Universität Karlsruhe
Fakultät für Informatik
Institut für Theoretische Informatik
Prof. Dr. A. Waibel

Schätzung der Oberkörperorientierung in einem Intelligenten Raum

Diplomarbeit

Lukas Rybok

November 2008

Betreuer: Dipl. Inform. M. Voit
Dr. R. Stiefelhagen
Prof. Dr. A. Waibel

Hiermit versichere ich, dass die vorliegende Arbeit ohne fremde Hilfe erstellt, keine anderen als die angegebenen Quellen benutzt und die den benutzten Quellen wörtlich entnommenen Stellen als solche kenntlich gemacht wurden.

Lukas Rybok

Karlsruhe, 17. November 2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Seitdem Computer angefangen haben in unser alltägliches Leben Einzug zu halten, müssen sie auf eine für den Menschen unnatürliche Weise über Tastatur und Maus bedient werden. Der Umgang mit ihnen bedarf einer gewissen Einarbeitung, da deren Bedienung für den Menschen nicht intuitiv ist. Dies ist mit ein Grund, für die teilweise immer noch vorhandene Abneigung gegenüber Rechnern in vielen Teilen der Gesellschaft, obwohl die Menschen viel von der Computerisierung profitieren können.
Daher ist es eine Vision für die Zukunft, Schnittstellen bereitzustellen mit denen die Menschen auf eine natürliche Art und Weise durch Sprache und Gestik mit Computern in allen Arbeitsbereichen interagieren können. Der Mensch soll nicht mehr der Maschine anpassen müssen, sondern umgekehrt. Solche Schnittstellen können zum Beispiel in einem *Intelligenten Raum* Verwendung finden. Ein solcher Raum setzt eine Reihe von Sensoren ein um rechnergestützt zu erkennen, was in ihm vorgeht. Er soll selbständig entscheiden können welche Informationen von Bedeutung sind und diverse Dienste für die Benutzer zur Verfügung stellen. Eine Aufgabe wäre es die Möglichkeit anzubieten im Raum befindliche Geräte per Sprachbefehl zu steuern oder den Benutzer auf Wunsch mit Informationen, die für ihn als interessant erachtet werden, zu versorgen.
Der Raum soll allerdings nicht nur Dienste auf Befehl ausführen, sondern den Benutzer auch proaktiv bei seinen Aufgaben unterstützen und dabei lernen, wie er die individuellen Wünsche des Benutzers erfüllen kann. So könnte er z.B. die Beleuchtung und Raumtemperatur auf die Vorlieben des jeweiligen Benutzers automatisch anpassen und lernen, wenn der Benutzer seine Präferenzen ändert.
Um seine unterstützende Rolle erfüllen zu können, ist für den Computer jedoch einiges an Weltwissen notwendig. Er sollte die Fähigkeit haben die menschlichen Kommunikationskanäle wie Gestik, Mimik oder Sprache erken-

nen und deuten zu können. Auch das Wissen über die Identität, Position oder Absichten der Benutzer ist von Bedeutung.

Einen Hinweis zur Erlangung dieses Wissens bietet die Orientierung des menschlichen Oberkörpers. So kann daraus bei der Verfolgung einer Person geschlossen werden, in welche Richtung sie sich vermutlich bewegen wird. Auch bei der Bestimmung von Personen oder Objekten, mit denen der Benutzer eines Intelligenten Raums interagiert kann das Wissen über seine Orientierung hilfreich sein. Denn damit sollte es möglich sein zu bestimmen, wohin der Schwerpunkt seiner Aufmerksamkeit langfristig ausgerichtet ist. Zum Beispiel kommt es immer wieder vor, dass Zuhörer eines Vortrags kurzfristig ihre Aufmerksamkeit auf andere Ziele lenken als auf den Vortragenden. Aber dennoch bleibt Ihr Oberkörper in Richtung des Sprechers orientiert, da dieser im Hauptaugenmerk der Zuhörer liegt.

Generell kann in Situationen, in denen Gruppen miteinander interagieren, das Wissen über die Orientierung der Teilnehmer Aufschluss geben über deren Gruppenzugehörigkeit. Im bereits angesprochenen Szenario eines Vortrags wäre es damit möglich rein visuell das Publikum vom Vortragenden zu unterscheiden.

1.1 Aufgabenstellung

Im Rahmen dieser Arbeit soll ein System entwickelt werden, welches basierend auf mehreren unterschiedlichen Kameraansichten, die Oberkörperorientierung einer sich in einem Intelligenten Raum befindlichen Person schätzt.

Informationen über die Personenposition sind in Form von Annotationen gegeben. Dies hat den Vorteil, dass Aussagen über die Schätzgenauigkeit des Systems gemacht werden können, ohne Störeinflüsse, die von Trackingproblemen stammen, beachten zu müssen. Außerdem würde die Entwicklung eines Personentrackers deutlich über den Rahmen dieser Arbeit hinausgehen.

Die Sensorausstattung umfasst vier digitale Kameras mit einer Auflösung von 640×480 Pixeln, die in den oberen Ecken des Raums fest angebracht sind. Deren Ausgaben sollen die Basis darstellen für die Hypothesenbildung über die Orientierung der beobachteten Person. In Abbildung 1.1 sind die Kameraanordnung, sowie das Weltkoordinatensystem skizziert.

Das System soll jedoch unabhängig von der Anzahl und Position der Kameras eine Schätzung für die Orientierung des Oberkörpers liefern können. Eine wichtige Rolle spielt dabei neben der Systemleistung auch die Laufzeit, da das entwickelte Verfahren in Echtzeit eingesetzt werden soll.

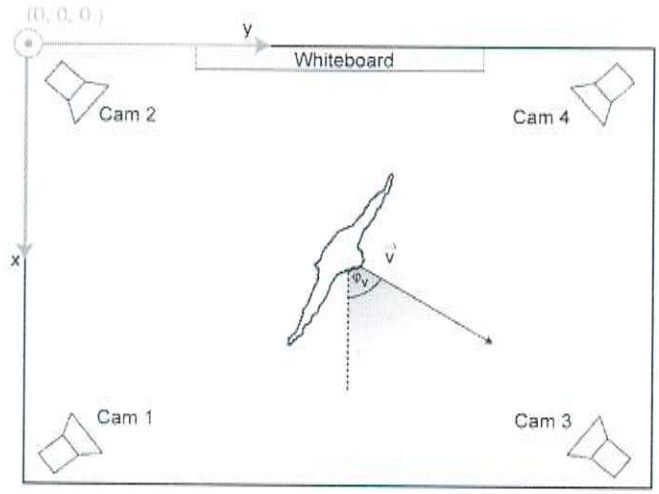


Abbildung 1.1: Skizze des Szenarios, in dem das System den Orientierungswinkel von menschlichen Oberkörpern bezüglich des im Bild angegebenen Raumkoordinatensystems schätzen soll.

1.2 Verwandte Arbeiten

Bisher sind keine Arbeiten bekannt, die sich mit dem alleinigen Schätzen der Oberkörperorienterung befassen. Allerdings kann dieses Problem mit Hilfe des *Articulated Body Tracking* gelöst werden, da es sich leicht auf die Orientierung des Oberkörpers schliessen lässt, wenn die Körperpose bekannt ist. Diesem Gebiet der digitalen Bildverarbeitung wird in der Forschung sehr viel Beachtung geschenkt, u.a. da er eine breite Reihe an praktischen Einsatzmöglichkeiten bietet, wie beim markerlosen Motion Capturing oder bei der Mensch-Roboter Interaktion.
Generell lassen sich mit *modellbasierten* und *ansichtsbasierten* Verfahren zwei Ansätze zur Lösung des Problems unterscheiden. Im folgenden werden einige Forschungsarbeiten, die sich jeweils einem der beiden Ansätze widmen, vorgestellt.
Bei modellbasierten Verfahren wird ein Hintergrundwissen über die Struktur des menschlichen Körpers vorausgesetzt, welches zur Merkmalssuche in der Bildinformation ausgenutzt wird. Diese Merkmale dienen wiederum der Rekonstruktion und dem Tracking der Körperpose.
Ein frühes modellbasiertes Verfahren wurde im Pfinder System [32] implementiert. Dieses baut dynamisch ein 2D Modell einer Person auf und nutzt zur Identifikation der einzelnen Körperregionen sowohl Informationen über die Zugehörigkeit von Objekten zum Vordergrund einer Szene, als auch über

die Farbe der einzelnen Regionen. Zur Initialisierung des Systems muss der Benutzer eine bestimmte Körperpose einnehmen, damit die anfängliche Position seiner Körperteile ermittelt werden kann.
In der Arbeit von Delamarre et al. [12] wird ein dreidimensionales Modell des menschlichen Körpers verwendet, welches in jedem Zeitschritt auf die Bildebene projiziert wird. Anschließend wird die Projektion auf der Personensilhouette ausgerichtet und das dadurch gewonnene Wissen verwendet um die neue Körperkonfiguration zu bestimmen. Die Benutzung von zwei Kameraansichten auf die Person soll helfen Unklarheiten aufzulösen, die z.B. durch Verdeckungen entstehen können. Ein Nachteil dieses Verfahrens ist jedoch, dass die initiale Körperpose bekannt sein muss.
Deutscher et al. [13] verwenden einen Partikelfilter zum artikularen Tracking der Körperpose. Als Merkmale dienen ein Kantenbild der Person sowie ihre Silhouette. Damit ist die Merkmalsdimension jedoch zu hoch für den Einsatz eines reinen Partikelfilteransatzes, da die Anzahl der benötigten Partikel exponentiell von der Merkmalsdimension abhängt. Daher wurde für deren Arbeit der *Annealed Particle Filter* entwickelt. Dieser nutzt bei der Suche nach dem globalen Maximum der Gewichtungsfunktion ein Verfahren der *simulierten Abkühlung* (engl. simulated annealing), mit dem lokale Maxima verlassen werden sollten damit bessere gefunden werden können. Dies führt dazu dass das Filter mit einer geringeren Anzahl von Partikeln eine vergleichbare Leistung hat, wie der Originalalgorithmus.
Das von Knoop et al. [17] entwickelte *VooDoo* System verwendet eine Swiss-Ranger *time-of-flight* Kamera zur Gewinnung räumlicher Informationen. Diese Spezialkamera liefert eine 3D Punktwolke der Szene, an der ein Körpermodell ausgerichtet wird. Dazu wird der *Iterative Closest Point* Algorithmus benutzt, mit dem iterativ zwei Punktwolken aneinander angepasst werden können (Registrierungsproblem).
Einen ähnlichen Ansatz verfolgen Ziegler et al. für das Tracking der artikularen Bewegung des Oberkörpers [34]. Zur Erstellung einer Punktwolke der Person wird jedoch eine Vordergrundsegmentierung angewandt, die auf Disparitätenbildern basiert. Diese stammen von vier im Raum verteilten Stereokamera-Paaren. Auf der Punktwolke wird auch bei diesem Ansatz ein Körpermodell ausgerichtet. Allerdings wird dazu das Registrierungsproblem zu einem linearen Schätzproblem umformuliert, welches mit einem *Unscented Kalman Filter* gelöst wird. Mit diesem Filter lassen sich, genau wie mit einem erweiterten Kalman Filter die Zustände eines nichtlinearen Systems schätzen. Es hat aber den Vorteil, dass keine Jacobi-Martizen berechnet werden müssen, was den Filterentwurf erleichtert.
Auch Iwasawa et al. [16] nutzen eine Multikameraumgebung und bestimmen darin die Oberkörperorientierung und die Position einzelner Körperteile

