eine Person beispielsweise den Arm aus befindet sich die Hand am weitesten entfernten Punkt vom Körper.

Der Vorteil dieser Vorgehensweise gegenüber den anderen Verfahren aus Abschnitt 2.2.4 aus dem Kapitel der verwandten Arbeiten ist, dass auch Hände detektiert werden können, die sich nicht zwischen Person und Kamera befinden. Beispielsweise verwenden Bernard et al. [BB10] ein Verfahren um das Tiefenbild in mehrere Tiefenbereiche zu unterteilen. Der vordere Tiefenbereich entspricht der Hand der hintere Tiefenbereich der Person. Wird die Hand in Richtung Sensor gestreckt stellt die Hand ein lokales Minimum dar und kann somit detektiert werden. Steht eine Person frontal zum Sensor und streckt die Hand zur Seite ist eine Detektion mit Hilfe der Entfernung zum Sensor nicht mehr möglich.

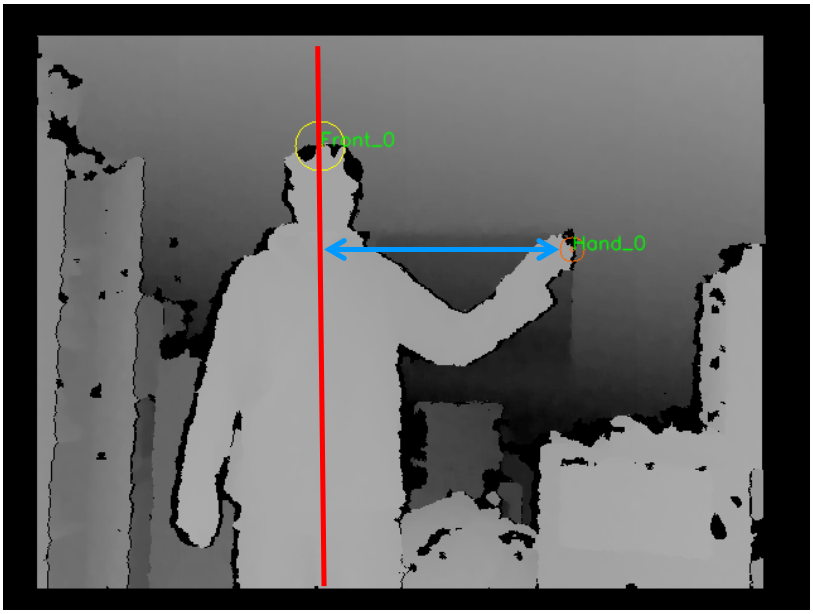
Aus diesem Grund verwendet das in dieser Arbeit entwickelte Verfahren den Abstand zwischen Hand und Körper zur Detektion der Hand. Der Körper wird hierzu auf eine Gerade reduziert. Die Gerade verläuft durch den Schwerpunkt $\mathbf{s}$ der Person parallel zur $y - Achse$ des Koordinatensystems der Punktewolke. Zur Bestimmung des Schwerpunktes werden alle 3D-Punkte I_P der Person betrachtet (Formel 4.3).

$$\mathcal{I}_P(v, w) = \begin{cases} \mathcal{I}(v, w), & \text{wenn } M_{Person}(v, w) = 1 \\ (0, 0, 0)^T, & \text{sonst} \end{cases} \tag{4.3}$$

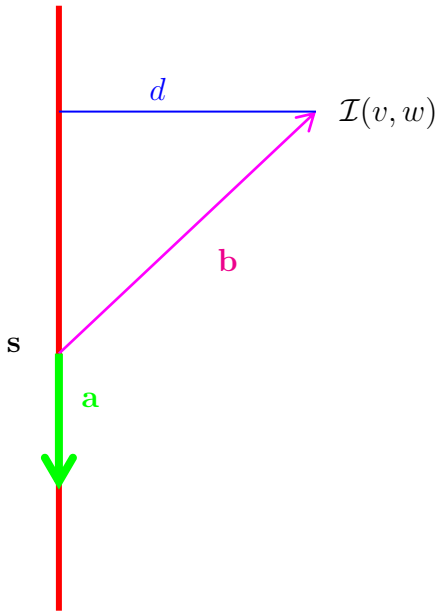
Den Schwerpunkt $\mathbf{s}$ bildet der Mittelwert aller Punkte (Formel 4.4)

$$\mathbf{s} = \frac{\sum_{v=0}^{V} \sum_{w=0}^{W} \mathcal{I}_P(v, w)}{\sum_{v=0}^{V} \sum_{w=0}^{W} M_{Person}(v, w)} \tag{4.4}$$

In Abbildung 4.2a ist diese Gerade in Rot eingezeichnet. Ziel ist es die in Blau eingezeichnete Entfernung in Metern zwischen der Geraden und jedem Punkt innerhalb einer Person zu bestimmen.



(a) Skizze der Situation



(b) Skizze der Berechnung

Abbildung 4.2: Bestimmung der Entfernung zwischen einem Punkt und einer Geraden.

Die Gerade ist definiert durch einen Punkt $\mathbf{s}$, der auf der Geraden liegt und dem Richtungsvektor $\mathbf{a}$. Da die Gerade parallel zur $y-$Achse verläuft beträgt der Richtungsvektor $\mathbf{a} = (0, 1, 0)^T$ (siehe Abbildung 4.2b). Zwischen dem Punkt $\mathbf{s}$ auf der Geraden und dem Punkt $\mathcal{I}(v, w)$ wird der Vektor $\mathbf{b}$ gebildet (Formel 4.5).

$$\mathbf{b} = \mathbf{s} - \mathcal{I}(v, w) \tag{4.5}$$

Die Entfernung d_{vw} einer Geraden zu dem Punkt $I(v, w)$ lässt sich mit Hilfe des Kreuzproduktes berechnen.

$$d_{vw} = \frac{\|\mathbf{a} \times \mathbf{b}\|}{\|\mathbf{a}\|} \tag{4.6}$$

Unter Verwendung von Formel 4.5, 4.6 und $\mathbf{a} = (0, 1, 0)^T$ ergibt sich Formel 4.7.

$$d_{vw} = \|(0, 1, 0)^T \times (\mathbf{s} - \mathcal{I}(v, w))\| \tag{4.7}$$

Mit Hilfe der Formel 4.7 wird eine Matrix M_D der Distanzwerte innerhalb der Region von M_{Person} bestimmt (siehe Formel 4.8).

$$M_D(v, w) = \begin{cases} d_{vw}, & \text{wenn } M_{Person}(v, w) = 1 \\ 0, & \text{sonst} \end{cases} \tag{4.8}$$

4.3 Hautfarbe

Neben der Eigenschaft der Position der Hand wird auch die Eigenschaft der Farbe der Hand verwendet. Aus dem Abschnitt 2.2.1, indem die verwandten Arbeiten zur Handdetektion mittels Farbe aufgeführt sind, sind verschiedene Vorgehensweisen bekannt.

Das Prinzip der Handdetektion durch Farbe beruht darauf die Hautfarbe zu detektieren. Es gibt dabei mehrere Möglichkeiten. Erstens die Hautfarben auf bestimmte Farbwerte festzulegen, zweitens die Hautfarben aus einem Datenbestand annotierter Bilder zu ermitteln und drittens die Hautfarbe aus der Gesichtsregion aus dem Bild zu bestimmen. Da durch die Personendetektion der Bereich der Person und somit auch der Bereich des Kopfes und des Gesichts bekannt sind wird in dem hier entwickelten Ansatz die Hautfarbe aus dem Gesichtsbereich extrahiert. Mit Hilfe der Hautfarbe des Gesichtes werden hautfarbene Bereiche der Person ermittelt.

4.3.1 Ermittlung der Gesichtsposition

Im ersten Schritt wird der Bereich des Kopfes genauer bestimmt. Es werden die Punkte des Bildes bestimmt, welche innerhalb der Person liegen ($M_{Person}(v, w) =$

1) und eine Maximalentfernung θ_{Kopf} zum höchsten Punkt des Kopfes $\mathcal{I}(v_h, w_h)$ nicht überschreiten (Formel 4.9).

$$M_{Kopf}(v,w) = \begin{cases} 1, & \text{wenn } M_{Person}(v,w) = 1 \text{ und } \|\mathcal{I}(v_h,w_h) - \mathcal{I}(v,w)\| < \theta_{Kopf} \\ 0, & \text{sonst} \end{cases} \tag{4.9}$$

Analog zu Formel 4.9 ermittelt Formel 4.10 ebenfalls den Kopfbereich, verwendet aber einen höheren Schwellwert $\theta_{KopfG} > \theta_{Kopf}$. Der so ermittelte größere Bereich des Kopfes wird später verwendet um hautfarbene Regionen des Kopfes auszuschließen.

$$M_{KopfG}(v,w) = \begin{cases} 1, & \text{wenn } M_{Person}(v,w) = 1 \text{ und } \|\mathcal{I}(v_h,w_h) - \mathcal{I}(v,w)\| < \theta_{KopfG} \\ 0, & \text{sonst} \end{cases} \tag{4.10}$$

In Abbildung 4.3a und 4.3b sind M_{Person} und M_{Kopf} als Binärbilder dargestellt. Zur Bestimmung des Gesichtsbereiches wird ein Rechteck um die gefüllten Bereiche des Binärbildes des Kopfes eingefasst. Dieses Rechteck, in Abbildung 4.3d und 4.3e in Grün eingezeichnet, beinhaltet den Kopfbereich der Person. Dieser Kopfbereich enthält neben dem Gesicht die Kopfhaare der Person. Aus diesem Grund wird proportional zum Kopfbereich (grünes Rechteck) ein Rechteck für das innere des Gesichtsbereiches erstellt (blaues Rechteck). Breite und Höhe des inneren Rechteckes entsprechen dabei 60% der Breite b_{Kopf} und 60% der Höhe h_{Kopf} des äußeren Rechteckes. Wie in Abbildung 4.3e skizziert ist das innere Rechteck horizontal aber nicht vertikal zentriert. Es wird ein Abstand von 30% der Höhe des äußeren Rechtecks eingehalten, damit die Kopfhaare nicht in den Gesichtsbereich fallen. Anhand der Farbwerte innerhalb des inneren Bereiches kann die Hautfarbe der Person ermittelt werden.

4.3.2 Erstellung des Farbhistogrammes

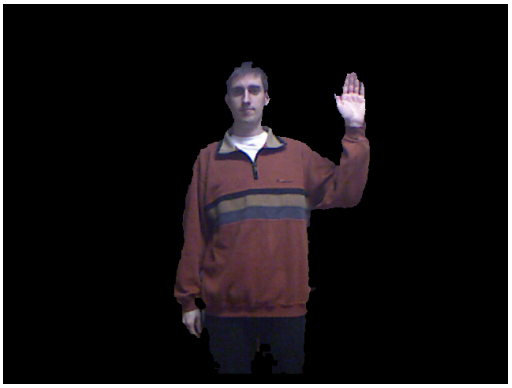
Der Sensor der Kinect liefet die Farbwerte als RGB-Bild. Die Punktewolke ist deckungsgleich zu den Farbwerten des RGB-Bildes des Sensors (Abbildung 4.3c und 4.3a). Es ist somit möglich zu jedem Punkt der Punktewolke den passenden RGB-Farbwert zu bestimmen.

Um die Eigenschaft der Farbe des Gesichtsbereiches zu speichern wird ein Farbhistogramm aus dem Gesichtsbereich erstellt. Wie aus Abschnitt 2.2.1 bekannt eignen sich zur Detektion der Hautfarbe besonders Farbräume, welche die Beleuchtungskomponente einzeln aufführen. Hierzu zählt der HSV-Farbraum, der sich in der Literatur [GZC+10, CPVC07] schon als sehr geeignet herausgestellt hat.

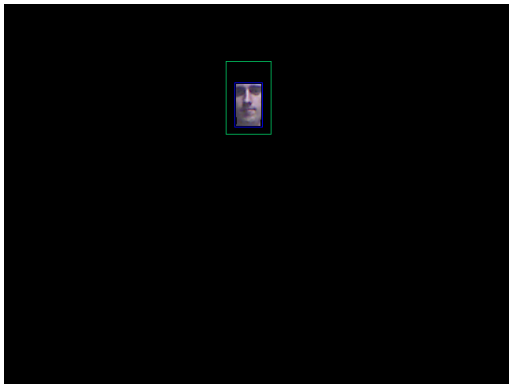
Das Farbhistogramm des HSV-Farbraums wird als 2D-Histogramm erstellt. Eine Achse des Histogramms stellt dabei den Farbwert H und die andere Achse

(a) Binärbild der Person

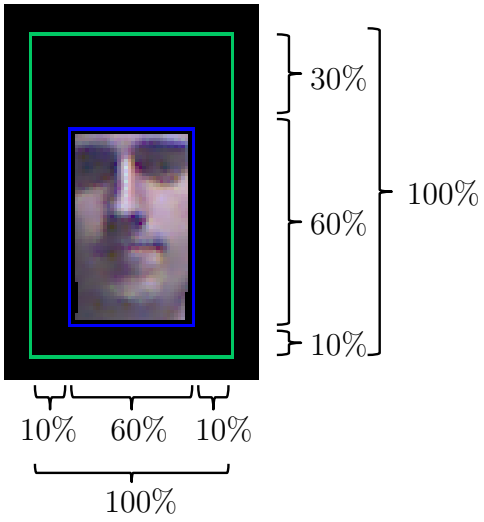
(b) Binärbild des Kopfbereiches



(c) Farbbild der Person



(d) Gesichtsbereich des Farbbildes



(e) Skizze zur Erzeugung des Gesichtsbereiches

Abbildung 4.3: Extraktion des Gesichtsbereiches einer Person.

die Sättigung S dar. Das Farbhistogramm des Gesichtsbereiches aus Abbildung 4.3d ist in Abbildung 4.4a dargestellt. Zum Vergleich ist das Farbhistogramm der gesamten Person in Abbildung 4.4b dargestellt. Die Gesichtsfarben bilden im 2D-Histogramm einen Häufungspunkt. Bei Farbwerten, die sich innerhalb des Häufungspunktes befinden handelt es sich um die Farbwerte der Hautfarbe. Das Histogramm kann teilweise verunreinigt sein, da es Farbwerte der Augenbrauen oder Augen enthalten könnte. Aus diesem Grund werden die Klassen des Farbhistogrammes ermittelt, welche die höchsten Werte aufweisen.