durch eine geometrische Analyse der Körpersilhouette. Zur Positionsbestimmung einzelner Gelenke wird zusätzlich ein genetischer Algorithmus eingesetzt. Ein Nachteil des Systems ist jedoch, dass es ein personenabhängiges Körpermodell benutzt, welches vorher eingelernt werden muss. Außerdem muss der Benutzer in der Initialisierungsphase eine vorgegebene Körperpose einnehmen.
Anders als die meisten Verfahren kann das von Mittal et al. [21] auch mit gegenseitigen Verdeckungen von Personen umgehen. Um dies zu erreichen werden die Personenkonturen in einer Multikamerumgebung anhand eines Farbmodells und Aufenthaltswahrscheinlichkeiten einzelner Körperteile extrahiert. Dadurch kann in jeder Kameraansicht der Umriss der verdeckenden Person gefunden werden. Die Silhouetten werden in Regionen aufgeteilt und deren 3D Position mit Hilfe der Epipolargeometrie bestimmt. Die so erhaltenen Körperteile werden mittels eines Beobachtungsmodells auf ein einfaches Zylindemodell des menschlichen Körpers übertragen.
Mikic et al. verwenden mehrere Kameras um mit Voxeln die Körpersilhouette in 3D beschreiben zu können [20]. Innerhalb der extrahierten Punktwolke werden die Körperteile mit Template Matching identifiziert. Die initiale Schätzung der Körperpose wird anhand von Position und Größe der gefundenen Körperteile mit einem Kalman Filter validiert. Selbes Filter, jedoch mit einem zusätzlichen Zustand (Geschwindigkeit des Torsos) und einer an den neu hinzugekommenen Zustand angepassten Transitionsmatrix wird zum Tracking der Körperpose eingesetzt.
Alle hier vorgestellten modellbasierten Verfahren haben gemeinsam, dass sie entweder ein zu einfaches Modell verwenden, das ihre Leistung stark einschränkt. Oder sie sind zu komplex, als dass sie sich für einen Einsatz in Echtzeit eignen würden. Außerdem ist die Loslösung der Oberkörperorientierung aus dem Körpermodell problematisch, weil ihre Bestimmung von Position und Orientierung der übrigen Körperteile abhängt.
Aus diesem Grund wurde für diese Arbeit ein ansichtsbasiertes Vorgehen gewählt. Dabei wird nur die Bildinformation zur Hypothesenbildung benutzt. Dieses hat jedoch den Nachteil, dass die Genauigkeit von ansichtsbasierten Schätzverfahren stark von der Umgebung abhängt. So wirken sich z.B. eine ungünstige Lichteinstrahlung oder ungeeignete Hintergründe oft negativ auf die Leistung aus.
Ein Beispiel für ein rein ansichtsbasiertes System wird von Rosales et al. in [24] gegeben. Darin wird die Personensilhouette mit einer Vordergrundsegmentierung extrahiert und durch *Hu-Momente*, einem Satz translations-rotetions- und skalierungsinverianten von Bild-Momenten, beschrieben. Die Abbildung des dadurch erhaltenen Merkmalsvektors auf die Körperpose wird durch ein künstliches Neuronales Netz erreicht. Zur Bildaufnahme wird eine

einzige Kamera verwendet. Dies führt zu einem Verlust der Tiefeninformationen, weshalb das System Schwierigkeiten hat unklare Körperhaltungen richtig zu deuten.
Dagegen nutzen Sun et al. [29] in ihrer Arbeit mehrere Kameraansichten einer Szene, was eine Beschreibung der Körperkontur mit Voxeln ermöglicht. Diese werden mit dreidimensionalen Shape-Contexts *(3DSC)* codiert und anschließend zu einen niedrigdimensionalen Merkmalsvektor mit einer probabilistischen PCA transformiert. Die Abbildung der Merkmale auf einen 55-dimensionalen Vektor, der die einzelnen Gelenkwinkel des menschlichen Körpers repräsentiert, geschieht mit einem *Relevance Vector Machine (RVM)* Regressor. Bei der RVM handelt es sich um ein Verfahren, welches zwar auf einem Bayes'schen Rahmenwerk aufsetzt, aber in seiner Funktion identisch ist mit einer SVM. Im Vergleich zur SVM wird die Trennhyperebene jedoch mit einer deutlich geringeren Anzahl von Stützvektoren beschrieben.
Einen ähnlichen Ansatz verfolgen Agarwal et al. [1] in ihrem System zur Schätzung der Körperpose. Allerdings wird darin nur eine einzige Kameransicht verwendet und die Personensilhouette wird zweidimensional mit *Histogram of Shape Contexts (HoSC)* codiert.
Ein Nachteil der vorgestellten ansichtsbasierten Verfahren ist, dass ihre Schätzung entweder auf nur einer Kameransicht basiert, was deren Leistung beschränkt. Oder es werden mehrere Ansichten verwendet, womit jedoch starke Einbußen in deren Geschwindigkeit verbunden sind.

1.3 Inhaltsübersicht

Das entwickelte System besteht aus vier Komponenten, auf die in Kapitel 2 im Detail eingegangen wird.
Zuerst wird, wie in Abschnitt 2.2 beschrieben, eine adaptive Vordergrundsegmentierung durchgeführt, um die Konturen von Personen zu finden. Innerhalb dieser Regionen wird die Silhouette des jeweiligen Oberkörpers mit den in Abschnitt 2.3 vorgestellten Verfahren lokalisiert und extrahiert.
Es können jedoch Fehler in Segmentierung (und Lokalisation) der Oberkörpersilhouette auftreten. Dadurch würde die weitere Verwendung der Konturinformationen für eine akkurate Orientierungsschätzung aber unmöglich gemacht, weshalb die extrahierte Silhouette zunächst bewertet wird. Sollte diese die angesetzten Gütekriterien nicht erfüllen, wird sie verworfen. In Abschnitt 2.4 wird erläutert, wie anhand von akzeptierten Silhouetten die Orientierung des Oberkörpers im entwickelten System geschätzt wird.
Zur (diskreten) Abbildung der Merkmale auf den Orientierungswinkel werden zwei Ansätze untersucht. Die Grundidee zum ersten Ansatz basiert auf einer

von Agarwal et al. [1] vorgestellten Arbeit. Genau wie darin wird auch in vorliegender Arbeit die Silhouette zu einem *Histogram of Shape-Contexts* transformiert (siehe Abschnitt 2.4.3). Sie dient jedoch, im Gegensatz zu [1], als Merkmal für einen Support Vector Machine Klassifikator. Weil sich die Merkmalsberechnung als sehr zeitaufwändig herausgestellt hat, wurde ein weiterer Ansatz zur Schätzung der Oberkörperorientierung entwickelt. Dabei wird die Silhouette mit *Shape-Context* Merkmalen codiert (siehe Abschnitt 2.4.2). Die Klassifikation geschieht in diesem Fall mit einem Nächster-Nachbar Verfahren.
Die zuvor genannten Schritte zur Ermittlung einer Hypothese über den Orientierungswinkel des Oberkörpers werden für jede Kameransicht separat durchgeführt und im letzten Schritt des vorgestellten Algorithmus zusammengefügt. Die Fusion der Einzelhypothesen zu einer Gesamtschätzung geschieht mit Hilfe eines Bayes-Filter Ansatzes. Auf die Einzelheiten dieses Fusionsverfahrens wird in Abschnitt 2.5 eingegangen.
In Kapitel 3 wird das entwickelte System anhand von Experimenten evaluiert. Für jedes der implementierten Verfahren wird ein Referenzparametersatz ermittelt, mit dem auf diversen Videosequenzen durchschnittlich die beste Systemleistung erzielt wird. Anschließend werden beide vorgestellte Schätzverfahren bezüglich ihrer experimentell ermittelten Ergebnisse miteinander verglichen.

Kapitel 2

Schätzen der Oberkörperorientierung

Im folgenden Kapitel wird das entwickelte System zur Schätzung der Oberkörperorientierung beschrieben. Es werden zwei Schätzverfahren vorgestellt und die einzelnen Schritte, die für die jeweilige Schätzung notwendig sind erläutert.

2.1 Übersicht

Es soll anhand verschiedener Ansichten auf eine Person den horizontalen Orientierungswinkel des Oberkörpers der Person geschätzt werden. Der Winkel wird gegen den Uhrzeigersinn beschrieben und nimmt diskrete Werte an.
Als Merkmal für die Schätzung wurde die Silhouette des Oberkörpers gewählt, da sie unempfindlich ist gegenüber Oberflächenstrukturen, wie Farbe und Textur der Kleidung. Ausserdem existiert eine Vielzahl an Verfahren mit denen sich Konturen von Personen effizient extrahieren lassen [15, 11, 9, 18, 30].
Das System besteht aus vier Teilen, die aufeinander aufbauen. In Abbildung 2.1 sind die einzelnen Schritte, die das System zur Schätzung des Orientierungswinkels durchführt, in einem Ablaufdiagramm beschrieben.
Zuerst wird eine Vordergrundsegmentierung separat in jeder Kameraansicht durchgeführt, um die Silhouette der Person zu finden, deren Oberkörperorientierung geschätzt werden soll. Anschließend wird der Teil der Personenkontur extrahiert, der den Oberkörper beschreibt. Diese Oberkörpersilhouette wird nach einfachen Kriterien bewertet, ob sie sich zur Orientierungsschätzung eignet und gegebenenfalls verworfen. Im darauf folgendem Schritt wird für jede der akzeptierten Silhouetten eine Hypothese für die Oberkörperorientie-

rung gebildet. Dazu werden zwei unterschiedliche Verfahren untersucht und miteinander verglichen.
Zum einen wird die Silhouette mit einem *Histogram of Shape-Contexts* (HoSC) codiert und dient anschließend als Merkmal für einen trainierten *Support Vector Machine* (SVM) Klassifikator. Dieser liefert für jede Orientierungswinkelklasse eine Konfidenz über die Zugehörigkeit des Merkmals zur jeweiligen Klasse zurück. Der Ansatz basiert auf der Arbeit von Agarwal et al. [1] und hat den Nachteil, dass sowohl das Trainieren des Klassifikators als auch die Codierung der Konturinformationen viel Rechenzeit in Anspruch nimmt.
Aus diesem Grund wurde eine weitere Methode zur Schätzung der Oberkörperorientierung entwickelt. Dabei wird die Silhouette direkt mit *Shape-Contexts* (SC) beschrieben und es wird auf die aufwändige Histogrammbildung verzichtet. Als Klassifikator wird bei diesem Ansatz ein Nächster-Nachbar Verfahren eingesetzt.
Im letzten Schritt werden die, für die unterschiedlichen Ansichten gemachten Einzelhypothesen zu einer Gesamthypothese verschmolzen.
Um die Problemstellung zu vereinfachen, werden für das System folgende Annahmen gemacht:

1. Es wird nur die Oberkörperorientierung von einer Person geschätzt
2. Der Oberkörper der beobachteten Person wird selten von anderen Personen verdeckt
3. es ist immer die Position des Oberkörpers im Weltkoordinatensystem bekannt

Annahme 2 wurde gemacht, damit eine reine Vordergrundsegmentierung zur Lokalisation von Personenkonturen ausreicht. Durch gegenseitiges Verdecken von mehreren Personen, würden ihre Umrisse zu einer einzigen Kontur verschmelzen. Das entwickelte System benötigt jedoch deutliche Personensilhouetten um genaue Schätzungen der Oberkörperorientierung machen zu können. Das Vorhandensein von Wissen über die Position des Oberkörpers in Form von Annotationen kann als eine Simulation eines Personentrackers angesehen werden. Wird jedoch ein echter Tracking-Algorithmus [4, 14, 19] dem System vorgeschaltet, lässt sich auch Annahme 3 umgehen.

2.2 Segmentierung

Im Szenario, in dem der entwickelte Oberkörperorientierungsschätzer eingesetzt werden soll, sind die verfolgten Personen signifikant dynamischer als

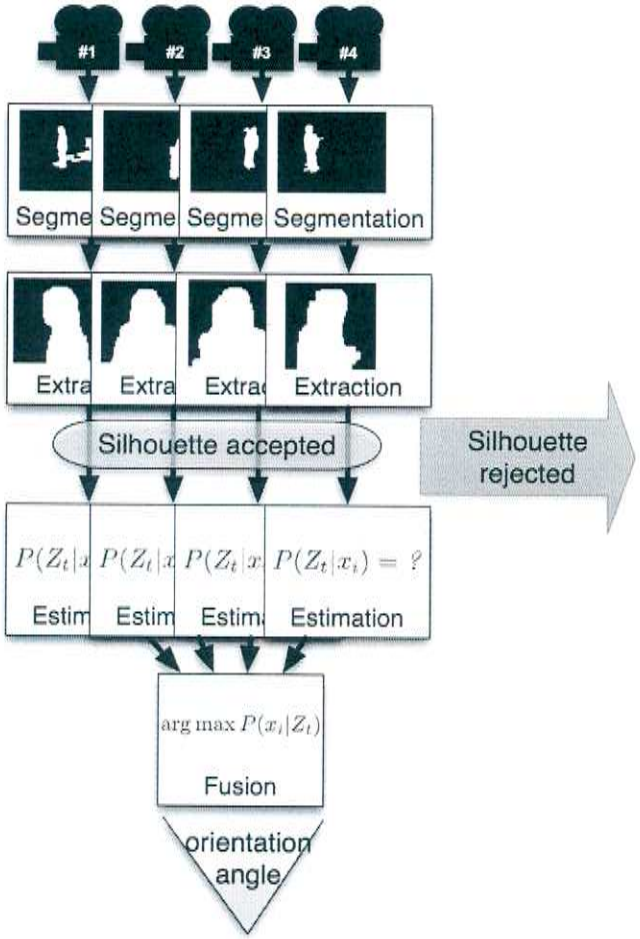
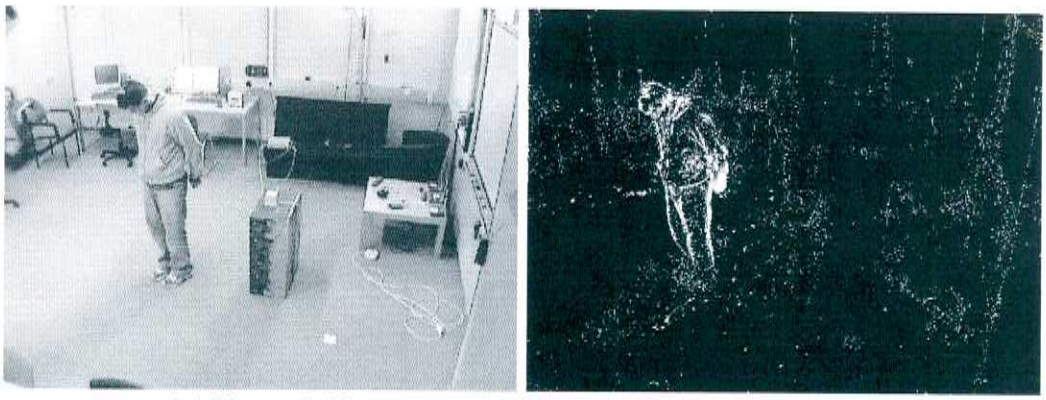


Abbildung 2.1: Ablaufdiagramm des Algorithmus zur Schätzung der Oberkörperorientierung.

der Hintergrund. Daher eignet sich die *Vordergrundsegmentierung* besonders gut um die Umrisse der verfolgten Personen zu detektieren.

Ein Verfahren, mit dem eine Klassifikation zwischen Hintergrund und Vordergrund durchgeführt werden kann und welches in einer erweiterten Form in dieser Arbeit eingesetzt wird, wird in [28] vorgestellt:

Dazu wird ein Modell des Hintergrunds verwendet, in welchem jedes Pixel durch eine Gaußmixturverteilung beschrieben ist. Dieses Hintergrundmodell wird auf einem leeren Bild der Szene gelernt und über die Zeit adaptiert, so dass Objekte, die über eine längere Zeit statisch sind, zum Hintergrundmodell hinzugefügt werden. Passt ein Pixel in den nachfolgenden Bildern nicht zu seiner Repräsentation im Hintergrundmodell, dann wird es als Vordergrund



(a) Kamerabild (b) Segmentierungsergebnis

Abbildung 2.2: Probleme des adaptiven Hintergrundmodells: Personen, die eine längere Zeit über beinahe statisch sind und deren Kleidung wenig texturiert ist, werden dem Hintergrundmodell hinzugefügt.

klassifiziert.

Eine Modifikation dieses Algorithmus, in der die Anzahl der Gaußmixturkomponenten dynamisch bis zu einem Maximalwert angepasst wird, wurde in [35] beschrieben. Dadurch ist das Verfahren im Vergleich zum Originalalgorithmus weniger rechenintensiv. Schließlich reicht es aus Bildausschnitte mit geringer Dynamik durch sehr wenige Mixturkomponenten ausreichend zu beschreiben, weshalb für sie auch weniger Berechnungen durchgeführt werden müssen. Außerdem führen diese Modifikationen zu einem etwas bessere Segmentierungsergebnis im Vergleich zum Originalalgorithmus [35].

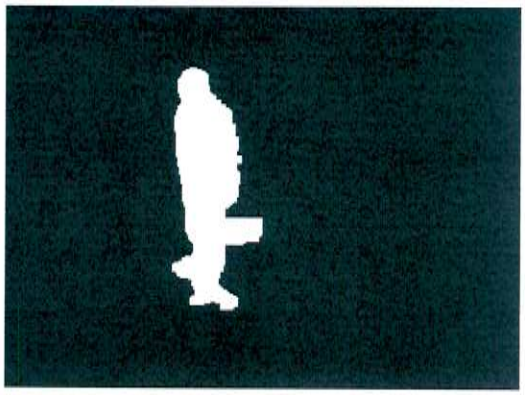
Das Adaptieren von längerfristig statischen Objekten als Hintergrund hat zwar den Vorteil, dass Veränderungen in der Hintergrundkonfiguration, z.B. durch neu geöffnete Türen oder auf dem Tisch abgelegte Objekte, im Modell angepasst werden, führt aber zu Problemen beim Einsatz des Verfahrens zur Personendetektion. Schließlich werden dadurch auch Personen zum Hintergrundmodell hinzugefügt, wenn sie eine längere Zeit über sich nicht bewegen. Dieses Problem kann aber auch auftreten, wenn die Personen nur geringe Bewegungen ausführen und wenig texturierte Kleidung tragen (siehe Abbildung 2.2b).

Um dem entgegenzuwirken wird dieser Algorithmus, wie im folgenden beschrieben und in Abbildung 2.3 anhand von Beispielen dargestellt, abgewandelt. Damit die beobachtete Person nicht Teil des Hintergrundmodells wird, werden alle zusammenhängenden Vordergrundregionen, die der Person zugeordnete Punkte enthalten bei der Modelladaption nicht berücksichtigt (die Zuordnung kann durch einen Personentracker erfolgen). Dies stellt zwar effizi-

(a) Kamerabild

(b) Endergebnis des Originalalgorithmus

(c) Ergebnis des modifizierten Algorithmus ohne Vorverarbeitung



(d) Endergebnis des modifizierten Segmentierungsalgorithmus

Abbildung 2.3: Vergleich zwischen dem Segmentierungsalgorithmus aus [35] und dem Ergebnis seiner Modifikation, bei der Vordergrundobjekte im Adaptionsschritt nicht berücksichtigt werden.

ent sicher, dass die beobachtete Person immer dem Vordergrund zugeordnet wird, hat aber den Nachteil, dass statische Objekte nun als Vordergrund klassifiziert werden, solange sie entweder mit der Person direkt in Kontakt sind oder von ihr teilweise verdeckt werden (z.B. Akten, Stuhl). Allerdings ist das Vorhandensein eines Risikos, dass die Personensilhouette lokal deformiert wird, weniger problematisch, als der sichere Verlust großer Teile der Silhouette.