Das Farbhistogramm kann als Matrix M_{Histo} aufgefasst werden. $M_{Histo}(h, s) = |b_{hs}|$ stellt einen Wert des Histogramms an der Indexstelle h und s dar. $|b_{hs}|$ ist definiert als die Anzahl der Werte innerhalb einer Histogrammklasse b_{hs}. Die Matrix hat dabei N Spalten und N Zeilen. Das Histogramm besitzt somit N^2 viele Klassen.

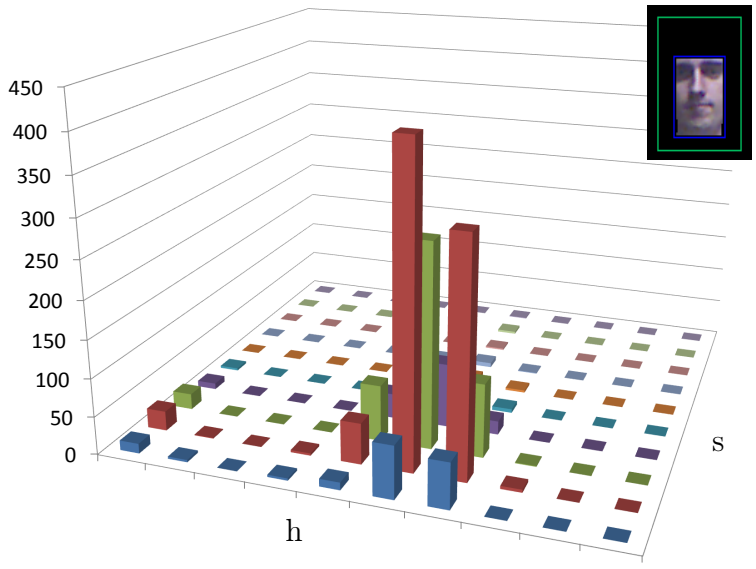
Ziel ist es die Menge Λ der Histogrammklassen zu ermitteln, welche zur Hautfarbendetektion am geeignetsten sind. Histogrammklassen, die nur wenig gefüllt sind, sind weniger aussagekräftig, da sie beispielsweise Farbwerte der Augenbrauen oder Farbrauschen beinhalten. Eine Möglichkeit die Menge Λ zu erstellen ist es, die Histogrammklassen zu verwenden, die eine Mindestanzahl von Werten aufweisen. Das Problem bei dieser Vorgehensweise ist, dass der Schwellwert stark von der Größe des Gesichtsbereiches und der gewählten Anzahl der Klassen N^2 des Histogramms abhängt. Um dieses Problem zu verhindern werden zur Menge Λ solange Klassen hinzugefügt bis die akkumulierte Gesamtanzahl der Histogrammwerte von Λ einen Schwellwert θ_{Histo} relativ zur gesamten Anzahl der im Histogramm befindlichen Werte erreicht hat. Liegt der Schwellwert θ_{Histo} beispielsweise bei $0,6$ befinden sich 60% der Histogrammwerte innerhalb der ausgewählten Klassen Λ. Im folgenden Pseudocode ist die Ermittlung der Menge Λ skizziert.

Algorithmus 2 Ermittlung der Menge der besten Histogrammklassen Λ

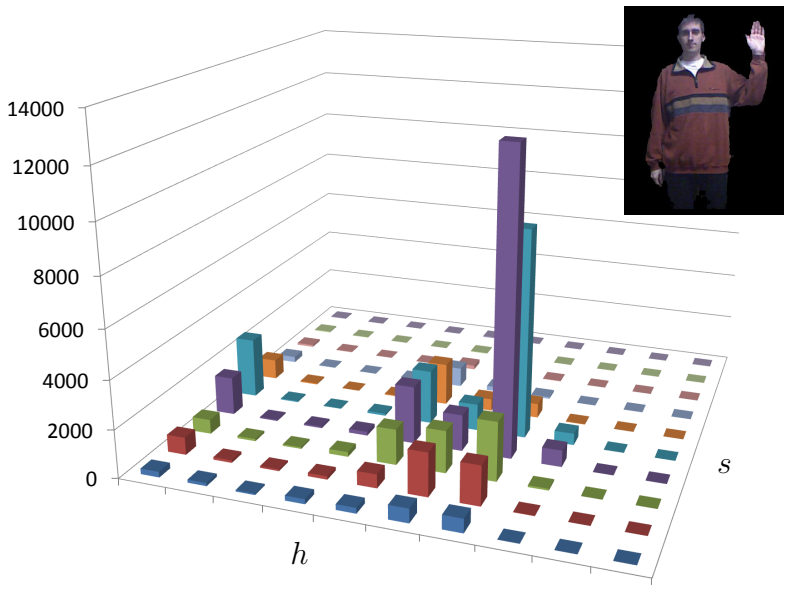
$\Lambda \leftarrow \emptyset$
$c \leftarrow 0$
while $\frac{c}{\sum_{h=0}^{N}\sum_{s=0}^{N} M_{Histo}(h,s)} < \theta_{Histo}$ **do**
  $b \leftarrow$ bestimme die Klasse b mit dem höchsten Wert $|b|$ für die gilt $b \notin \Lambda$
  $c \leftarrow c + |b|$
  $\Lambda \leftarrow \Lambda \cup b$
end while

4.3.3 Hautfarbenbewertungswert

Nachdem die Menge Λ der Klassen des Farbhistogramms aus dem Gesicht extrahiert wurde, wird das Histogramm zur Detektion hautfarbener Bereiche eingesetzt.



(a) Histogramm aus dem Gesichtsbereich



(b) Histogramm aus dem Personenbereich

Abbildung 4.4: 2D-Histogramm der h und s Werte des HSV-Farbraums.(Die verwendeten Farben zur Darstellung dienen nur zur besseren Unterscheidung der einzelnen Klassen. Die Farben stehen in keinem Verhältnis zum Farbwert der Klasse.)

Ein Pixel an der Stelle (v, w) wird als Hautfarbe betrachtet, wenn der Farbwert im HSV-Farbraum in der Klasse des Histogramms der Menge Λ liegt (Formel 4.11). b_{vw} stelllt die Histogrammklasse zum Farbwert an der Stelle (v, w) des Farbbildes dar.

$$M_{Haut}(v,w) = \begin{cases} 1, & \text{wenn } b_{vw} \in \Lambda \text{ und } M_{Person}(v,w) = 1 \\ & \quad \text{und } M_{KopfG}(v,w) = 0 \\ 0, & \text{sonst} \end{cases} \tag{4.11}$$

Das entstehende Binärbild zu $M_{Haut}(v, w)$ ist in Abbildung 4.5*b* dargestellt. Dieses Binärbild wird in mehreren Schritten aufbereitet. Um das Rauschen zu verringern wird auf das Binärbild ein Medianfilter angewendet (Abbildung 4.5c). Im Anschluss daran werden mit einer Dilatation gefolgt von einer Erosion des Binärbildes kleine Lücken geschlossen (Abbildung 4.5d). Mit Hilfe einer erneuten Erosion werden kleinere Bereiche unterdrückt. Das erstellte Binärbild in Abbildung 4.5e wird im Folgenden durch $M_{HautB}(v, w)$ dargestellt. Mit Hilfe von M_{HautB} wird eine Matrix $M_{Hautwert}$ erstellt. Die Grundidee ist bei der Bewertung eines Pixels die Nachbarschaft des Pixels zu betrachten. Je mehr hautfarbene Pixel sich innerhalb der Nachbarschaft befinden, desto wahrscheinlicher ist es, dass das Pixel zu einer Hand gehört. Die Größe der Nachbarschaft wird abhängig von der Breite des Kopfes in Pixeln bestimmt. In Formel 4.12 ist dargestellt wie sich der Bewertungswert berechnet. Der Wert k entsprich dabei 25% der Kopfbreite b_{Kopf} in Pixeln.

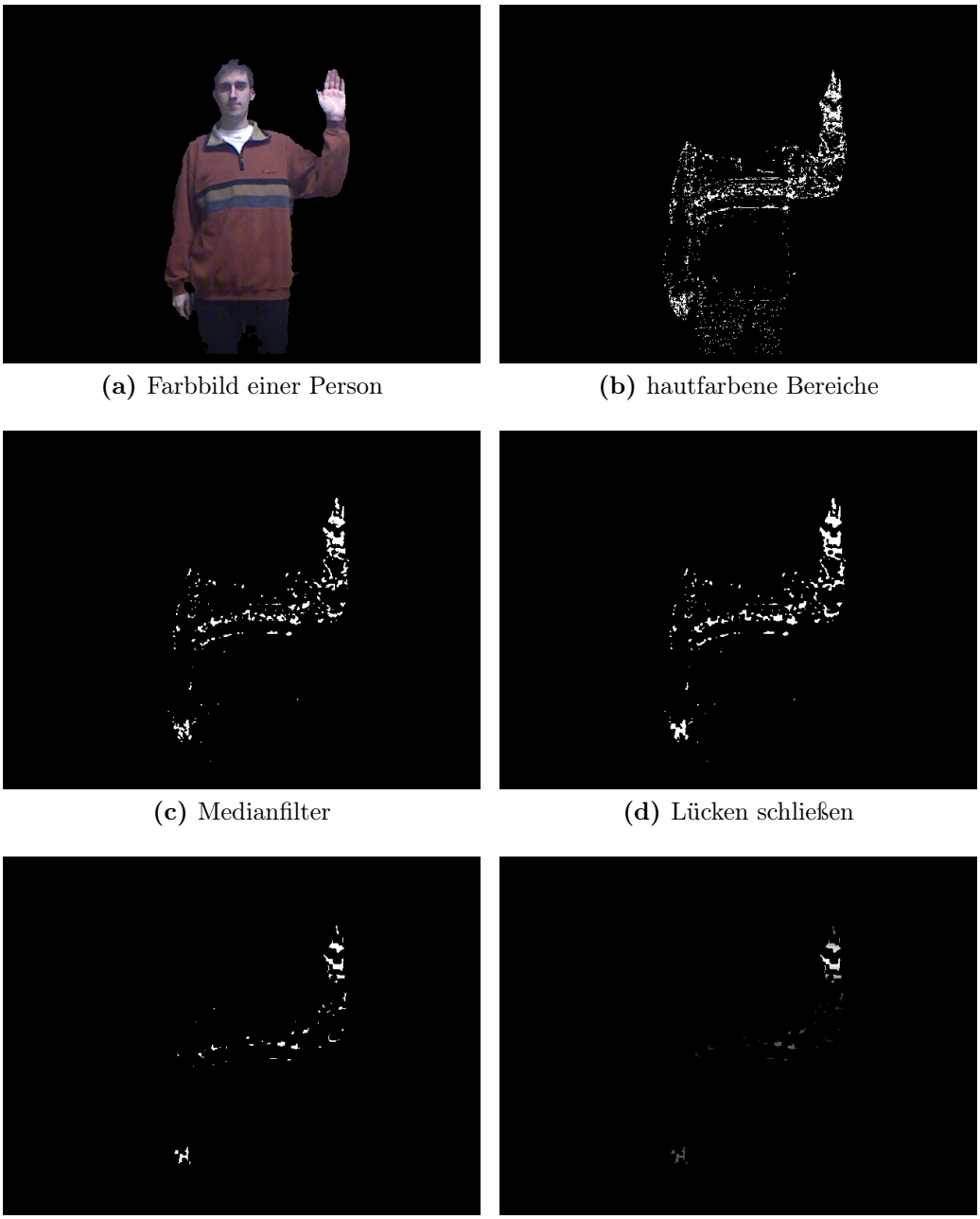
$$M_{Hautwert}(v,w) = \sum_{i=v-k}^{v+k} \sum_{j=w-k}^{w+k} M_{Haut}(i,j) \tag{4.12}$$

Die Matrix $M_{Hautwert}$ ist in Abbildung 4.5f visualisiert, indem die Werte auf ein Grauwertbild abgebildet wurden. Mit Hilfe dieser Daten kann nun im Folgenden die Hand detektiert werden.

4.4 Handdetektion

Zur Detektion der Hand wird die Entfernung der Hand zum Körper (aus Abschnitt 4.2) und die Information der Hautfarbe (aus Abschnitt 4.3) verwendet.