Durch Fehler in der Segmentierung kann es dazu kommen, dass die Regionen einzelner Objekte nicht mehr zusammenhängend werden und dadurch trotzdem Teile einer Person zum Hintergrundmodell hinzugefügt werden. Daher werden durch das Anwenden der morphologischen Operatoren *Erosion* und *Dilatation* kleinere Segmentierungsfehler ausgeglichen. Danach werden größe-

re Löcher innerhalb einer Vordergrundregion geschlossen, indem die äußeren Umrisse der Region ausgefüllt werden.
Das Resultat der Segmentierung ist in Abbildung 2.3d dargestellt. Zwar erhält man durch die vorgestellte Vordergrundmentierung die Silhouette der beobachteten Person, aber auch deren Schatten. Dies ist jedoch für die Oberkörperextraktion unbedenklich, da Schatten (in dieser Arbeit zugrunde liegendem Szenario) selten auf Höhe des Oberkörpers auftreten.

2.3 Oberkörperextraktion

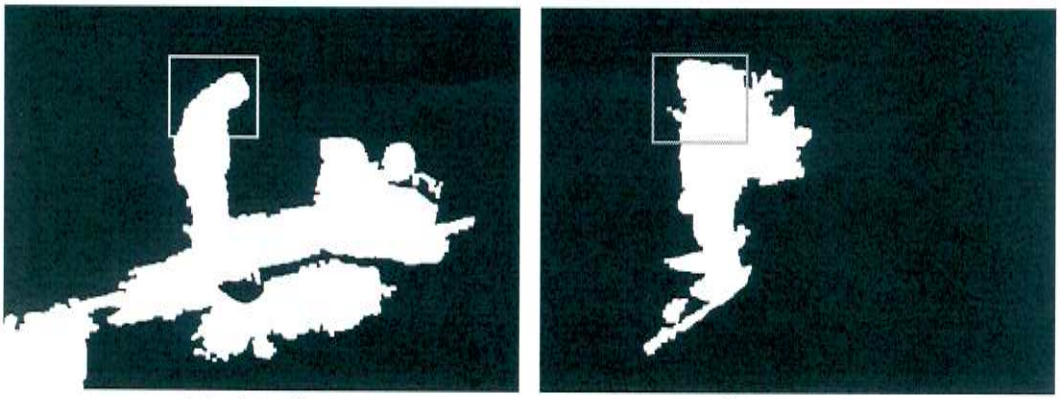
Durch die Vordergrundsegmentierung erhält man die Umrisse der beobachteten Person. Da jedoch nur die Oberkörperregion für die Schätzung interessant ist, muss diese zuerst innerhalb der jeweiligen Silhouette lokalisiert werden. Diese kann durch Segmentierungsfehler stark von ihrer wahren Form abweichen, weshalb die Güte der extahierten Oberkörpersilhouette bestimmt und bewertet wird.

2.3.1 Lokalisierung des Kopfes

Zum Finden der Oberkörpersilhouette innerhalb der Körperregion wird zuerst nach der Position von dem höchsten Punkt des Kopfes $P_i = (x_i, y_i)$ $(i = 0 \ldots \#Kameras)$ in den einzelnen Kameransichten gesucht. Dieser wird als derjenige Punkt angenommen, der innerhalb der Körperregion die niedrigste y-Koordinate besitzt. Sollten mehrere solche Punkte existieren, dann wird derjenige Punkt für P_i ausgewählt, der sich genau in der Mitte zwischen allen Kandidaten für P_i befindet.
Sobald in mindestens zwei Kameransichten jeweils ein solcher Punkt P_i lokalisiert wurde, kann mittels Triangulation die 3D-Position des Kopfes P_{Kopf} ermittelt werden. Weil die einzelnen P_i jedoch nur ungefähr mit P_{Kopf} korrespondieren, kann die auf ihrer Basis geschätzte Kopfposition stark von ihrer wahren Position abweichen. Daher wird auch der bei der Triangulation entstehende Fehler untersucht, wenn eine Ansicht ausgelassen wird. Ergibt sich durch die Auslassung eine Verbesserung des Triangulationsfehlers, dann wird dies solange wiederholt, bis entweder der Fehler nicht mehr kleiner wird oder durch die Auslassung nur eine Ansicht übrig bleibt.

2.3.2 Extraktion des Oberkörpers

Ausgehend von P_{Kopf} kann die Mitte der Oberkörperregion gefunden werden. Sie wird festgelegt als derjenige Punkt, der sich eine durchschnittliche



(a) akzeptiert (b) verworfen

Abbildung 2.4: Extraktion und Bewertung von Oberkörpersilhouetten. Eine Silhouette wird akzeptiert, wenn der Anteil ihrer Fläche im Vergleich zur Größe des Suchfensters innerhalb eines Toleranzrahmens liegt[1].

Kopfhöhe ($250mm$) tiefer entlang der z-Achse befindet als P_{Kopf}.
Um diesen Punkt wird anschließend eine dreidimensionale Box zentriert. Deren Ausmaße wurden derart gewählt, dass sie den gesamten Oberkörper unabhängig von seiner Orientierung umschließen kann. Die Boxgröße wurde empirisch ermittelt und beträgt jeweils $250mm$ in Breite und Tiefe und $260mm$ in der Höhe.
Zur Extraktion der Oberkörperregion muss diese Box auf die einzelnen Kameraansichten projiziert werden. Die grösste innerhalb des dadurch entstehenden Suchfensters gefundene Region wird als Oberkörper angenommen.

2.3.3 Bewertung der Silhouette

Da das implementierte Verfahren zur Orientierungsschätzung empfindlich reagiert gegenüber grösseren Fehlern in der Silhouettensegmentierung, wird die extrahierte Silhouette nur zur Hypothesenbildung herangezogen, wenn sie das im folgenden beschriebene Gütekriterium erfüllt.
Solche Fehler können auftreten, wenn die Kleidung eine ähnliche Farbe wie der Hintergrund hat, die Position des Oberkörpers an der falschen Stelle angenommen wird (z.B. durch über den Kopf gehobene Arme), oder wenn die Umrisse von anderen Vordergrundobjekten mit der Personensilhouette verschmelzen.
Zur Gütemessung wird der Anteil der Silhouettenfläche mit der Größe des

[1] auf die Detektion von Schatten wird verzichtet, da sie keinen Einfluss hat auf die Qualität der Oberkörpersegmentierung haben

Oberkörper-Suchfensters in Relation gesetzt. Liegt der sich ergebende Wert ausserhalb eines Toleranzrahmens $[\tau_{min}, \tau_{max}]$, der anhand von Trainigsdaten ermittelt wurde, dann wird keine Orientierungsschätzung für den extrahierten Oberkörper durchgeführt. Diese Schwellwerte werden derart gewählt, dass die Rate der falsch akzeptierten Silhouetten niedrig ist, um sicherzustellen, dass sich das Ergebnis der Hypothesenfusion möglichst nur aus Schätzungen zusammensetzt, die auf zuverlässigen Daten basieren.
Ein positiver Nebeneffekt des Verwerfens fehlerhafter Silhouetten ist, dass keine Rechenzeit für unnötige Berechnungen aufgewendet werden muss.

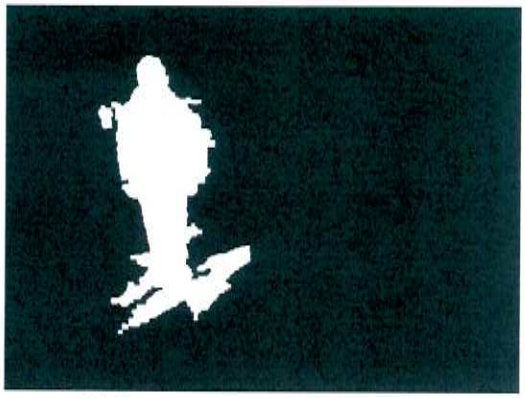
2.4 Hypothesenbildung

Auf dem Gebiet des Maschinellen Sehens wird zur Bildauswertung zwischen zwei Vorgehensweisen unterschieden:
Bei *modellbasierten* Ansätzen wird ein Vorwissen über die Objektform ausgenutzt und darauf basierend ein Objektmodell erstellt, welches zur Auswertung der Szene genutzt wird. Der Vorteil der Verfahren, die diesen Ansatz verfolgen ist, dass diese üblicherweise unabhängig von äußeren Störeinflüssen, wie Beleuchtungsunterschieden sind.
Dagegen verwenden *bildbasierte* Verfahren bloß die visuelle Information über die Szene. Sie weisen in der Regel eine bessere Generalisierungsfähigkeit auf als modellbasierte Verfahren, weshalb in dieser Arbeit ein bildbasiertes Vorgehen zur Schätzung der Oberkörperorientierung gewählt wurde.

2.4.1 Übersicht

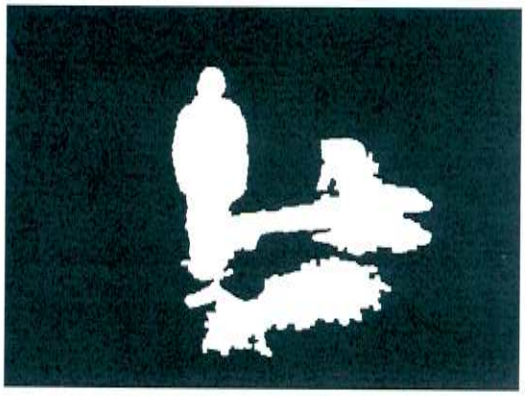
Anstatt die komplette Bildinformation über den Oberkörper zu verwenden, wird nur deren Silhouette genutzt, da sie sich gut aus Bildern extrahieren läßt, unempfindlich ist gegenüber Variationen in der Oberflächenstruktur (z.B. Farbe und Textur der Kleidung) und trotzdem viele Informationen enthält, die für die Orientierungsschätzung relevant sind. Verloren geht jedoch die Unterscheidung zwischen Vorder- und Hinteransicht auf den Körper (siehe Abbildung 2.5).
Außerdem schränken durch eine fehlerhafte Segmentierung entstehende Artefakte, wie Schatten, die Effizienz von Silhouetten als Objektbeschreibung ein, da jede lokale Störung das Gesamtergebnis beeinträchtigt. Dieses Problem tritt auch bei anderen bei *globalen* Konturrepräsentationen, wie Momenten auf [24, 7].
Dagegen sind *lokale* Umrissrepräsentationen robust gegenüber kleineren lokalen Fehlern, weshalb für diese Arbeit mit Merkmalen, die auf *Shape-Contexts*

(a) Kamerabild: Frontansicht



(b) Segmentierung: Frontansicht

(c) Kamerabild: Rückansicht



(d) Segmentierung: Rückansicht

Abbildung 2.5: Anhand von Silhouetten lassen sich in der Regel Aussagen über die Körperorientierung machen. Die Unterscheidung zwischen Vorder- und Hinteransicht auf den Körper basieren auf dessen Kontur erscheint jedoch nicht intuitiv.