Die Entfernung der einzelnen Punkte zur vertikalen Geraden einer Person ist in M_D gespeichert und in Abbildung 4.6a dargestellt. M_D enthält Entfernungswerte, die in Metern angegeben sind. Der Bewertungswert zur Beurteilung der Hautfarbe liegt in $M_{Hautwert}$ vor (Abbildung 4.6b). Um die Bewertungswerte aus Entfernung und Hautfarbe zu kombinieren werden beide auf einen Bereich zwischen 0 und 1 normiert. Bei der Entfernung entspricht 1 Meter 100%. Eine Normalisierung ist



(a) Farbbild einer Person

(b) hautfarbene Bereiche

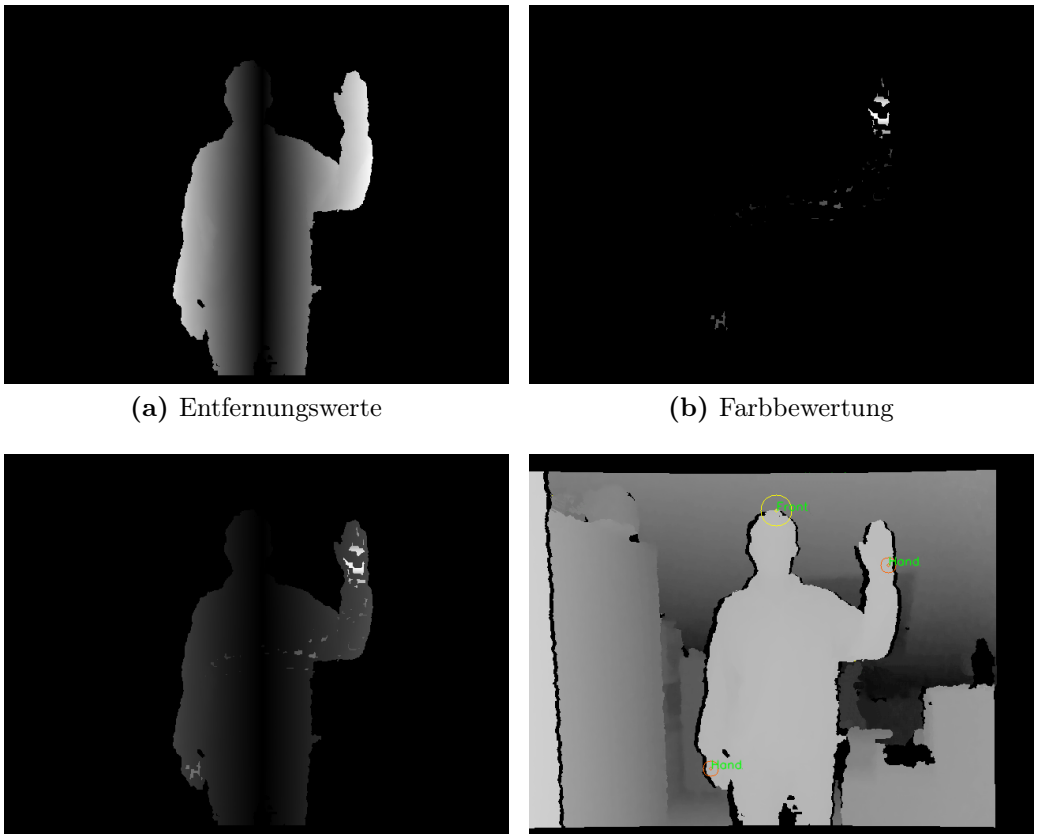
(c) Medianfilter

(d) Lücken schließen

(e) Erosion

(f) Bewertungswert der Hautfarbe

Abbildung 4.5: Verarbeitung der gefundenen hautfarbenen Bereiche.



(a) Entfernungswerte **(b)** Farbbewertung

(c) Kombination aus Farbe und Entfernung **(d)** Ergebnis der Detektion

Abbildung 4.6: Detektion der Hand mit Hilfe der Entfernung und Farbe der Hand.

durch die Angabe in Metern somit nicht mehr notwendig. Der Bewertungswert der Hautfarbe wird normalisiert indem durch den Maximalwert $max(M_{Hautwert})$ dividiert wird.

Die normierten Bewertungswerte werden durch Addition zu einem Bewertungswert zusammengefasst. Mit Hilfe der Faktoren α und β kann der Einfluss der Bewertungswerte verändert werden. Sind $\alpha = \beta = 1$ werden sowohl Farbe als auch Entfernung gleich gewertet. Wird beispielsweise β verringert wird der Einfluss der Farbe verringert. Die Berechnung der kombinierten Bewertung $M_{Bewertung}$ wird in Formel 4.13 berechnet.

$$M_{Bewertung}(v,w) = \alpha \cdot M_D(v,w) + \beta \cdot \frac{M_{Hautwert}(v,w)}{max(M_{Hautwert})} \tag{4.13}$$

Abbildung 4.6c visualisiert den Bewertungswert.

Durch die Verwendung des Faktors β ist es möglich bestimmte Situationen gesondert zu behandeln. Eine solche Situation trifft ein, wenn die Hautfarbende-

tektion fehlschlägt. Es kann beispielsweise vorkommen, dass große Teile der Kleidung als Hautfarbe erkannt werden. Dies kann unterschiedliche Ursachen haben. Die ermittelten Hautfarben des Gesichtes können durch starke Abschattung beeinträchtigt werden. Wird der Gesichtsbereich nicht richtig bestimmt könnten farbige Teile des Kragens der Person mit in das Histogramm des Gesichtes aufgenommen werden. Weiterhin könnte die Farbe der Kleidung der Person ihrer Hautfarbe ähneln. Die Person könnte viele hautfarbene Flächen besitzen beispielsweise indem sie kurze Kleidung trägt. Eine Ermittlung der Hände über die Farbe wird dadurch fehleranfälliger. Aus diesem Grund wird das Verhältnis γ der der Fläche des umschließenden Rechteckes des Kopfes zur Fläche des Kopfbereiches der Person bestimmt (Formel 4.14). Das Rechteck des Kopfes hat die Höhe h_{Kopf} und Breite b_{Kopf} und wurde in Abschnitt 4.3.1 ermittelt.

$$\gamma = \frac{\sum_{v=0}^{V} \sum_{w=0}^{W} M_{Haut}(v, w)}{h_{Kopf} \cdot b_{Kopf}} \tag{4.14}$$

Liegt dieses Verhältnis γ unterhalb des Schwellwertes $\theta_{Hautanteil}$ wird β auf 0 gesetzt und die Hautfarbenerkennung deaktiviert. Der Bewertungswert enthält dann nur die Entfernung der Hand.

Um die Hand mit Hilfe des Bewertungswertes zu detektieren wird das Maximum innerhalb der Matrix $M_{Bewertung}$ bestimmt. Der Bereich in der Nachbarschaft des Maximums und der Bereich des Maximums selbst werden ausgeschlossen und auf den restlichen Werten der Matrix das verbleibende Maximum ermittelt. Die beiden gefundenen Maxima stellen die Hände der Person dar.

Neben der Position der Hände in 3D-Koordinaten ist zudem die Entfernung zum Körper bekannt. Aufbauend auf diesen Daten kann eine Filterung der gefundenen Hände durchgeführt werden. Soll z.B. in einem Anwendungsfall eine aufzeigende Person gefunden werden könnten alle gefunden Hände verworfen werden, welche sich nicht oberhalb der Kopfhöhe befinden. Besteht der Anwendungsfall darin die Richtung zu ermitteln in der eine Person die Hand ausstreckt könnte diese über die bekannte Position des Kopfes und die Position der Hand ermittelt werden. Die Entfernung die eine Hand vom Körper hat kann genutzt werden um die Bedeutung der Hand zu bewerten. Streckt eine Person ihre Hand weit aus könnte dies z.B. bedeuten, dass sie eine Richtung angeben will oder etwas greifen möchte. Diese Hand hat eine größere Bedeutung als eine Hand, die nur am Körper anliegt.

Kapitel 5

Evaluation

In diesem Kapitel wird die Evaluation der Personen- und Handdetektion durchgeführt. Im ersten Abschnitt wird die Personendetektion und im zweiten Abschnitt die Handdetektion evaluiert.

5.1 Personendetektion

Für die Evaluation wurden mit der Kinect Szenen aufgenommen auf denen eine oder mehrere Personen zu sehen sind. Die Personen bewegten sich durch den Raum oder saßen auf einem Stuhl. Da sich die Personen bewegten wurden sowohl dem Sensor abgewandte als auch dem Sensor zugewandte Personen erfasst.

Zur Evaluation und zu Erstellung der Trainingsdaten wurden aus dem RGBD-Stream einzelne Stichproben der gesamten Punktewolke gezogen. Mit Hilfe eines Annotationswerkzeuges, beschrieben in Anhang A, wurde die Punktewolke annotiert. Zu jeder Punktewolke ist die Position jeder sichtbaren Person und die Ausrichtung der Person bekannt. Die annotierten Daten wurden in zwei Datensätze, Datensatz A und Datensatz B, getrennt.

Insgesamt gibt es mehrere Möglichkeiten die ermittelten Merkmale zur Klassifikation zu verwenden. Eine Möglichkeit besteht darin, das Breiten- und Reliefmerkmal zu kombinieren, indem die Merkmalsvektoren zu einem Merkmalsvektor zusammengefasst werden. Mit dem kombinierten Merkmal wird eine SVM trainiert um anschließend die Objekte zu klassifizieren. Das Ergebnis dieser Vorgehensweise ist, dass viele Objekte fälschlicherweise als Person klassifiziert wurden.

Aus diesem Grund wurde der Test wiederholt und nur das Breitenmerkmal zur Klassifikation verwendet. In diesem Fall wurden ebenfalls viele Objekte fälschlicherweise als Person klassifiziert. Aus Abschnitt 3.4.2 zur Klassifikation ist bekannt, dass es möglich ist den Merkmalsraum mit Hilfe des ersten Werts des Merkmalsvektors zu visualisieren. Abbildung 5.1 zeigt den Merkmalsraum des Brei-

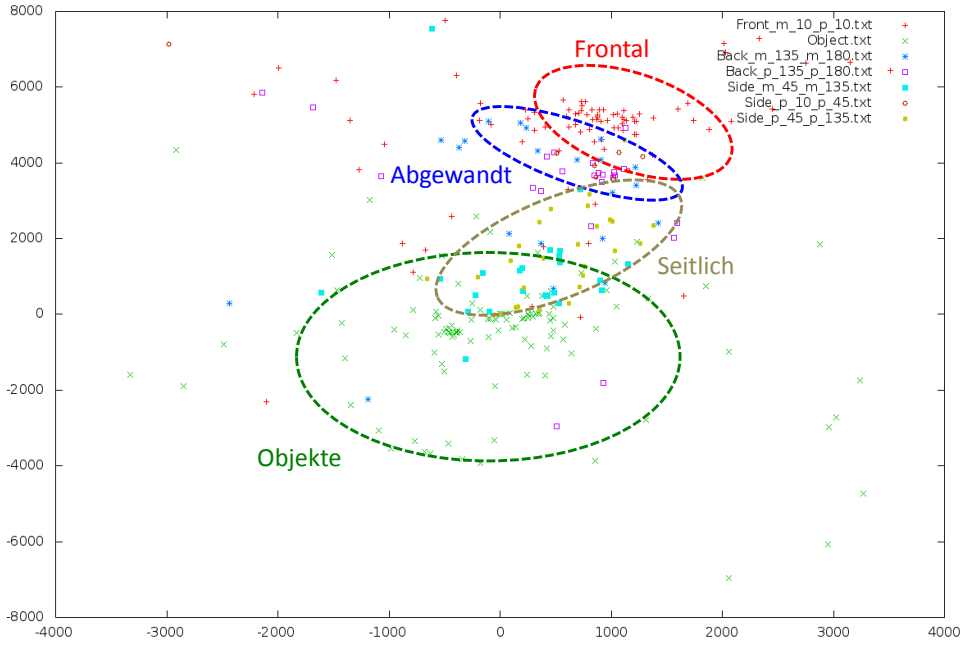


Abbildung 5.1: Darstellung des Merkmalsraums des Breitenmerkmals.

tenmerkmals. Die Aufgabe der SVM ist es diesen Merkmalsraum einzuteilen um Objekte von Personen zu trennen. Zur besseren Darstellung der zu klassifizierenden Bereiche wurden diese farbig umkreist. In Grün dargestellt ist der Bereich, in dem sich verstärkt Breitenmerkmale eines Objektes befinden. Im rot umrandeten Bereich befinden sich die Merkmale einer frontal zum Sensor stehenden Person. Merkmale einer abgewandten Person befinden sich im blau markierten Bereich. Der verbleibende Bereich, in Braun umrandet, stellt die Merkmale seitlich zum Sensor stehender Personen dar.

Anhand dieser Abbildung sind die Stärken und Schwächen des Breitenmerkmals zu erkennen. Das Breitenmerkmal ist gut geeignet um vollständig frontal bzw. vollständig abgewandte Personen von Objekten zu trennen. Sobald die Person leicht seitlich zum Sensor steht, ist das Merkmal nicht mehr aussagekräftig genug um Personen gegenüber von Objekten abzugrenzen.

Das Reliefmerkmal ist durch dieses Problem nicht betroffen. Steht eine Person beispielsweise um 45° gedreht beeinflusst dies nicht das Reliefmerkmal. Zu erkennen ist dies in Abbildung 5.2, in der der Schulterbereich nicht durch das Relief erfasst wird. Eine leichte Drehung hat somit keine Auswirkung auf das Relief.

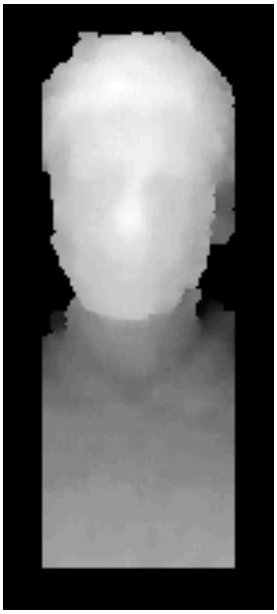


Abbildung 5.2: Tiefenbereich aus dem die Reliefinformation extrahiert wird.

Es wurden weitere Kombinationsmöglichkeiten der Verwendungen beider Merkmale getestet. Beispielsweise die Anwendung mehrerer SVMs, die nacheinander mit Hilfe des Breiten- und Reliefmerkmals klassifizieren. Das Ergebnis der Klassifikation wies einen höheren Anteil an falsch klassifizierten Personen auf als die alleinige Anwendung des Reliefmerkmals zur Klassifikation.

Aus diesem Grund wurde zu Klassifikation der Person nur das Reliefmerkmal eingesetzt. Für das Reliefmerkmal wurden dazu 3 Klassen erstellt. Die erste Klasse enthält alle Personen die zugewandt zum Sensor stehen. Das Relief wird über den Kinn- und Halsbereich der Person gebildet. Zu dieser zugewandten Klasse zählen zudem Personen, die z.B. 45° zum Sensor gedreht sind. Sobald das Relief durch den Schulterbereich der Person gebildet wird gilt die Person als abgewandt zum Sensor. In diesem Falle steht die Person im Winkel von 90° zum Sensor. Wird das Relief über den Rückenbereich gebildet zählt dies ebenfalls als abgewandte Person.

Eine Person gilt als detektiert, wenn das Merkmal in der Klasse der zugewandten oder abgewandten Person klassifiziert wurde.

Zur Klassifikation wurde das Reliefmerkmal mit einer Merkmalslänge von 5 eingesetzt. In Tabelle 5.1 sind die weiteren gewählten Schwellwerte für die Evaluation aufgelistet.

Zur Bewertung des Klassifikators wurde die Genauigkeit (engl. precision) und die Trefferquote (engl. recall) berechnet. Insgesamt lässt sich das Ergebnis einer Klassifikation in 4 Werte fassen:

Parameter	Wert
$\theta_{halbeKopfhoehe}$	0.18m
$\theta_{minKopfbreite}$	0.15m
$\theta_{maxKopfbreite}$	0.35m
$\theta_{BoxLinks}$	30%
$\theta_{BoxRechts}$	30%
$\theta_{BoxOben}$	10%
$\theta_{BoxUnten}$	20%
$\theta_{Merkmalslaenge}$	5

Tabelle 5.1: Parameter der Personendetektion

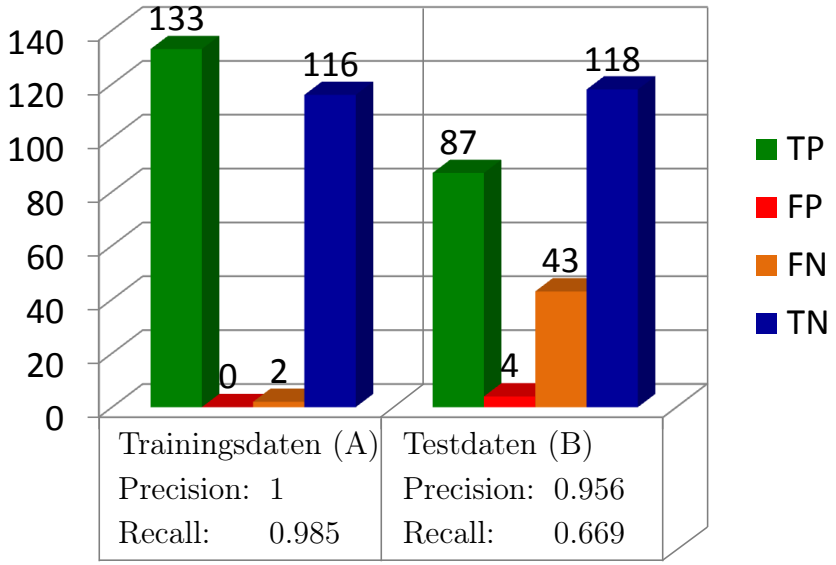
- Der Wert tp (engl. true positive) steht für die Anzahl der Personen, die als Person klassifiziert wurden.
- Der Wert fp (engl. false positive) steht für die Anzahl der Objekte, die als Person klassifiziert wurden.
- Der Wert fn (engl. false negative) steht für die Anzahl der Personen, die als Objekt klassifiziert wurden.
- Der Wert tn (engl. true negative) steht für die Anzahl der Objekte, die als Objekt klassifiziert wurden.

Zur Berechnung der Genauigkeit und der Trefferquote werden folgende Formeln (5.1 und 5.2) verwendet [TS09]:

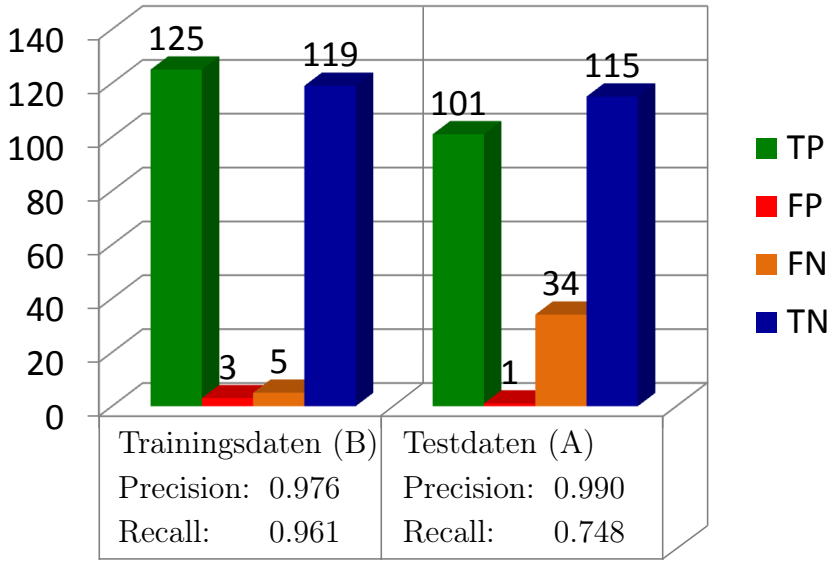
$$precision = \frac{tp}{tp + fp} \tag{5.1}$$

$$recall = \frac{tp}{tp + fn} \tag{5.2}$$

Das Ergebnis der Evaluation des Klassifikators ist in Abbildung 5.3 dargestellt. Abbildung 5.3a zeigt das Evaluationsergebnis, indem die SVM mit Datensatz A trainiert wurde. Im linken Bereich ist das Evaluationsergebnis auf den Trainingsdaten und im rechten Bereich das Evaluationsergebnis auf den Testdaten zu sehen. Abbildung 5.3b zeigt das Ergebnis wenn auf Datensatz B trainiert und auf Datensatz A getestet wurde. Anhand der Evaluation ist zu erkennen, dass nur wenige Objekte fälschlicherweise als Person klassifiziert wurden. Somit wird eine hohe Genauigkeit der Klassifikation von über 95% erreicht. Aufgrund der Anzahl der Personen die fälschlicherweise als Objekt klassifiziert wurden ist die Trefferquote geringer und liegt zwischen 66% bis 74%.

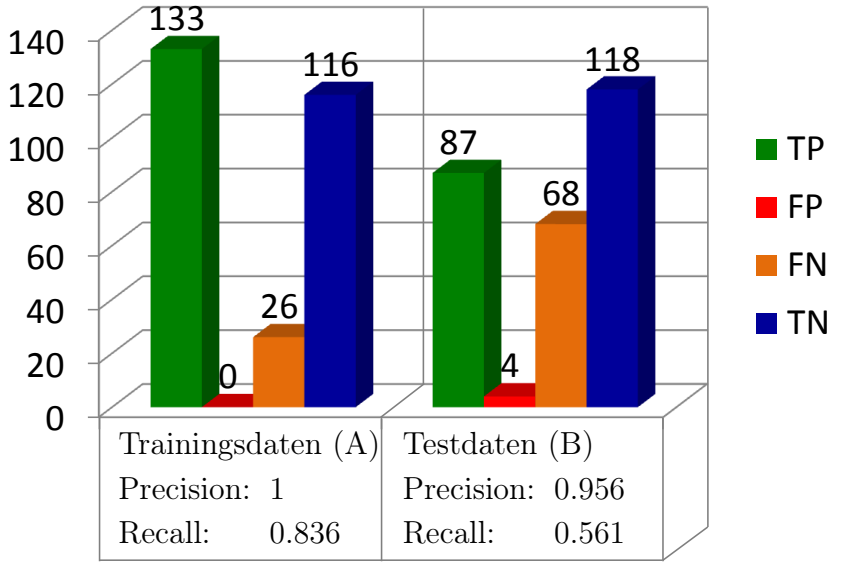


(a)

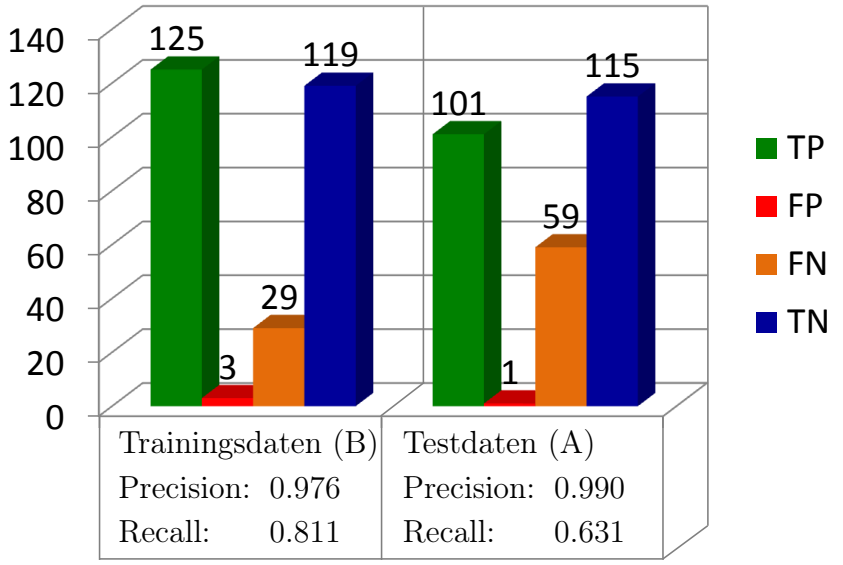


(b)

Abbildung 5.3: Evaluation des Klassifikators



(a)



(b)

Abbildung 5.4: Gesamtevaluation des Klassifikators mit der Kandidatensuche

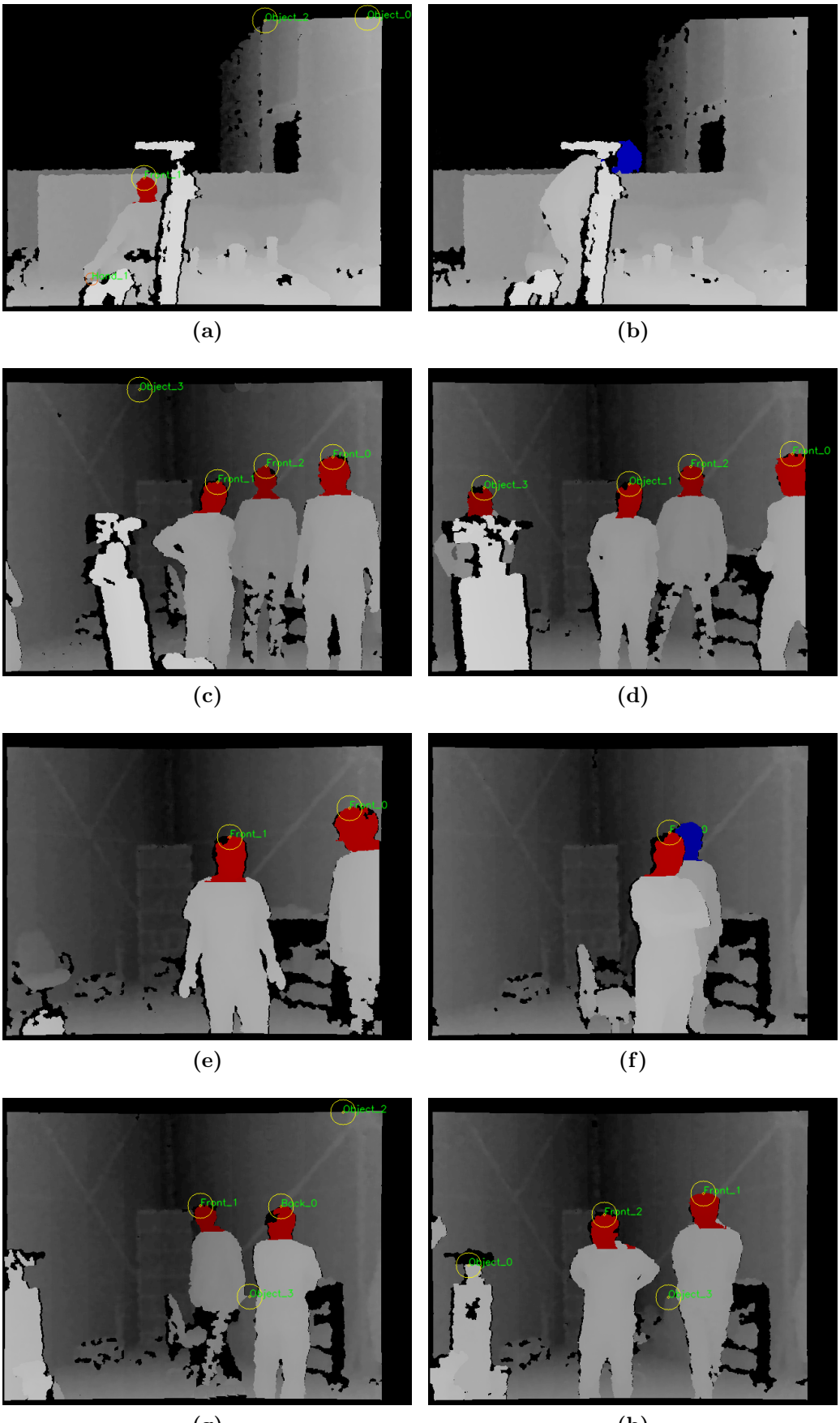


Abbildung 5.5: Ergebnis der Personendetektion.

Eine der Gründe, weshalb eine Person als Objekt klassifiziert wird, ist das die Person teilweise verdeckt ist. Zu sehen ist dies in Abbildung 5.5d, in der die Person am linken Rand durch ein Objekt im Vordergrund verdeckt wird. Durch das Objekt wird das Reliefmerkmal nicht korrekt bestimmt und eine Klassifikation schlägt fehl.

Eine zu starke Verdeckung der Person sorgt wie in Abbildung 5.5b und 5.5f dafür, dass eine Person durch den Kandidatenfinder nicht detektiert wird. In der Abbildung sind diese Fälle blau markiert.

Personen, die während der Kandidatensuche nicht als mögliche Kandidaten markiert werden, werden nicht klassifiziert. Zur Evaluation des gesamten Systems muss deshalb der Verlust der nicht detektieren Personen durch die Kandidatensuche hinzuaddiert werden. Das Ergebnis der Evaluation des Gesamtsystems ist in Abbildung 5.4 dargestellt.

Neben dem reinen Klassifikationsergebnis, ob es sich um eine Person oder ein Objekt handelt, wird zudem bestimmt ob eine Person zugewandt oder abgewandt zum Sensor steht (Abbildung 5.5g und 5.5h).

Die durchschnittliche Zeit für die Ermittlung der Kandidaten und anschließende Klassifikation beträgt 1,1 Sekunden[1].