[5] basieren eine lokale Beschreibung der Silhouette gewählt wurde.

Für die Schätzung der Oberkörperorientierung werden zwei Verfahren untersucht und miteinander verglichen. Einerseits dienen die Shape-Contexts in einer komprimierten Form, als *Histogram of Shape-Contexts* (HoSC) [1] als Merkmal für einen SVM Klassifikator (siehe Abschnitt 2.4.4. Im zweiten untersuchten Ansatz werden sie konkateniert zu einen einzigen Merkmalsvektor mit einem Nächster-Nachbar Verfahren, wie in Abschnitt 2.4.5 beschrieben, klassifiziert.

Damit der Klassifikator nicht für jede Kameraansicht neu trainiert werden muss, werden die geschätzten Winkel relativ zum jeweiligen Kamerakoordinatensystem angegeben. Eine Transformation der Winkel ins Weltkoordina-

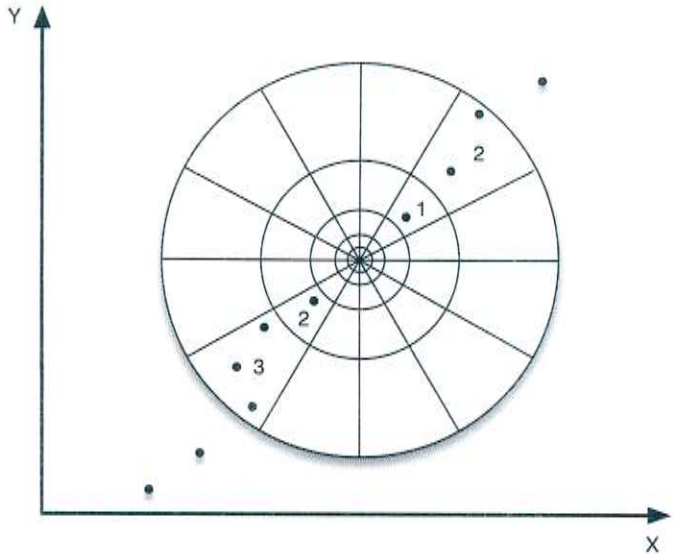
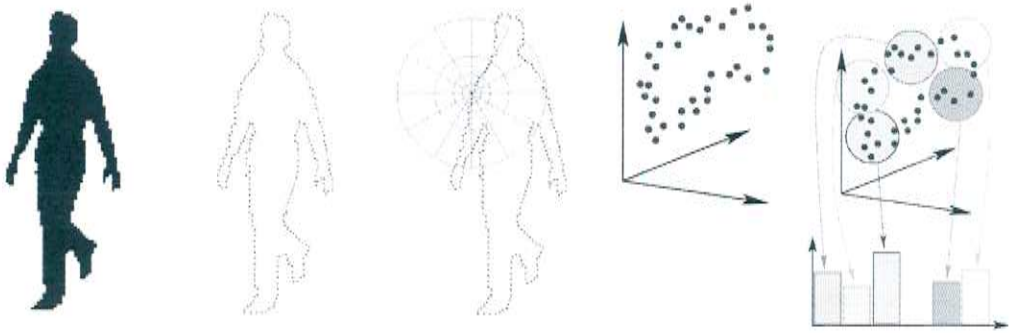


Abbildung 2.6: Transformation von Konturpunkten zu Shape-Contexts: für jeden Punkt wird ein log-polares Histogramm über die Position seiner Nachbarpunkte gebildet. Der Wert der einzelnen Histogrammklassen ergibt sich aus der Anzahl der Nachbarpunkte, die sich im jeweiligen Ringabschnitt befinden. Alle Punkte, die sich weiter vom Bezugspunkt befinden, als der Radius des Shape-Contexts, werden nicht berücksichtigt.

tensystem geschieht im Fusionsschritt. Als Klassengröße werden verschiedene Werte anhand des resultierenden Schätzfehlers miteinander verglichen.

2.4.2 Shape-Contexts

In [5] werden Shape-Contexts zur Beschreibung von Konturinformationen vorgestellt. Sie werden definiert als auf dem logarithmischen Polarkoordinatensystem basierende Histogramme von Umrisspunkten. Ein Beispiel für einen Shape-Context ist in Abbildung 2.6 gegeben. Shape-Contexts sind skalierungs- und translationsinvariant und eignen sich daher gut zur Codierung von Oberkörpersilhouetten als Merkmale für die Orientierungsschätzung. Ihre Grundidee wird in Abbildung 2.7 veranschaulicht und im folgenden erläutert.
Eine Oberkörpersilhouette wird durch die Menge ihrer Punkte repräsentiert. Diese Punkte werden zunächst im gleichen Abstand zueinander abgetastet, so dass man eine Menge $\mathcal{P} = \{p_1, \ldots, p_n\}, p_i \in \mathbb{R}^2$ von n Punkten erhält (siehe Abbildung 2.7b). Anhand von Experimenten hat sich dafür ein Wert



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

Abbildung 2.7: Codierung einer Silhouette mit einem Histogram of Shape-Contexts: (a) Extraktion der Objektsilhouette (b) Abtastung der Silhouettenpunkte (c) Berechnung von Shape-Contexts für jeden Silhouettenpunkt (d) Verteilung der Shape-Contexts im Merkmalsraum (e) Vektorquantisierung der Verteilung und Histogrammbildung (aus [2]).

von $n = 50$ als vernünftig erweisen.
Für jedes p_i kann ein Shape-Context berechnet werden, indem ein Histogramm h_i über die relative Position aller verbleibenden $n - 1$ Punkte $q \in \mathcal{P}$ im log-polaren Raum in Beziehung zu p_i berechnet wird, mit:

$$h_i(k) = |\{q \neq p_i : (q - p_i) \in Klasse(k)\}| \tag{2.1}$$

Der Wert der einzelnen Histogrammklassen ergibt sich also aus der Anzahl der Punkte, die innerhalb der entsprechenden Ringsektoren liegen. Befinden sich Punkte ausserhalb des äussersten Histogrammrings, dann werden sie nicht für den jeweiligen Shape-Context berücksichtigt, wodurch Lokalitätseigenschaft dieses Konturdeskriptors gewährleistet wird. Darüberhinaus reagieren Shape-Contexts empfindlicher auf nahe Nachbarpunkte, da die Größe der Histogrammklassen logarithmisch von ihrer Entfernung zum Bezugspunkt abhängt.
Aus der Abhängigkeit des Histogrammradius von der durchschnittlichen Distanz zwischen allen Punkten aus P, ergibt sich die Invarianz der Shape-Contexts gegenüber von Skalierung der Konturen. Die Translationsinvarianz folgt daraus, dass die Histogramme jeweils relativ zu einem Bezugspunkt berechnet werden.
Eine Möglichkeit Shape-Contexts zu berechnen wird in Algorithmus 1 in Pseudocode beschrieben. Dabei wird zur Umrechnung von kartesischen Koordinaten in Polarkoordinaten die *atan2*-Funktion eingesetzt, da es mit dem

gewöhnliches Arkustangens nicht direkt möglich ist den Winkel im korrekten Quadranten zu ermitteln.
Die Polarkoordinatentransformation wird beschleunigt durch die Ausnutzung der Symmetrieeigenschaft der Euklidischen Norm bei der Distanzberechnung zwischen zwei Punkten. Die atan2-Funktion ist zwar nicht symmetrisch, allerdings kann $atan2(q, p)$ trotzdem schnell ermittelt werden, nachdem $atan2(p, q)$ berechnet wurde (siehe Algorithmus 1).
Ein Shape-Context ist durch vier Parameter bestimmt. Die Anzahl und Größe der Histogrammklassen wird beeinflusst durch die Parameter $\#ring$ (Anzahl der Ringe) und $\#wedge$ (Anzahl der Sektoren). Deren Werte wurden anhand der Trainingsdaten experimentell bestimmt und beste Ergebnisse wurden erzielt mit $\#ring = 5$ und $\#wedge = 8$.
Über die Parameter r_{min} und r_{max} läßt sich der Radius des innersten bzw. des äußersten Ringes innerhalb des Histogramms steuern. Da für alle $q \in \mathcal{P}$ die Distanzen $q - p_i$ der Punkte zueinander mit der Durchschnittsdistanz normalisiert werden, fließen in einen Shape-Context nur Punkte ein, deren Distanz zu p_i maximal die r_{max}-fache mittlere Distanz zwischen allen Punkten beträgt.