5.2 Handdetektion

Im Anschluss an die Detektion der Person wird die Detektion der Hände durchgeführt. Die Detektion der Hände wird nur durchgeführt, wenn eine Person zugewandt zum Sensor detektiert wird. Zur Evaluation wurde ein Datensatz erstellt, indem insgesamt 1065 Hände sichtbar waren.

Es wurden unterschiedliche Handpositionen aufgenommen. Die Hand kann verschiedene Formen annehmen. Die genaue Form der Hand hat wenig Einfluss auf die Erkennung der Hand. Entscheidend ist die Fläche der hautfarbenen Bereiche der Hand, die mit dem Sensor erfasst wird. Eine flache Hand bildet z.B. eine geringere Fläche wenn nur die Handkante der Hand sichtbar ist. Wie in Abbildung 5.8 zu sehen, ist eine Detektion dennoch möglich.

Um die Erkennung der Hautfarbe zu erschweren wurde unterschiedliche Kleidung getragen. Kleidung mit beispielsweise blauer oder grüner Farbe unterscheidet sich deutlich von der Hautfarbe und stellt kein Problem dar. Aus diesem Grund wurde unter anderem Kleidung in orange- bzw. brauntönen getragen. Wie in Abbildung 5.7a zu sehen, können die Hände anhand der Farbe korrekt detektiert werden.

Die Erkennung der Hautfarbe kann fehlschlagen (siehe Abbildung 5.6). In diesem Falle ist das Gesicht leicht abgeschattet und die Gesichtsfarbe entspricht eher

[1] Verwendet wurde ein Intel Core 2 Duo T7500 Prozessor mit 2.2GHz

einem Grauton. Die ermittelten Farben des Gesichtes entsprechen eher den Farben der Kleidung als den Farben der besser beleuchteten Hände (Abbildung 5.6b). Ein solcher Fehlerfall wird durch das System erkannt, da die ermittelte, vermeintlich hautfarbene Fläche im Verhältnis zur erwarteten Fläche der Hand zu groß ist. Die Hautfarbenerkennung wird deaktiviert und die Hand wird über die Entfernung zum Körper bestimmt (Abbildung 5.6c).

Für die Evaluation wurden die Hände in 3 Kategorien eingeteilt. Zu den Händen welche nahe am Körper anliegen zählen Hände, welche ca. 10cm vom Körper entfernt sind. Hände, die ca. zwischen 10cm und 50cm vom Körper entfernt sind, besitzen eine mittlere Entfernung zum Körper. Hände, die einen Abstand von über 50cm vom Körper aufweisen, gehören zu den weit ausgestreckten Händen. Das Evaluationsergebnis ist in Tabelle 5.3 abgebildet. Die verwendeten Parameter für die Evaluation sind in Tabelle 5.2 aufgezählt.

Parameter	Wert
$\theta_{disNachbar}$	0.5m
$\theta_{disKopf}$	1.5m
θ_{Kopf}	0.28m
θ_{KopfG}	0.4m
N^2	100
θ_{Histo}	65%
$\theta_{Hautanteil}$	50%

Tabelle 5.2: Parameter der Handdetektion

	nahe am Körper	mittlere Entfernung	weit ausgestreckt
gefunden	136	331	253
nicht gefunden	177	155	13
Verhältnis	43%	68%	95%

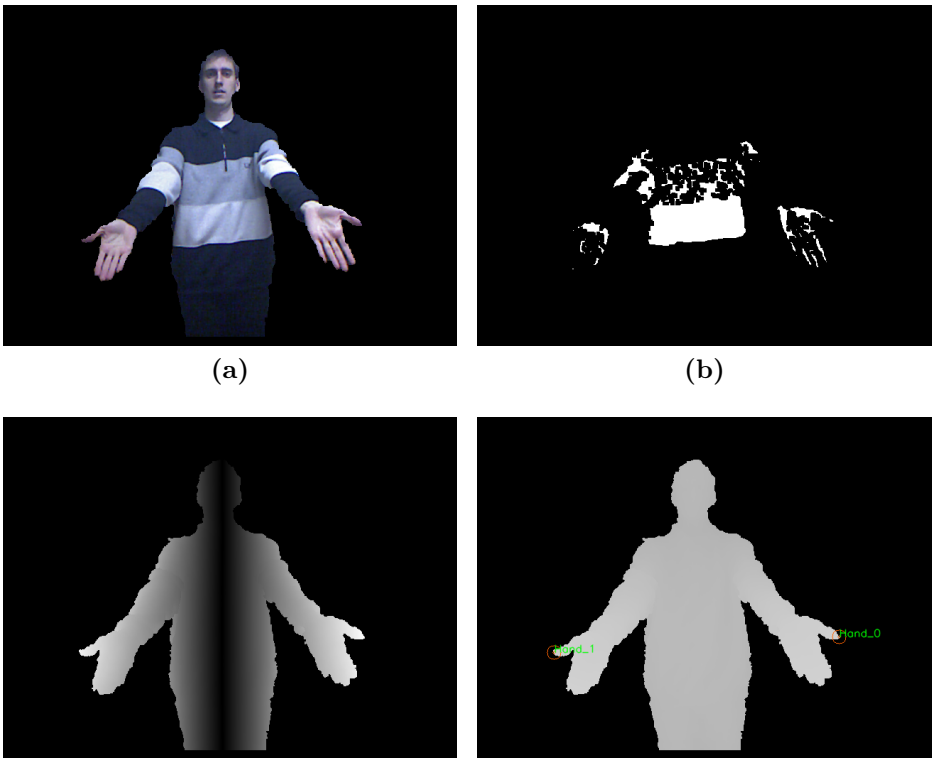
Tabelle 5.3: Ergebnis der Handdetektion bei der Suche nach 2 Händen einer Person.

Anhand der Tabelle 5.3 ist zu erkennen, dass eine Hand besser erkannt wird je weiter sie vom Körper entfernt ist. Von insgesamt 313 nahe anliegenden Händen wurden 136 Hände gefunden. Dies entspricht einem Prozentsatz der gefundenen Hände von 43%. Auf mittlerer Entfernung zum Körper wurden 68% der Hände detektiert. Am besten wurden mit 95% weit ausgestreckte Hände detektiert.

Der Grund für den geringen Anteil der gefundenen nahe anliegenden Hände liegt daran, dass häufig zwei Hände der Person zu sehen sind. Durch die Vorgehensweise bei der Ermittlung der Hand werden die Hände anhand der Farbe und der Entfernung zum Körper bewertet. Die Hände werden durch die lokalen Maxima der Bewertungsmatrix ermittelt. Das Maximum der Bewertungsmatrix stellt die erste gefundene Hand dar und das zweite lokale Maximum, welches sich nicht in der Nachbarschaft zum Maximum befindet, die zweite Hand. Durch diese Vorgehensweise wird häufig die zweite zu ermittelnde Hand nicht gefunden.

Zum Vergleich wurde derselbe Test erneut ausgeführt. Im Gegensatz zum vorherigen Test wurde nur eine Hand pro Person ermittelt. Diese eine Hand der Person wurde in 92% der Fälle detektiert. Hält eine Person eine Hand in z.B. mittlerer Entfernung und eine Hand in naher Entfernung wird die Hand in mittlerer Entfernung bevorzugt erkannt.

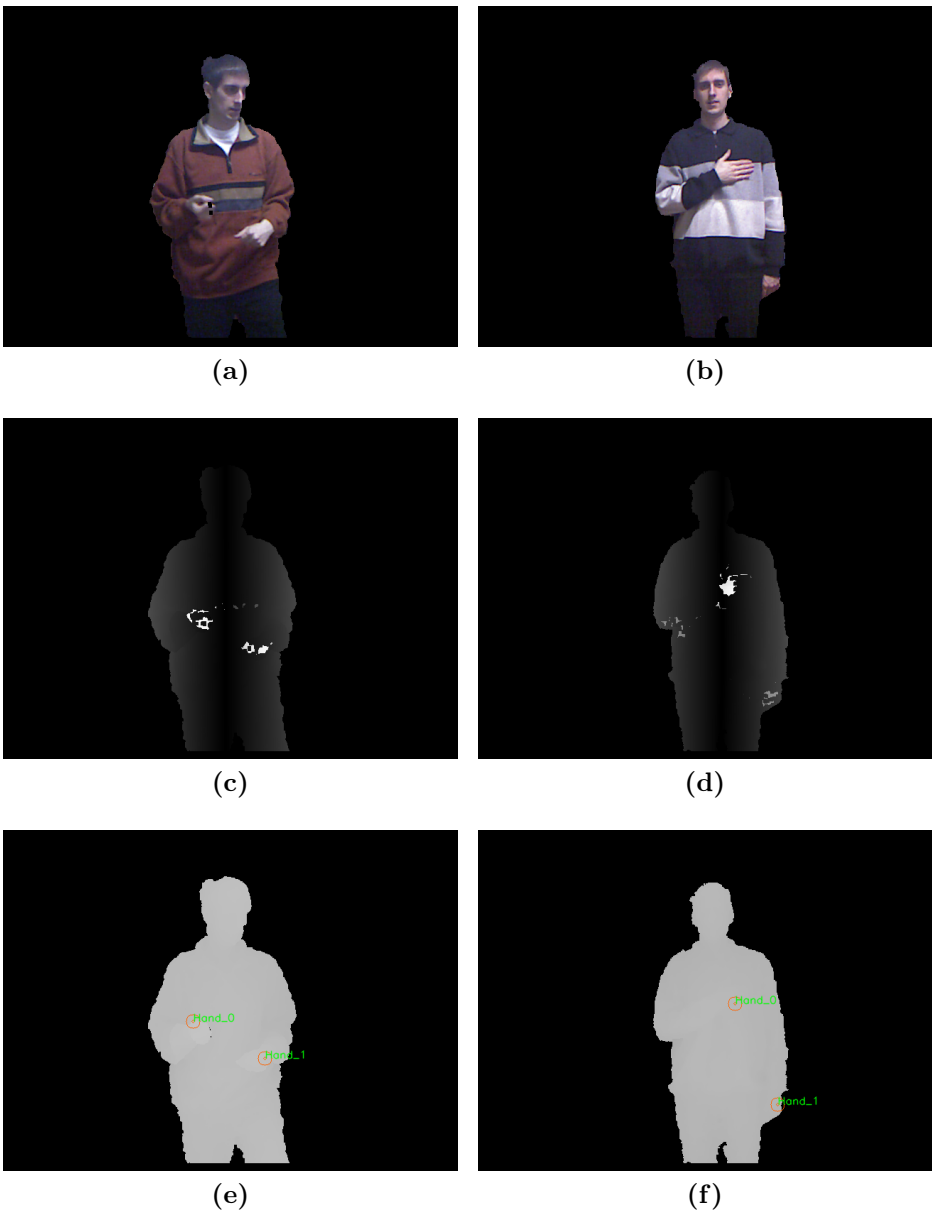
Das Ergebnis zeigt, dass es möglich ist beide Hände einer Person zu detektieren. Die Detektion wird besser, je weiter die Person die Hand vom Körper ausstreckt. Die Eigenschaft der Entfernung unterstützt das Finden der aussagekräftigsten Hand.



(a) **(b)**

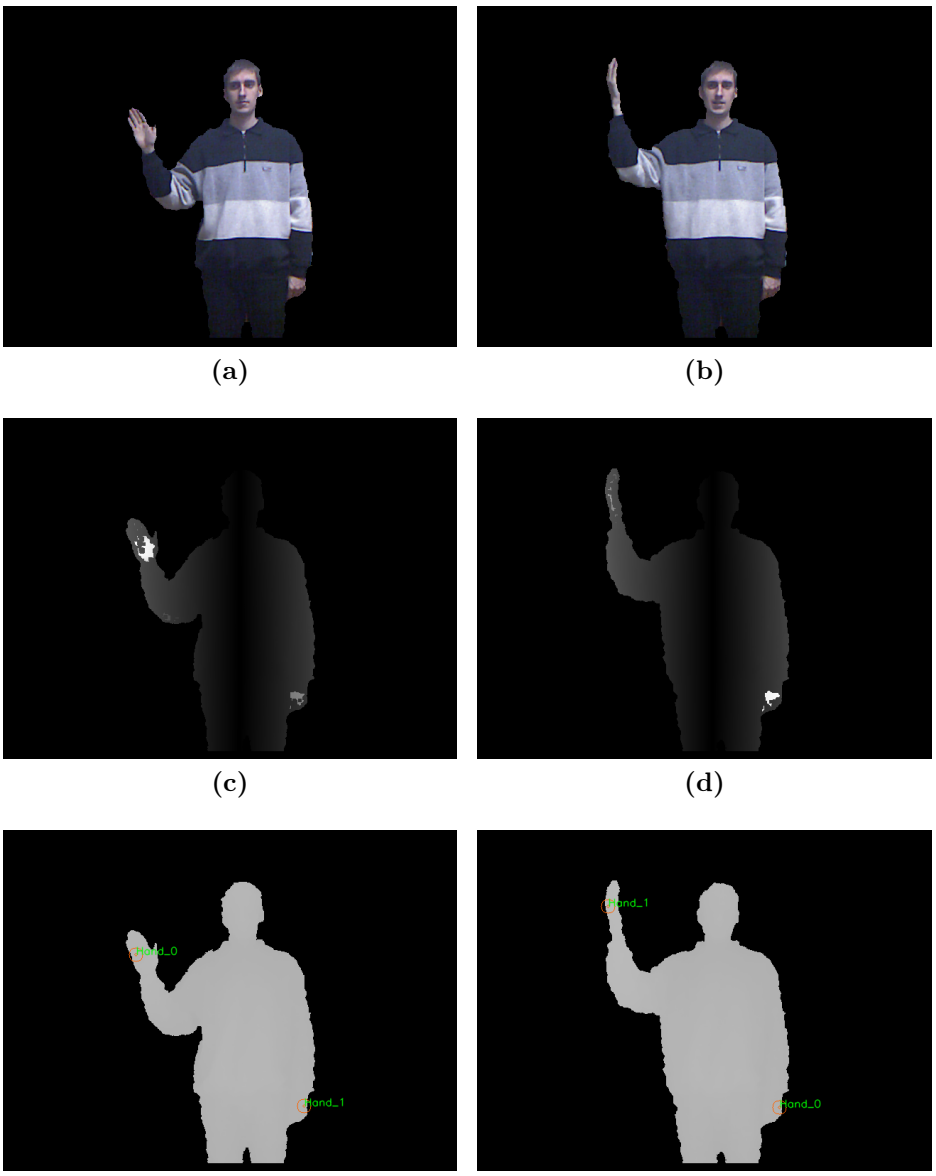
(c) **(d)**

Abbildung 5.6: Versagen der Hautfarbendetektion



(a) (b) (c) (d) (e) (f)

Abbildung 5.7: Beispiele für nahe am Körper liegende Hände



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Abbildung 5.8: Beispiel des Einflusses der Perspektive auf die Größe der Hand.

Kapitel 6

Fazit

In diesem Kapitel wird im ersten Abschnitt der Inhalt der Arbeit kurz zusammengefasst und das Ergebnis der Evaluation bewertet. Der zweite Abschnitt gibt einen kurzen Ausblick.

6.1 Zusammenfassung

Das gesamte entwickelte System zur Personen- und Handdetektion lässt sich anhand von Abbildung 6.1 zusammenfassen. Im ersten Schritt wurde eine Kandidatensuche auf der Punktewolke durchgeführt. Die Punktewolke wurde in Schichten zerlegt und innerhalb jeder Schicht mit Hilfe eines einfachen Personenmodells mögliche Kandidaten einer Person ermittelt. Die Kandidatensuche ermöglicht es den Suchraum stark einzugrenzen. Anstelle alle Punkte innerhalb der Punktewolke auf mögliche Personen hin zu untersuchen werden nur wenige Personenkandidaten erzeugt.

Auf Basis dieser Kandidaten wird mit dem neu entwickelten Reliefmerkmal das Relief einer möglichen Person bestimmt. Bei der Erstellung des Merkmals wird besonders auf die Robustheit des Merkmals Wert gelegt. Durch die Bildung des Mittelwertes einzelner Zeilen konnte so ein Merkmal erzeugt werden, welches nicht durch leichte Bewegungen des Kopfes beinträchtig wird.

Neben dem Reliefmerkmal wurde auch das Breitenmerkmal entwickelt, welches die Breite der Person erfasst. Es hat sich herausgestellt, dass dieses Merkmal schlechter zur Personendetektion geeignet ist, da vor allem seitlich stehende Personen schlecht erkannt werden.

Mit Hilfe einer Diskreten Fourier-Transformation wird das Reliefmerkmal für eine Klassifikation aufbereitet. Durch die Nutzung der Fourier-Transformation wird es ermöglicht die Merkmalslänge zu verkürzen, da im Frequenzraum eine Ordnung

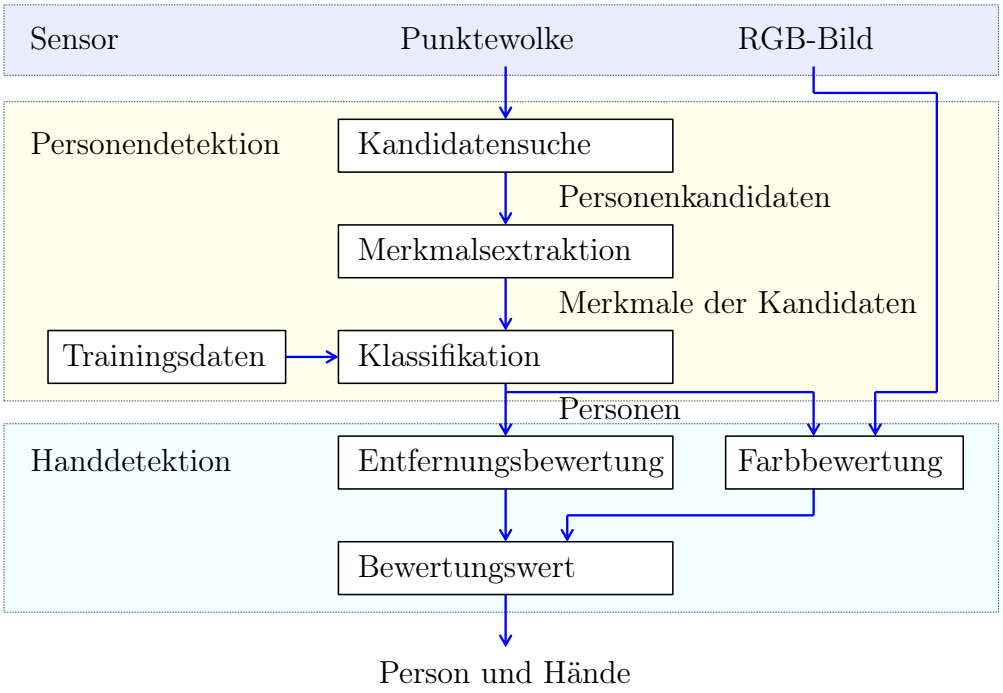


Abbildung 6.1: Systemüberblick

nach Frequenzen vorliegt. Zugleich wird durch die Entfernung des Gleichanteils eine Normalisierung des Reliefmerkmals erreicht.

Unter Verwendung eines Annotationswerkzeugs wurden Testdaten annotiert und Merkmale extrahiert, die zum Training einer nicht linearen SVM verwendet werden. Die SVM wurde verwendet um Personen zu klassifizieren. Die Klassifikation ermöglicht dabei auch die Erkennung, ob eine Person abgewandt oder zugewandt zum Sensor steht.

Auf den als zugewandt klassifizierten Personen wird anschließend eine Detektion der Hand durchgeführt. Es wird hierzu eine Bewertung anhand der Entfernung der Hand zum Körper und der Farbe der Hand erstellt. Sowohl Farbe als auch Entfernung werden kombiniert um anschließend die Hände zu detektieren.

Die Personendetektion wurde entwickelt und anhand der Parameter so eingestellt um eine möglichst hohe Genauigkeit zu erzielen. Bereits vorhandene Systeme, die beispielsweise die Beine einer Person über Laserscanner detektieren, weisen in diesem Bereich Schwächen auf. Wird ein Objekt fälschlicherweise als Person detektiert schlagen auf der Personendetektion aufbauende Aufgaben fehl. Aus diesem Grund wird bevorzugt eine Person nicht zu detektieren als fälschlicherweise ein Objekt als Person aufzufassen. In der praktischen Anwendung nimmt der Sensor über die Zeit immer neue Punktewolken der Szene auf und stellt sie dem Algorith-

mus zur Klassifikation zur Verfügung. Wird eine Person somit in einem einzelnen Frame nicht erkannt bedeutet dies nicht, dass die Person nie gefunden wird. Ist eine Person beispielsweise durch eine andere Person zu stark verdeckt und bewegt sich eine Person, kann die verdeckte Person, die zuvor nicht detektiert wurde, detektiert werden. Durch die hohe Genauigkeit von über 95% wird sichergestellt, dass es sich bei einer gefundenen Person auch um eine Person handelt. Es können somit weitere Aufgaben, die auf der Personendetektion aufbauen, durchgeführt werden.

Zu diesen Aufgaben zählt die Detektion der Hand. In der praktischen Anwendung sind hierbei vor allem die Hände von Bedeutung, die sich vom Körper abheben. Bei der Kommunikation durch die Hand zeigt eine Person beispielsweise in eine Richtung oder streckt die Hand entgegen um einen Gegenstand entgegenzunehmen. Die Evaluation hat hierbei gezeigt, dass die Detektion der Hand umso besser funktioniert umso weiter die Hand vom Köper entfernt ist. Es wurde auch gezeigt, dass es mit Hilfe der Farbe möglich ist auch am Körper anliegende Hände zu detektieren. Im Gegensatz zu anderen Ansätzen, welche nur Hände detektieren die zwischen Sensor und Person steht, erkennt dieser Ansatz auch Hände, welche zur Seite ausgestreckt werden.

6.2 Ausblick

Die Personen- und Handdetektion stellt darüber hinaus auch einige Daten zur Verfügung, welche für anschließende Aufgaben nützlich sind. Hierzu zählt auch das freigestellte RGB-Bild der Person. Auf Basis dieses Bildes könnte eine Personenerkennung durchgeführt werden. Denkbar wäre eine Unterscheidung von Personen anhand der Farben oder eine Extraktion bestimmter Merkmale. Zu diesen Merkmalen zählen beispielsweise SIFT-Merkmale, die innerhalb der Person ermittelt werden oder Merkmale einer Gesichtserkennung auf dem Gesichtsbereich der Person. Mit Hilfe der Information des höchsten Punktes des Kopfes ist neben der Ermittlung der Haarfarbe auch die Bestimmung der Größe realisierbar.

Vorstellbar wäre zudem ein Filtersystem auf Basis der erhobenen Daten, welches komplexere Anfragen bearbeitet. Eine solche Anfrage könnte z.B. sein die Person zu bestimmen, welche eine Hand aufweist deren 3D-Koordinate über der 3D-Koordinate des höchsten Punktes des Kopfes liegt. Auf diese Weise werden die aufzeigenden Personen bestimmt.

Die ermittelte Entfernung der Hand zum Köper kann zudem als Maß für die Wichtigkeit der Hand verwendet werden. Besteht die Aufgabe beispielsweise darin ein Objekt an eine Person zu übergeben, wird das Objekt an die Position der Hand übergeben, welche am weitesten vom Körper entfernt ist.

Darüber hinaus kann die Handdetektion als initiale Detektion für weitere Aufgaben dienen. Eine einmal detektierte Hand könnte mit Hilfe eines Algorithmus

verfolgt und so die Gesten einer Hand erfasst werden. Die Form der detektierten Hand könnte weiter ausgewertet werden. Mit Hilfe der relativen Position der Hand zum Kopf ist eine Bestimmung der Richtung der Hand, in der eine Person zeigt, erreichbar.

Auch die Personendetektion könnte zur Verfolgung der Person eingesetzt werden. Berücksichtigt werden könnte auch die Ausrichtung der Person. Abgewandte Personen bewegen sich wahrscheinlich vom Sensor weg.

Neben diesen auf der Personendetektion aufbauenden Verfahren könnte das entwickelte Reliefmerkmal auch zur Klassifikation anderer Objekte eingesetzt werden. Beispielsweise indem das Relief unterschiedlicher Möbelstücke erfasst wird um diese zu klassifizieren.

Anhang A

Annotationswerkzeug

In diesem Kapitel wird das entwickelte Annotationswerkzeug und seine Verwendung beschrieben. Für die Erstellung von Trainingsdaten und zur Evaluation der Ergebnisse ist es von Vorteil die Punktewolke zu annotieren.

Aus dem RGBD-Stream der Kinect werden in regelmäßigen Abständen Stichproben gezogen. Die RGBD-Punktewolke wird mit Hilfe der *Point Cloud Library* als pcd-Datei gespeichert. Eine pcd-Datei enthält die gesamte Punktewolke und die RGB-Daten der Kamera. Die einzelnen pcd-Dateien werden mit einem Zeitstempel versehen und in einem Ordner gespeichert. Wird das Annotationswerkzeug aufgerufen, wird die erste pcd-Datei innerhalb des Ordners geöffnet und angezeigt. Der Inhalt der Punktewolke wird über Dreitafelprojektion dargestellt.

Für die Dreitafelprojektion wird die Punktewolke jeweils auf ihre Achsen projiziert. Insgesamt entstehen somit 3 unterschiedliche Ansichten (siehe Abbildung A.1a). Oben links zu sehen ist die Vorderansicht, diese entsteht durch die Projektion aller z-Koordinaten auf die $y, x-$Ebene. Innerhalb der Ansicht werden zur besseren Orientierung die $z-$Werte auf Grauwerte abgebildet. Die Ansicht rechts daneben stellt die Seitenansicht dar. Die $x-$Werte wurden auf die $y, z-$Ebene abgebildet. Rechts unten wird die Draufsicht durch die Projektion auf die $z, x-$Ebene dargestellt.

Durch Definition von 6 projizierenden Geraden ist es möglich eine Bounding-Box innerhalb der 3D-Punktewolke zu definieren. Diese Geraden sind in Abbildung A.1b vergrößert abgebildet. Mit Hilfe der 6 Schieberegler lassen sich die Geraden und somit die Bounding-Box einstellen. Zur besseren Kontrolle des gewählten Bereiches wird in der Tiefenansicht der markierte Bereich rot eingezeichnet. Durch einen Rechtsklick auf Tiefenansicht unten rechts wird zwischen der Tiefenansicht und dem RGB-Bild der Kamera gewechselt.

Der beispielhafte Ablauf einer Annotation kann wie folgt aussehen:

1. Der *newPerson*-Button wird geklickt um eine neue Person zu erstellen.

2. Die Bounding-Box der Person wird ausgewählt.
3. Mit dem *addPerson*-Button wird die Person zur Liste der annotierten Personen hinzugefügt.
4. Der *newHead*-Button wird gedrückt um den Kopf für die Person zu erstellen.
5. Die Bounding-Box des Kopfes wird ausgewählt.
6. Über en *addHead*-Button wird der Kopf der Peson hinzugefügt.
7. Die Bounding-Box des Kopfes wird ausgewählt.
8. Der *newHand*-Button wird gedrückt um die Hand für eine Person zu erstellen.
9. Die Bounding-Box der linken Hand wird ausgewählt.
10. Über den *addHandL*-Button wird die linke Hand der Person hinzugefügt.

Zusätzlich kann anhand der Draufsicht die Ausrichtung (der Gier-Winkel) der Person annotiert werden. Der Standpunkt der Person wird durch den Mittelpunkt der Boundig-Box bestimmt und konstant gehalten. Die Ausrichtung wird über die Angabe der Blickrichtung durch einen Klick auf die Draufsicht bestimmt (Abbildung A.3). Durch einen Klick auf den *addRotation*-Button wird die Ausrichtung der Person hinzugefügt.