2.4.3 Histogram of Shape-Contexts

Ein Nachteil der Nutzung von Shape-Contexts als Merkmal für einen Klassifikator ist deren hohe Dimension, da eine Objektsilhouette durch eine Menge von n Vektoren beschrieben wird, die jeweils eine Dimension von $D = \#ring * \#wedge$ haben. Mit den in dieser Arbeit gewählten Werten für die Anzahl der abgetasteten Silhouettenpunkte $n = 50$ und der Shape-Context Dimension $D = 40$ ergibt sich eine Merkmalsdimension von 2000.
Weil bei der Schätzung der Oberkörperorientierung die Geschwindigkeit des Verfahrens auch eine wichtige Rolle spielt, eignen sich die Shape-Contexts schlecht für den Einsatz in komplexen Klassifikatonsverfahren. In [1] wird jedoch ein Verfahren vorgestellt, wie sich Shape-Context Merkmale zu einem einem niedrigdimensionalen Shape-Context Histogram (*engl. Histogram of Shape-Contexts, HoSC*) komprimieren lassen.
Dazu wird zunächst Vektorquantisierung auf die Menge der Shape-Contexts angewandt. Bei der Vektorquantisierung werden Merkmalsvektoren durch einen Vektor repräsentiert, der ihnen am ähnlichsten ist. Diese repräsentativen Merkmalsvektoren sind in einer Tabelle (*Codebuch*) zusammengefasst. Das Codebuch kann erstellt werden durch die Ballung aller Mitglieder des Merkmalsraums und dem Verwenden der Clusterzentren als Einträge im Codebuch.
In [22] wird Vektorquantisierung zur Komprimierung von Shape-Contexts

Algorithmus 1 : Berechnung von Shape Contexts:
r_{min} und r_{max} geben die Radien des Histogramms an; $\#wedge$ bezeichnet die Anzahl der Sektoren des Histogramms und $\#ring$ die Anzahl der Ringe

Data : $\mathcal{P} = \{p_1, \ldots, p_n\}, p_i \in \mathbb{R}^2$; // Menge von abgetasteten Silhouettenpunkten
Result : $\mathcal{SC}(\mathcal{P})$; // Menge von Shape Contexts
begin
  // berechne alle Distanzen und Winkel zwischen den Eingabepunkten
  foreach $p, q \in \mathcal{P}$ **do**
    $dist(p,q) \longleftarrow |p - q|$; $dist(q,p) \longleftarrow dist(p,q)$;
    $\Delta_x \longleftarrow p_x - q_x$; $\Delta_y \longleftarrow p_y - q_y$;
    $angle(p,q) \longleftarrow atan2(\Delta_y, \Delta_x)$;
    // benutze die Eigenschaften von atan2 zur schnellen Winkelberechnung
    if $\Delta_x = \Delta_y = 0$ **then**
      $angle(q,p) \longleftarrow 0$
    else if $\Delta_x = 0$ **then**
      $angle(q,p) \longleftarrow -angle(p,q)$
    else if $\Delta_y < 0$) **then**
      $angle(q,p) \longleftarrow angle(p,q) + \pi$
    else if $\Delta_y \geq 0$ **then**
      $angle(q,p) \longleftarrow angle(p,q) - \pi$
  end
  $Normalize(dist(\mathcal{P}))$; // normalisiere die Punktdistanzen mit der durchschnittlichen Distanz zwischen den Punkten
  $binSize = \frac{\#ring-1}{log(r_{max})-log(r_{min})}$; // Größe einer Histogrammklasse im log-polaren Raum
  // berechne Shape Contexts durch Bildung von Histogrammen über die Silhouettenpunkte
  foreach $p, q \in \mathcal{P}$ **do**
    $ring \longleftarrow \lfloor (dist(p,q) - log(r_{min})) \cdot binSize) \rfloor$;
    if $ring < 0$ **then** $ring \longleftarrow 0$ **else** $ring++$;
    $wedge \longleftarrow \lfloor \frac{angle(p,q)*\#wedge}{2\pi} \rfloor$;
    $SC_p(ring, wedge)++$;
  end
end

erstmals vorgestellt. Die Menge der Shape-Contexts aller Trainingsdaten bildet eine multivariate Verteilung im Merkmalsraum, die mit dem k-Mittelwerte Algorithmus geballt werden kann (siehe Abbildung 2.7d). Dazu wird nach k Vektoren im Merkmalsraum iterativ gesucht, welche jeweils im Zentrum eines Clusters liegen. Diese Clusterzentren bilden zusammen mit den ihnen entsprechenden Indizes das Codebuch und werden *Shapeme* genannt.
Da es sich bei Shape-Contexts um Histogramme handelt wird zur Suche der Shapeme die χ^2-Distanz verwendet, die häufig [8, 27] als Ähnlichkeitsmaß für Histogramme genannt wird. Die Distanz χ^2-Distanz zwischen den Histogrammen h_i und h_j errechnet errechnet sich über:

$$\chi^2(h_i(k), h_j(k)) = \frac{1}{2}\sum_{i=1}^{D}\frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)} \tag{2.2}$$

Diese Codebuch-Vektoren werden im nächsten Schritt verwendet um eine mit Shape-Contexts SC beschriebene Kontur in einen HoSC-Mekrmalsvektor zu transformieren. Dabei repräsentiert jede Klasse des Shape-Context Histogramms einen Codebuch-Vektor CB. Zum Berechnen des Histogramms könnte man den Wert jeder Klasse bestimmen als die Anzahl der Shape-Contexts, die dem entsprechenden Codebook-Vektor am ähnlichsten sind. Um Quantisierunkseffekte zu vermeiden wird jedoch statt dessen *softvoting* verwendet, d.h. jeder Shape-Context steuert einen Beitrag zu mehr als einer Histogrammklasse bei. Die Zugehörigkeitsrate η_i eines Shape-Contexts zum i-ten Shape-Context kann nach [23] berechnet werden mit:

$$\eta_i(SC) = \frac{min_{r=1\ldots k}|SC - CB_r|^2}{|SC - CB_i|^2} \tag{2.3}$$

Damit HoSC-Merkmale miteinander vergleichbar sind, auch wenn sie auf einer unterschiedlichen Anzahl von Shape-Contexts basieren, werden die Histogramme zum Schluss normalisiert.

2.4.4 Orientierungsschätzung durch SVM-Klassifikation

Die Schätzung der Oberkörperorientierung basierend auf HoSC Merkmalen erfolgt mit einem Support Vector Machine (SVM) Klassifikator [26]. Die SVM soll die Winkelklasse der Oberkörperdrehung prädizieren auf Basis eines HoSC, durch den eine Oberkörpersilhouette codiert wird. Ausserdem sollen Konfidenzen über die Stärke der Zughörigkeit eines Merkmalsvektors zu jeder Winkelklasse mit geschätzt werden.

Mit einer einzigen SVM lässt sich nur über 2-Klassenprobleme entscheiden. Bei der Orientierungsschätzung soll aber die Zugehörigkeit eines HoSC zu einer von w Winkelklassen geprüft werden. Daher wird eine *One-versus-One* Strategie benutzt, bei der $\frac{w(w-1)}{2}$ Klassifikatoren ihre Hypothesen zum Klassifikationsergebnis beitragen.
Dadurch wird auch die in [33] beschriebene Bestimmung der Klassifikationskonfidenzen ermöglicht, die für die Fusion der Einzelhypothesen(siehe Abschnitt 2.5) benötigt werden.
Zur Orientierungsschätzung kommt eine *C-SVM* [6] zum Einsatz. Wie alle SVMs ist sie bestimmt durch eine Kernelfunktion und zusätzlich durch den Parameter C, der steuert, wie stark Fehlklassifikationen bestraft werden sollen. Dabei ist zu beachten, dass ein zu hoher Wert von C kann zu *Overfitting* führen kann. Overfitting bedeutet, dass dass ein Klassifikator sich an die Trainingsdaten zu sehr anpasst und deshalb schlechter generalisiert. Als Implementierung einer SVM wird in dieser Arbeit die *LIBSVM* Bibliothek [10] verwendet.
Als Kernel wurde eine *radiale Basis-Funktion (RBF)* ausgewählt. Eine lineare Basis-Funktion lässt sich zwar schneller berechnen als eine RBF, weist jedoch nur eine gute Klassifikationsleistung auf, wenn die zu klassifizierenden Daten linear separierbar sind. Der Geschwindigkeitsvorteil des linearen Kernels sollte außerdem nicht zu stark ins Gewicht fallen, da durch die HoSC-Codierung der Merkmale ihre Dimension sehr klein (=50) ist. Dagegen lassen sich die Daten mit dem Polynomkernel genau wie mit der RBF nichtlinear auf einen höherdimensionalen Raum abbilden. Die RBF zeichnet sich jedoch gegenüber dem Polynomkernel dadurch ab, dass sie durch weniger Parameter bestimmt ist. Dadurch ist die Modelbestimmung für die SVM weniger komplex, wenn ein RBF als Kernel verwendet wird. Die RBF ist definiert als:

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \tag{2.4}$$

wobei x_i und x_j zwei Merkmale beschreiben und σ die Breitedes Kernels bestimmt.
Bei der Modellauswahl für die SVM muss also ein Parameterpaar (C, γ) gefunden werden, so dass die Mißklassifikationsrate bei unbekannten Daten minimal wird. Zur Bestimmung eines geeigneten Parameterpaars wird hier auf eine Gittersuche zurückgegriffen. Es gibt zwar schnellere Verfahren zur Modellauswahl, aber die Gittersuche hat den großen Vorteil, dass sie sehr einfach parallelisierbar ist, was bei einem Einsatz auf modernen Multikernprozessoren vorteilhaft ist [10].
Bei der Gittersuche wird zuerst festgelegt welche Werte jeder der beiden Parameter annehmen darf. Anschließend werden die SVMs mit jeder Kombi-

nation der Parameterwerte trainiert und die Klassifikationsgenauigkeit ermittelt. Das Parameterpaar, welches zur geringsten Missklassifikationsrate führt, wird als Modell für die SVM ausgewählt. Besonders gut eignen sich für die Gittersuche exponentiell wachsende Sequenzen von C und γ, wobei um beste Ergebnisse zu erzielen zuerst ein grobes Gitter gewählt werden sollte, welches anschließend verfeinert wird [10]. Damit der Klassifikator nicht nur auf den Trainingsdaten eine gute Genauigkeit erreicht, sondern auch bei unbekannten Daten, wird für jede Suchepisode n-fach Kreuzvalidierung verwendet.
Da sich durch numerisch stark verschiedene Dimensionen einzelner Merkmalskomponenten die Klassifikationsgenauigkeit einer SVM verschlechtern kann [25], werden die Merkmalsvektoren linear auf den Wertebereich von $[0, 255]$ skaliert.