Während die Annotation durchgeführt wird, werden die annotierten Bereiche einer Baumstruktur im unteren rechten Bereich hinzugefügt (Abbildung A.4b). Der *clear*-Button löscht alle Elemente der Baumstruktur. Der *save*-Button speichert die Elemente in eine XML-Datei, welche bis auf die Dateiendung den selben Namen der pcd-Datei erhält. Existiert zu einer gegebenen pcd-Datei eine xml-Datei, wird diese automatisch geladen. Die Elemente der Baumstruktur können mit der Maus ausgewählt werden. Wird ein Element ausgewählt, wird die Bounding-Box des Elementes angezeigt. Ein Element kann geändert werden, indem es ausgewählt wird und die entsprechende Bounding-Box neu gesetzt wird. Durch den Klick eines *add*-Buttons wird das entsprechende Element neu gesetzt.

Der *LoadNext*-Button und *LoadPrevious*-Button dient der Navigation. Soll ein Testdatensatz annotiert werden, kann so einzeln durch die Einzelbilder der Kinect navigiert werden. Für eine schnellere Navigation innerhalb der Einzelbilder ist der Schiebebalken am unteren Rand des Steuerungsbereiches angebracht. Der zweite Schiebebalken dient zur Einstellung des Maßstabes der Projektion.

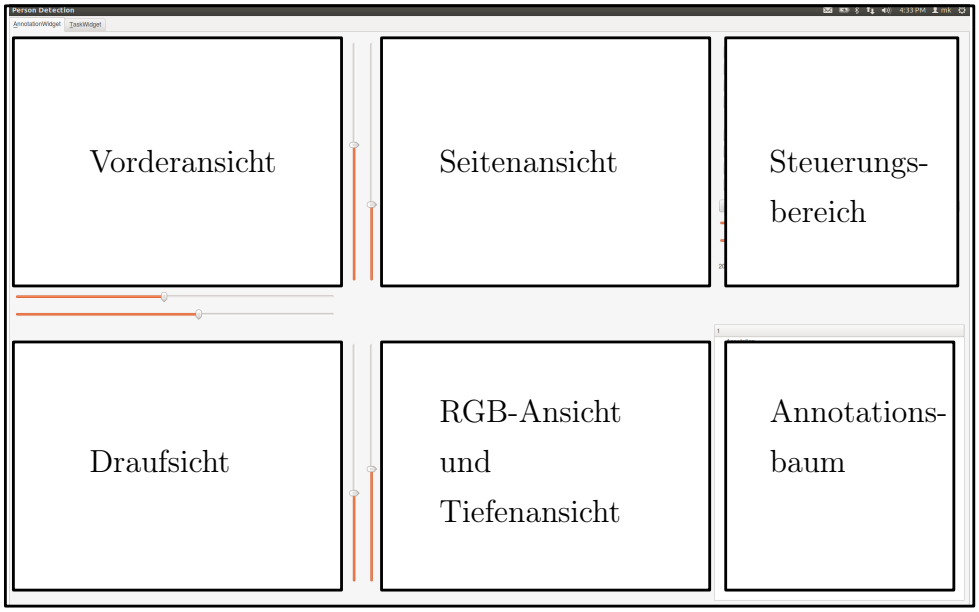
Der *guideMode*-Button führt eine zuvor festgelegte Reihenfolge von Ereignissen aus. In der praktischen Anwendung kommt es häufig vor, dass für jede Person der

Kopf, die Ausrichtung und die Hände der Person bestimmt werden. Anstelle einzeln immer wieder die selbe Reihenfolge von Befehlen auszuführen wird der Anwender durch die Annotation geleitet. Hierzu werden einzelne Anweisungen an den Nutzer direkt auf dem Button eingeblendet.

Der *nextPerson*-Button kann verwendet werden um bei mehreren annotierten Personen die nächste Person auszuwählen. Alternativ kann auch auf die Baumansicht geklickt werden.

Zur schnelleren Erstellung der Bounding-Box ist es möglich mit einem Linksklick auf der RGB-Ansicht bzw. Tiefenansicht eine grobe Bounding-Box zu setzen. Die Größe der geschätzten Bounding-Box häng von der Auswahl des neu hinzuzufügenden Elementes ab. Mit Hilfe der Schieberegler kann die Bounding-Box präziser eingestellt werden.

Das Annotationswerkzeug beschreibt die Szene, indem es eine Baumstruktur aufbaut und diese als XML-Datei speichert. Da es leicht erweiterbar ist, kann es auch für andere Anwendungsgebiete eingesetzt werden. Denkbar wäre eine Erweiterung, indem der Name der detektierten Person vermerkt wird. Auf diese Weise könnte eine Personenerkennung evaluiert werden.



(a) Aufbau des Annotationswerkzeugs

(b) Projizierende Geraden zur Einstellung der Bounding-Box

Abbildung A.1: Aufbau und Funktion des Annotationswerkzeugs.

Abbildung A.2: Steuerungsbereich des Annotationswerkzeugs

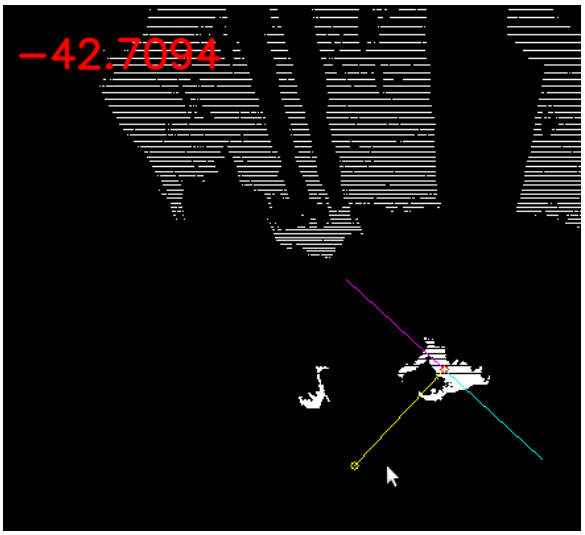
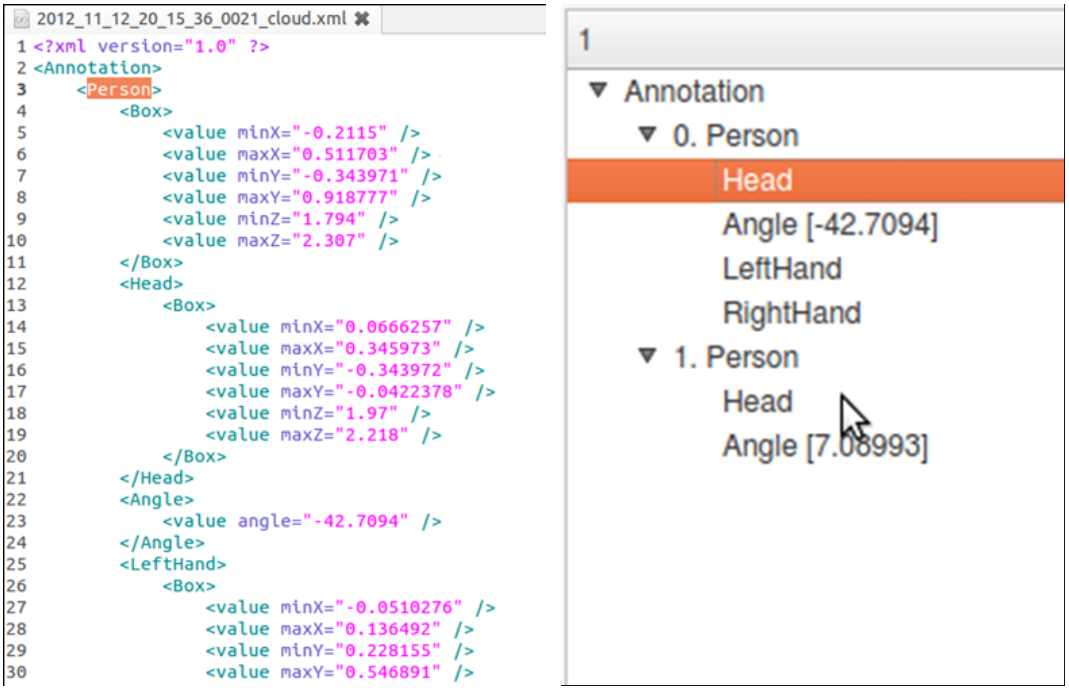


Abbildung A.3: Annotation des Gierwinkels der Person



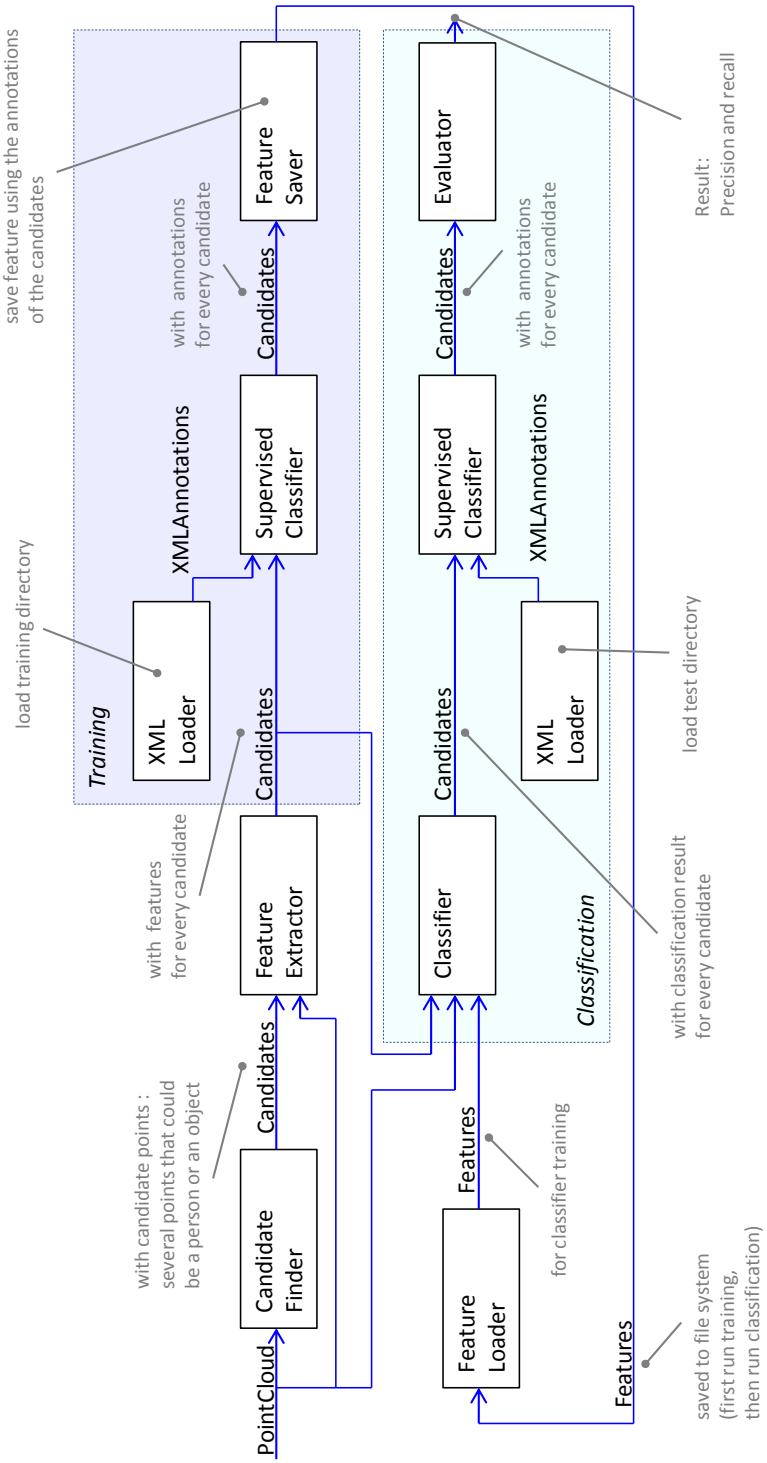
(a) XML-Datei **(b)** Annotationsbaum

Abbildung A.4: Darstellung der Annotation als Baumstruktur.

Anhang B

Systemübersicht

In Abbildung B.1 ist eine ausführlichere Systemübersicht der Personendetektion abgebildet. Es wird der Ablaufplan des Systems dargestellt. Die Übersicht ist in Englisch angegeben, da sich die entsprechenden Benennungen auch in der Implementierung wiederfinden.