2.4.5 Orientierungsschätzung durch Nächster-Nachbar Klassifikation

Die Transformation von Konturpunkten zu HoSC-Merkmalen hat sich als sehr rechenintensiv erwiesen. Deshalb wird im Rahmen dieser Arbeit ein alternatives Verfahren zur Schätzung der Oberkörperorientierung mit der SVM Klassifikation von HoSC Merkmalen verglichen. Anstatt die Silhouette mit einem HoSC zu codieren, sollen die Shape-Contexts direkt den Merkmalsvektor für einen Schätzer bilden. Hierzu werden alle Shape-Contexts, die eine Silhouette beschreiben, zu einem einzigen Merkmalsvektor zusammengefügt. Dies geschieht unter Beachtung der Abtastungsreihenfolge der einzelnen Konturpunkte.
Da die Dimension dieses Merkmals je nach Parameterwahl der Shape-Contexts bei $2000 - 12000$ liegen kann, eignet sich diese Silhouettencodierung schlecht zur SVM Klassifikation. Daher wird statt dessen ein Nächster-Nachbar Ansatz für die Orientierungsschätzung verwendet.
Beim Training des Klassifikators werden alle Oberkörpersilhouetten, deren Orientierungswinkel ähnlich zueinander sind zu einer Klasse zusammengefasst. Als guter Wert für die Klassengröße wurde $4°$ experimentel ermittelt. Als Repräsentant jeder Orientierungswinkelklasse x_i $(i = 1 \dots w)$ wird der Mittelpunkt C_i über alle ihr zugehörigen Merkmalsvektoren berechnet.
Zur Klassifikation einer neuen Merkmalsinstanz z_t zum Zeitpunkt t, wird die Distanz von z_t zu allen Clusterzentren berechnet. Das Merkmal wir dann derjenigen Klasse zugeordnet, deren Repräsentant dem Merkmal am nächsten ist. Für die Fusion der einzelnen Hypothesen ist jedoch nicht nur der wahrscheinlichste Orientierungswinkel von Interesse, sondern auch die Konfidenz über die Zugehörigkeit von z_t zu jeder Klasse x_i. Diese ergibt sich aus dem

Abstand des Merkmalsvektors zum jeweiligen Clustermittelpunkt durch:

$$P(z_t|x_i) = \frac{\sum_{j=1}^{w} \chi^2(z_t, C_j)}{\chi^2(z_t, C_i)} \tag{2.5}$$

2.5 Fusion

Für jede Kameraansicht wird eine Schätzung für die Orientierung des Oberkörpers abgegeben. Um von der der Multikameraumgebung profitieren zu können, müssen diese Einzelhypothesen jedoch zu einer Gesamthypothese fusioniert werden.
Durch dieses Vorgehen sollte eine Steigerung der Schätzgenauigkeit des Systems erreicht werden, da es bei der Schätzung leicht zu Verwechslungen zwischen Vorder- und Hinteransicht auf den Oberkörper kommen kann. Das Einbinden von Hypothesen, die auf unterschiedlichen Ansichten basieren, könnte diese Verwechslungen auflösen. Auch können Segmentierungsfehler die Zuverlässigkeit der Schätzung einer Quelle stark beeinflussen, so dass die Zusammensetzung der Endhypothese aus unterschiedlichen Einzelschätzungen wünschenswert ist.

2.5.1 Bayes-Filter Fusion

Zur Fusion der einzelnen Schätzungen zu einer Gesamthypothese wird ein Bayes-Filter Ansatz verwendet, der in [31] vorgestellt wird. Die einzelnen Winkelklassen x_i, die die Oberkörperorientierung bezüglich des Raumkoordinatensystems beschreiben, bilden zusammen den Zustandsraum $X = \{x_i\}$. Gesucht ist die Wahrscheinlichkeit $P(x_i|Z_t)$, dass bei der Beobachtung eines Merkmals $Z_t = \{z_{j,t}\}$ in allen Kameransichten, der durch x_i beschriebene Orientierungswinkel vorliegt. Diese a-Posteriori Wahrscheinlichkeit lässt sich unter Anwendung des Bayestheorems [3] berechnen mit:

$$P(x_i|Z_t) = \frac{P(Z_t|x_i)P(x_i)}{P(Z_t)} \tag{2.6}$$

Die Gesamthypothese über den Orientierungswinkel $\hat{x}_t$ zum Zeitpunkt t bildet dann der Maximum a-Posteriori (MAP) Zustand, der gegeben ist durch:

$$\hat{x_t} = \arg\max_{x_i \in X} P(x_i|Z_t) = \arg\max_{x_i \in X} k \cdot P(Z_t|x_i)P(x_i) \tag{2.7}$$

Da die Auftrittswahrscheinlichkeit der Merkmale $k = P(Z_t)^{-1}$ für alle Zustände x_i gleich ist, braucht sie bei der Auswertung obiger Gleichung nicht berechnet werden.
Allerdings fließen die einzelnen a-Posteriori Wahrscheinlichkeiten auch in den Fusionsschritt zum Zeitpunkt $t+1$ mit ein (siehe Abschnitt 2.5.3), weshalb der Normalisierungsfaktor $P(Z_t)$ trotzdem ausgewertet werden muss. Dieser ergibt sich aus dem Satz der totalen Wahrscheinlichkeit zu:

$$P(Z_t) = \sum_{i=1}^{|X|} P(Z_t|x_i)P(x_i) \tag{2.8}$$

Auf die anderen Faktoren aus Gleichung 2.7 wird in den folgenden Abschnitten eingegangen werden.

2.5.2 gemeinsame Messbeschreibung

Die in Abschnitt 2.4 beschriebenen Verfahren zur Orientierungsschätzung des Oberkörpers liefern $n \in [1, \ldots \#Kameras]$ Einzelkonfidenzen $P(z_{j,t}|x_i)$ $(j = 1 \ldots n)$ für die Schätzung des diskretisierten Orientierungswinkels x_i bezüglich des Kamerakoordinatensystems der jeweiligen Kameraansicht. Diese fließen gemeinsam in die klassenbedingte Wahrscheinlichkeit $P(Z_t|x_i)$ ein, durch die Mittelung der Konfidenzen aller Klassifikatoren:

$$P(Z_t|x_i) = \frac{1}{n} \sum_{j=1}^{n} P(Z_t|\phi_j(x_i)) \tag{2.9}$$

wobei die einzelnen kamerarelativen Drehwinkel über die Transformationsfunktion ϕ_j in eine Darstellung bezüglich des Raumkoordinatensystems überführt werden.
Die Idee hinter diesem Vorgehen zur Berechnung einer gemeinsamen klassenbedingten Wahrscheinlichkeit ist, dass eine Hypothese x_i umso besser bewertet werden sollte, je mehr Kameraansichten diese stützen.

2.5.3 Einbindung zeitlicher Information

Die a-Priori Wahrscheinlichkeit für das Auftreten eines Zustands $P(x_i)$ zum Zeitpunkt t hängt vom Zustand x' zum Zeitpunkt $t-1$ ab, da die natürliche Oberkörperbewegung einigen Einschränkungen unterworfen ist. Sie lässt sich berechnen mit:

$$P(x_i) = \sum_{x' \in X} P(x_i|x')P(x'|Z_{t-1}) \tag{2.10}$$

wobei die a-Posteriori Wahrscheinlichkeit $P(x'|Z_{t-1})$ aus dem zeitlich vorhergehenden Fusionsschritt stammt. Fehlt diese Information, da keine frühere Schätzung existiert (z.B. bei der Initialisierung), dann werden alle Zustände als gleichwahrscheinlich angenommen ($P(x_i) = \frac{1}{|X|}$). Ansonsten wird ein Gauß-Kernel mit Standardabweichung σ angewandt um den Zustandsübergang zu modellieren, wodurch sich $P(x_i)$ ergibt zu:

$$P(x_i) = \sum_{x' \in X} N_{0;\sigma}(x_i - x')P(x'|Z_{t-1}) \tag{2.11}$$

Damit wird in die Bildung der Endhypothese das Wissen eingebracht, dass die Bewegung des Oberkörpers im Allgemeinen sehr gleichmäßig ist. Bei der Schätzung sollten also kleinere Änderungen der Orientierung bevorteilt und starke Änderungen als unrealistisch, und daher als Fehlklassifikationen, eingestuft werden.
Die Standardabweichung σ hängt stark vom Bewegungsstil einer Person und der Aufnahmerate der Kamera ab und wird experimentel anhand der Trainingsdaten ermittelt.

Kapitel 3

Ergebnisse und Auswertung

Im Rahmen dieser Arbeit wurden zwei Verfahren zur Schätzung der Oberkörperorientierung in einem System implementiert. Im folgenden werden beide Verfahren hinsichtlich ihrer Klassifikationsrate, des mittleren Schätzfehlers und ihres Rechenaufwands analysiert und miteinander verglichen.

Das System wurde bezüglich seiner Leistung untersucht, durch die Evaluation des mittleren Fehlers der fusionierten Gesamthypothese auf einer Menge von Videosequenzen, in denen die Position und Orientierung des Oberkörpers manuell annotiert wurden.

Da durch die Bewertung der Silhouettenqualität, die einzelnen Orientierungsschätzungen nicht immer auf allen vier Kameransichten basieren, muss die Anzahl der Einzelhypothesen aus denen sich die Gesamthypothese zusammensetzt, bei der Auswertung des Systems auch berücksichtigt werden.

Die Tests laufen nach den *leave-one-out* Prinzip ab, bei dem immer jeweils eine Sequenz aus dem vorhandenen Datensatz zur Evaluation des Systems herangezogen wird, während das Training auf den übrigen Sequenzen geschieht. Um Aussagen über die Generalisierungsfähigkeit des implementierten Verfahren machen zu können, ist die jeweilige Person aus der Testsequenz nicht Mitglied der Trainingsdaten.

Die Leistung des Systems wird von mehreren Parametern bestimmt, weshalb zunächst ein Referenzparametersatz ermittelt wurde, mit dem gute Ergebnisse bei der Orientierungsschätzung erzielt werden können. Anschließend wurde der Einfluss der einzelnen Parameter auf die Systemleistung analysiert. Wegen des geringen Umfanges an Trainingsdaten, wurde dabei allerdings auf eine Aufteilung der Daten in eine Test- und Validierungsmenge verzichtet. Statt dessen wurden die Referenzparameter auf dem Testdatensatz ermittelt, um zu untersuchen welche Systemleistung durch eine geeignete Parameterwahl möglich ist.

(a) Kamera 1 (b) Kamera 2

(c) Kamera 3 (d) Kamera 4

Abbildung 3.1: Ausgaben der vier Kameras bei jeweils unterschiedlichen Videosequenzen

3.1 Testszenario

Zur Untersuchung des entwickelten Systems wurden Videosequenzen verwendet, die in einem möblierten Raum aufgenommen wurden, um ein möglichst realistisches Szenario zu modellieren. Pro Sequenz ist immer nur eine Testperson von Interesse. Diese bewegt sich in einem kleinen vordefinierten Bereich im Raum zufällig um die eigene Achse und führt dabei diverse Kopf- und Oberkörperbewegungen durch.

Die Sequenzen stammen aus einem Datensatz, der ursprünglich für das Testen von Verfahren zur Schätzung der Kopforientierung angelegt wurde. Aus diesem Grund ist jeweils ein Magnettracker auf dem Kopf der Personen angebracht, der zur Aufzeichnung der tatsächlichen Kopfdrehung verwendet wurde. Für die Aufgabe der Oberkörperorientierungsschätzung ist dies nur in einigen wenigen Ausnahmefällen störend. Einige Beispielaufnahmen der vier zugrunde liegenden Kameras sind in Abbildung 3.1 dargestellt.