(a)

Abbildung B.1: Systemübersicht

Literaturverzeichnis

[AL92] Ai, Wayne I. ; Langley, Pat: Induction of One-Level Decision Trees. In: *Proceedings of the Ninth International Conference on Machine Learning*, Morgan Kaufmann, 1992, S. 233–240

[AMB07] Arras, Kai O. ; Martínez Óscar ; Burgard, Mozos W.: Using Boosted Features for Detection of People in 2D Range Scans. In: *In Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2007

[BB10] Bernard, Arnaud ; Bing, Benny: Automatic Hand Reference Acquisition Using a Stereo 3D Webcam. In: *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*. New York, NY, USA : ACM, 2010 (ICDSC '10), S. 232–233

[BBCT05] Bennewitz, Maren ; Burgard, Wolfram ; Cielniak, Grzegorz ; Thrun, Sebastian: Learning Motion Patterns of People for Compliant Robot Motion. In: *Internationl Journal of Robotics Research* 24 (2005), S. 31–48

[BBR+07] Bertozzi, M ; Broggi, A ; Rose, M. D. ; Felisa, M. ; Rakotomamonjy, A. ; Suard, F.: A Pedestrian Detector Using Histograms of Oriented Gradients and a Support Vector Machine Classifier. In: *Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference*. Seattle, WA, USA, 2007

[BMH+09] Bajracharya, Max ; Moghaddam, Baback ; Howard, Andrew ; Brennan, Shane ; Matthies, Larry H.: Results from a Real-time Stereo-based Pedestrian Detection System on a Moving Vehicle. In: *Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking*. Kobe (Japan), 2009

[BMP01] Belongie, Serge ; Malik, Jitendra ; Puzicha, Jan: *Matching Shapes*. 2001

[CAHS06] Coogan, Thomas ; Awad, George ; Han, Junwei ; Sutherland, Alistair: Real Time Hand Hesture Recognition Including Hand Segmentation and Tracking. In: *Proceedings of the Second international conference on Advances in Visual Computing - Volume Part I*. Berlin, Heidelberg : Springer-Verlag, 2006 (ISVC'06), S. 495–504

[CM02] Comaniciu, Dorin ; Meer, Peter: Mean shift: A Robust Approach Toward Feature Space Analysis. In: *In PAMI*, 2002, S. 603–619

[CPVC07] Cerlinca, Tudor I. ; Pentiuc, Stefan G. ; Vatavu, Radu D. ; Cerlinca, Marius C.: Hand Posture Recognition for Human-Robot Interaction. In: *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*. New York, NY, USA : ACM, 2007 (WMISI '07), S. 47–50

[CSP09] Choi, Junyeong ; Seo, Byung-Kuk ; Park, Jong-Il: Robust Hand Detection for Augmented Reality Interface. In: *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*. New York, NY, USA : ACM, 2009 (VRCAI '09), S. 319–321

[CXL+12] Chai, Xiujuan ; Xu, Zhihao ; Li, Qian ; Ma, Bingpeng ; Chen, Xilin: Robust Hand Tracking by Integrating Appearance, Location and Depth Cues. In: *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*. New York, NY, USA : ACM, 2012 (ICIMCS '12), S. 60–65

[DSM+11] Doliotis, Paul ; Stefan, Alexandra ; McMurrough, Christopher ; Eckhard, David ; Athitsos, Vassilis: Comparing Gesture Recognition Accuracy Using Color and Depth Information. In: *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA : ACM, 2011 (PETRA '11), S. 20:1–20:7

[DT05] Dalal, Navneet ; Triggs, Bill: Histograms of Oriented Gradients for Human Detection. In: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*. Washington, DC, USA : IEEE Computer Society, 2005 (CVPR '05), S. 886–893

[DWS+09] Dollar, Piotr ; Wojek, Christian ; Schiele, Bernt ; Perona, Pietro ; Darmstadt, Tu: Pedestrian detection: A benchmark. In: *In CVPR*, 2009

[FHM02] Fod, Ajo ; Howard, Andrew ; Mataric, Maja J.: Laser-Based People Tracking. In: *In Proc. of the IEEE International Conference on Robotics & Automation (ICRA*, 2002, S. 3024–3029

[FMR08] Felzenszwalb, Pedro ; Mcallester, David ; Ramanan, Deva: A Discriminatively Trained, Multiscale, Deformable Part Model. In: *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2008*, 2008

[FP09] Fosso, Fabio ; Porta, Marco: A Vision-based Attentive User Interface with (Semi-)Automatic Parameter Calibration. In: *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing.* New York, NY, USA : ACM, 2009 (CompSysTech '09), S. 33:1–33:6

[FSV07] Francke, Hardy ; Solar, Javier Ruiz-del ; Verschae, Rodrigo: Real-time Hand Gesture Detection and Recognition using Boosted Classifiers and Active Learning. In: *Proceedings of the 2nd Pacific Rim conference on Advances in image and video technology.* Berlin, Heidelberg : Springer-Verlag, 2007 (PSIVT'07), S. 533–547

[FXZ+12] Feng, Ziyong ; Xu, Shaojie ; Zhang, Xin ; Jin, Lianwen ; Ye, Zhichao ; Yang, Weixin: Real-time Fingertip Tracking and Detection using Kinect Depth Sensor for a new Writing-in-the Air System. In: *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service.* New York, NY, USA : ACM, 2012 (ICIMCS '12), S. 70–74

[GZC+10] Ghosh, Soumita ; Zheng, Jianmin ; Chen, Wenyu ; Zhang, Jane ; Cai, Yiyu: Real-time 3D Markerless Multiple Hand Detection and Tracking for Human Computer Interaction Applications. In: *Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry.* New York, NY, USA : ACM, 2010 (VRCAI '10), S. 323–330

[HHKP12] Hegger, Frederik ; Hochgeschwender, Nico ; Kraetzschmar, Gerhard K. ; Ploeger, Paul G.: People Detection in 3d Point Clouds using Local Surface Normals. Sankt Augustin (Germany), 2012

[HHRB11] Holz, Dirk ; Holzer, Stefan ; Rusu, Radu B. ; Behnke, Sven: Real-Time Plane Segmentation using RGB-D Cameras. In: *Proceedings of the 15th RoboCup International Symposium* Bd. 7416. Istanbul, Turkey : Springer, July 2011 (Lecture Notes in Computer Science), S. 307–317

[HK10] Hordern, Daniel ; Kirchner, Nathan: Robust and Efficient People Detection with 3-D Range Data using Shape Matching. Sydney (Australia), 2010

[IF11] Ikemura, Sho ; Fujiyoshi, Hironobu: Real-time Human Detection using Relational Depth Similarity Features. In: *Proceedings of the 10th Asian conference on Computer vision - Volume Part IV*. Berlin, Heidelberg : Springer-Verlag, 2011 (ACCV'10), S. 25–38

[IMKK04] Ikeda, Hitoshi ; Maeda, Masahiro ; Kato, Noriji ; Kashimura, Hirotsugu: Classification of Human Actions using Face and Hands Detection. In: *Proceedings of the 12th annual ACM international conference on Multimedia*. New York, NY, USA : ACM, 2004 (MULTIMEDIA '04), S. 484–487

[JR02] Jones, Michael J. ; Rehg, James M.: Statistical Color Models with Application to Skin Detection. In: *Int. J. Comput. Vision* 46 (2002), Januar, Nr. 1, S. 81–96. – ISSN 0920–5691

[LC09] Lee, Byungsung ; Chun, Junchul: Manipulation of Virtual Objects in Markerless AR System by Fingertip Tracking and Hand Gesture Recognition. In: *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. New York, NY, USA : ACM, 2009 (ICIS '09), S. 1110–1115

[LF04] Liu, Xia ; Fujimura, Kikuo: Hand Gesture Recognition using Depth Data. In: *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*. Washington, DC, USA : IEEE Computer Society, 2004 (FGR' 04), S. 529–534

[LLS04] Leibe, Bastian ; Leonardis, Ales ; Schiele, Bernt: Combined Object Categorization and Segmentation With An Implicit Shape Model. In: *In ECCV workshop on statistical learning in computer vision*, 2004, S. 17–32

[LSS05] Leibe, Bastian ; Seemann, Edgar ; Schiele, Bernt: Pedestrian Detection in Crowded Scenes. In: *In CVPR*, 2005, S. 878–885

[LZTS04] Lee, Dah-Jye ; Zhana, Pengcheng ; Thomasa, Aaron ; Schoenbergerb, Robert: Shape-based Human Detection for Threat Assessment. In: *Visual Information Processing XIII*, 2004

[MKH09] Mozos, Oscar M. ; Kurazume, Ryo ; Hasegawa, Tsutomu: Multi-Layer People Detection using 2D Range Data. In: *In Proceedings of the*

IEEE ICRA 2009 Workshop on People Detection and Tracking. Kobe (Japan), 2009

[MSZ04] Mikolajczyk, K ; Schmid, C ; Zisserman, A: Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In: *The 8th ECCV, pages 69-81*. Prague, Czech Republic, 2004

[Ots79] Otsu, N: A Threshold Selection Method from Gray-level Histograms. In: *IEEE Transactions on Systems, Man and Cybernetics, Vol. 9, No. 1.*, 1979, S. 62–66

[PN05] Premebida, Cristiano ; Nunes, Urbano: Segmentation and Geometric Primitives Extraction from 2D Laser Range Data For Mobile Robot Applications. In: *in Robotica 2005, Scientific meeting of the 5th National RoboticsFestival*, 2005

[RMYZ11] Ren, Zhou ; Meng, Jingjing ; Yuan, Junsong ; Zhang, Zhengyou: Robust Hand Gesture Recognition with Kinect Sensor. In: *Proceedings of the 19th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2011 (MM '11), S. 759–760

[RYZ11] Ren, Zhou ; Yuan, Junsong ; Zhang, Zhengyou: Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera. In: *Proceedings of the 19th ACM international conference on Multimedia*. New York, NY, USA : ACM, 2011 (MM '11), S. 1093–1096

[SA11] Spinello, L. ; Arras, K. O.: People Detection in RGB-D Data. In: *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2011

[SAAS08] Stefan, Alexandra ; Athitsos, Vassilis ; Alon, Jonathan ; Sclaroff, Stan: Translation and Scale-invariant Gesture Recognition in Complex Scenes. In: *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA : ACM, 2008 (PETRA '08), S. 7:1–7:8

[SATS10] Spinello, L. ; Arras, K. O. ; Triebel, R. ; Siegwart, R.: A Layered Approach to People Detection in 3D Range Data. In: *Proc. of The AAAI Conference on Artificial Intelligence: Physically Grounded AI Track (AAAI)*, 2010

[SBFC03] Schulz, Dirk ; Burgard, Wolfram ; Fox, Dieter ; Cremers, Armin B.: *People Tracking with a Mobile Robot Using Sample-Based Joint Probabilistic Data Association Filters.* 2003

[SBL11] Schiffer, Stefan ; Baumgartner, Tobias ; Lakemeyer, Gerhard: A Modular Approach to Gesture Recognition for Interaction with a Domestic Service Robot. In: *Proceedings of the 4th international conference on Intelligent Robotics and Applications - Volume Part II.* Berlin, Heidelberg : Springer-Verlag, 2011 (ICIRA'11), S. 348–357

[SM09] Satake, Junji ; Miura, Jun: Robust Stereo-Based Person Detection and Tracking for a Person Following Robot. In: *Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking.* Kobe (Japan), 2009

[Spi08] Spinello, Luciano: Human Detection using Multimodal and Multidimensional Features. In: *In Proceedings of the International Conference in Robotics and Automation (ICRA).* Pasadena, USA, 2008

[STS10] Spinello, Luciano ; Triebel, Rudolph ; Siegwart, Roland: Multiclass Multimodal Detection and Tracking in Urban Environments. In: *Int. J. Rob. Res.* 29 (2010), Oktober, Nr. 12, S. 1498–1515. – ISSN 0278–3649

[TC05] Topp, Elin A. ; Christensen, Henrik I.: *Tracking for Following and Passing Persons.* 2005

[TS09] Theodoridis Sergios, Koutroumbas K.: *Pattern Recognition.* Elsevier LTD, Oxford, 2009. – 629–639 S.

[TSTC03] Thayananthan, A. ; Stenger, B. ; Torr, P. H. S. ; Cipolla, R.: Shape Context and Chamfer Matching in Cluttered Scenes. In: *Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition.* Washington, DC, USA : IEEE Computer Society, 2003 (CVPR'03), S. 127–133

[UO99] Utsumi, A. ; Ohya, J.: Multiple-hand-gesture tracking using multiple cameras. In: *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* Bd. 1, 1999. – ISSN 1063–6919, S. 2 vol. (xxiii+637+663)

[VJS05] Viola, Paul ; Jones, Michael J. ; Snow, Daniel: Detecting Pedestrians Using Patterns of Motion and Appearance. In: *Int. J. Comput. Vision* 63 (2005), Juli, Nr. 2, S. 153–161. – ISSN 0920–5691

[WKSE11] Wachs, Juan P. ; Kölsch, Mathias ; Stern, Helman ; Edan, Yael: Vision-based Hand-gesture Applications. In: *Commun. ACM* 54 (2011), Februar, Nr. 2, S. 60–71. – ISSN 0001–0782

[WZ09] Wen, Jiajun ; Zhan, Yinwei: Vision-based Two Hand Detection and Tracking. In: *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. New York, NY, USA : ACM, 2009 (ICIS '09), S. 1253–1258

[WZZ+10] Wu, Shuangqing ; Zhang, Yin ; Zhang, Sanyuan ; Ye, Xiuzi ; Cai, Yiyu ; Zheng, Jianmin ; Ghosh, Soumita ; Chen, Wenyu ; Zhang, Jane: 2D Motion Detection Bounded Hand 3D Trajectory Tracking and Gesture Recognition Under Complex Background. In: *Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry*. New York, NY, USA : ACM, 2010 (VRCAI '10), S. 311–318

[XCA11] Xia, Lu ; Chen, Chia-Chih ; Aggarwal, J. K.: Human Detection Using Depth Information by Kinect. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. Austin, TX, USA, 2011, S. 15–22

[XPCR05] Xavier, Joao ; Pacheco, Marco ; Castrom, Daniel ; Ruano, Antonio: Fast Line Arc Circle and Leg Detection from Laser Scan Data in a Player Driver. In: *in Proc. of the IEEE Int. Conference on Robotics & Automation (ICRA 05)*, 2005

[YFMT08] Yuan, Miaolong ; Farbiz, Farzam ; Manders, Corey M. ; Tang, Ka Y.: Robust Hand Tracking using a Simple Color Classification Technique. In: *Proceedings of The 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*. New York, NY, USA : ACM, 2008 (VRCAI '08), S. 6:1–6:5

[ZAA11] Zhang, Zhong ; Alonzo, Rommel ; Athitsos, Vassilis: Experiments with Computer Vision Methods for Hand Detection. In: *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA : ACM, 2011 (PETRA '11), S. 21:1–21:6

[Zha04] Zhang, Qilong: Extrinsic Calibration of a Camera and Laser Range Finder. In: *In IEEE International Conference on Intelligent Robots and Systems (IROS*, 2004, S. 2004

[ZK07] Zivkovic, Zoran ; Kröse, Ben: Part Based People Detection using 2D Range Data and Images. In: *In IEEE RSJ International Conference on Intelligent Robots and Systems*. San Diego, USA, 2007, S. 214–219

[ZL02] Zhang, Dengsheng ; Lu, Guojun: A comparative Study of Fourier Descriptors for Shape Representation and Retrieval. In: *Proc. of 5th Asian Conference on Computer Vision (ACCV*, Springer, 2002, S. 646–651