Insgesamt standen für die Experimente 9 solcher Videosequenzen, mit 8 unterschiedlichen Personen und jeweils ca. 2700 Bildern pro Aufnahme, zur Verfügung. Dies entspricht einer ungefähren Länge von 3 Minuten pro Sequenz. Auch wenn die aufgenommene Bewegung der beobachteten Personen in einem realen Szenario nicht auftreten würde, eignen sich dich die Videos trotzdem zur Evaluation des entwickelten Systems, da in ihnen eine gleichmäßige Verteilung über alle Orientierungswinkelklassen vorherrscht.
Ein Nachteil der verwendeten Testdaten ist, dass in keiner der Sequenzen Aufnahmen einer komplett leeren Szene enthalten sind. Die Momentaufnahme der leeren Szene ist jedoch für die Bildung eines Hintergrundmodells bei der Vordergrundsegmentierung notwendig. Daher musste ein solches Modell manuell erstellt werden. Dies geschah durch Ausschneiden und wieder Zusammenfügen einiger Bildausschnitte verschiedener Aufnahmen. Die Verwendung dieses synthetischen Hintergrundmodells birgt jedoch eine große Fehlerquelle in sich. Nicht entfernte Schatten von Personen oder kleine Abweichung von der Beleuchtung bzw. der Möbelposition im Hintergrundmodell führen direkt zu Segmentierungsfehlern in den Testsequenzen.
Eine weitere Fehlerquelle kommt dadurch zustande, dass die Testperson in einer der Kameraansichten eine andere Person verdeckt. Dadurch ist für diese Ansicht keine exakte Silhouettenextraktion möglich, solange die Verdeckung besteht. Allerdings ist diese Verdeckung nur in einem Teil der Videosequenzen dauerhaft.
Insgesamt lassen sich die zur Evaluation verwendeten Aufnahmen in drei Klassen einteilen:

1. das synthetische Hintergrundmodell passt gut zur Videosequenz und es kommen nur kurzfristige Verdeckungen von Personen zustande (Person 3, 5, 7, 8)

2. die Verwendung des synthetischen Hintergrundmodells führt zu leichten Segmentierungsfehlern und Verdeckungen von Personen sind möglich (Person 1a, 1b, 4, 6)

3. die Bekleidungsfarbe der Testperson weist eine hohe Ähnlichkeit auf mit der Farbe der Gegenstände, die sich hinter der Person befinden, wodurch es in einigen Kameraansichten zu starken Segmentierungsfehlern kommt (Person 2)

Die Verwendung solcher fehlerbehafteter Daten zur Evaluation des Systems hat trotzdem den Vorteil, dass damit seine Stärken und Schwächen in unterschiedlichen Situationen untersucht werden können, die auch in einer realistischen Umgebung möglich sind.

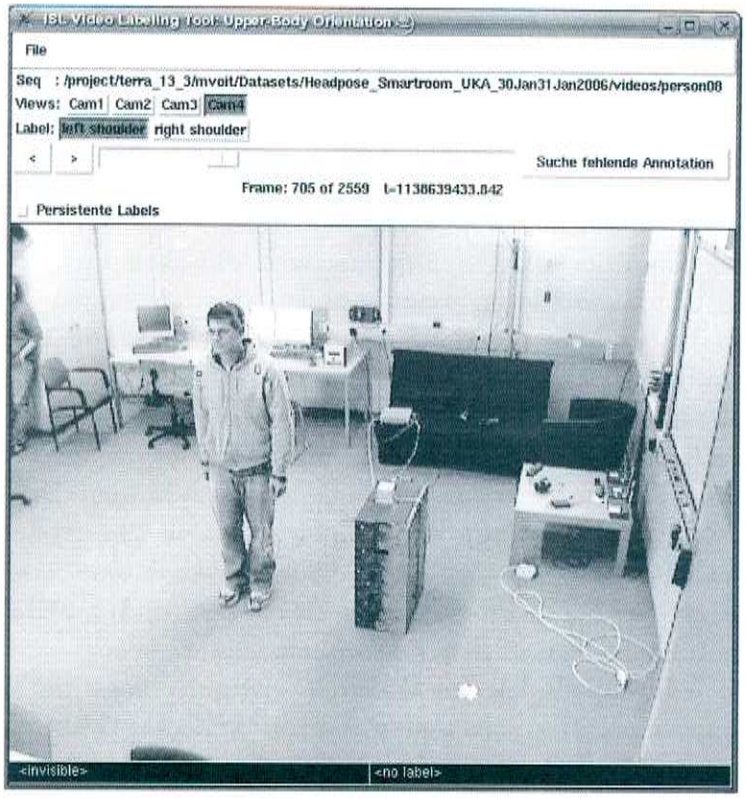
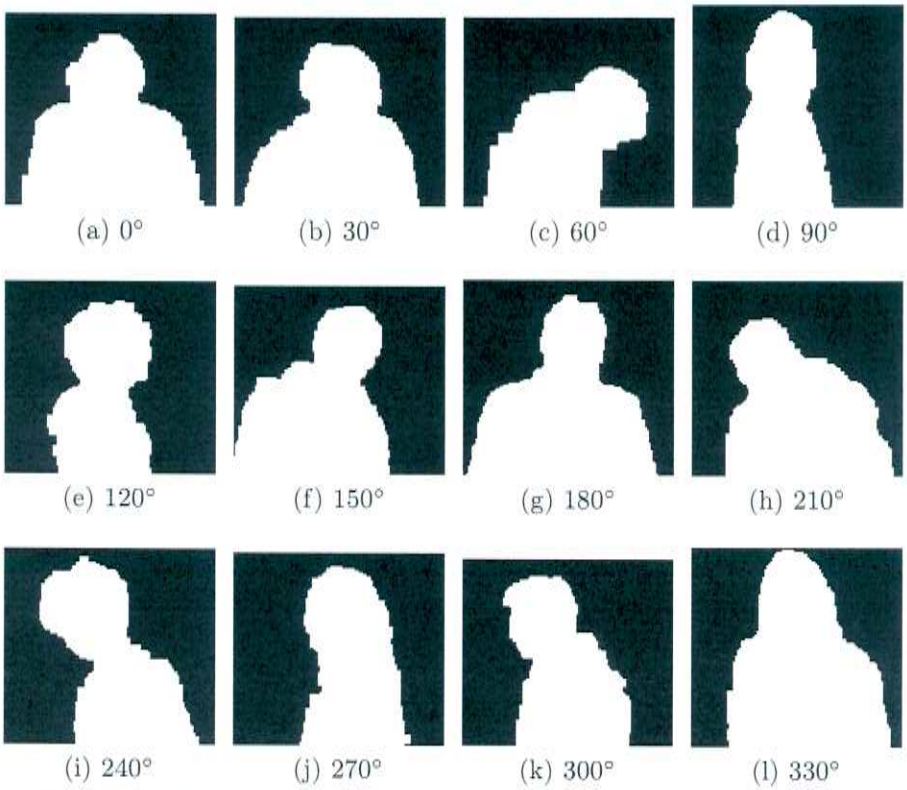


Abbildung 3.2: manuelle Annotation der Videosequenzen: Die Position der beiden Schultern wurde in jeder Kameraansicht per Mausklick markiert.

Da zur Schätzung der Oberkörperorientierung überwachte Lernverfahren eingesetzt werden, musste die Oberkörperorientierung der Testpersonen in den Evaluationssequenzen annotiert werden. Die Annotation geschah manuell mit dem in Abbildung 3.2 dargestellten Software-Werkzeug.

Dazu wurde in jedem fünften Bild die Position der beiden Schultern der Testperson markiert. Da dies für alle vier Kameraansichten erfolgte, lässt sich aus den einzelnen 2D Positionen der Schultern deren 3D Position im Raum rekonstruieren. Anhand dieser 3D Koordinaten kann anschließend die Position und Orientierung des Oberkörpers errechnet werden. Damit beide Werte für jedes Bild der Videosequenzen verfügbar sind, werden sie durch lineare Interpolation der vorhandenen Werte geschätzt.

Zur Erstellung eines Trainingsdatensatzes wurden die Oberkörpersilhouetten mit dem in Abschnitt 2.3 vorgestellten Verfahren automatisch aus den Videosequenzen extrahiert. Um sicher zu stellen, dass der Datensatz keine fehlerhaften Silhouetten enthält wurden die Konturbilder in einem zweiten Schritt manuell gefiltert.



(a) 0° (b) 30° (c) 60° (d) 90°

(e) 120° (f) 150° (g) 180° (h) 210°

(i) 240° (j) 270° (k) 300° (l) 330°

Abbildung 3.3: Beispielbilder aus dem Trainingsdatensatz für einige Winkelklassen

Weil zwischen zwei aufeinanderfolgenden Aufnahmen sich die Form der jeweiligen Oberkörperregion kaum voneinander unterscheidet, wurden nur Silhouettenbilder in den Trainingsdatensatz aufgenommen, die in jedem fünften Bild extrahiert wurden. Allerdings hängt die Leistung der untersuchten Klassifikatoren stark von der Anzahl unterschiedlicher Trainingmuster ab. Zur Verbesserung der Ergebnisse wurden Silhouettenbilder im Trainingsdatensatz synthetisch erzeugt durch das horizontale Spiegeln der bestehenden Trainingsdaten.

Insgesamt standen damit 16445 Abbildungen von Oberkörpersilhouetten zum Training der Klassifikatoren zur Verfügung, von denen einige exemplarisch in Abbildung 3.3 aufgeführt sind.

3.2 Ermittlung von Referenzparametern

Im entwickelten System kann eine Vielzahl an Parametern variiert werden. Eine feingranulare Untersuchung, wie stark sie sich auf die Schätzgenauigkeit auswirken ist daher aus kombinatorischen Gründen nicht möglich. Aus diesem Grund wurde mit einer groben Gittersuche für jedes der beiden implementierten Verfahren jeweils ein Parametersatz ermittelt, mit dem das System auf allen verfügbaren Testsequenzen den durchschnittlich kleinsten Fehler aufweist.
Während der Parameterbestimmung wurde für den Gauß'schen Kernel die Standardabweichung in der Fusion der einzelnen Klassifikationshypothesen zunächst bei einem festen Wert von $\sigma = 10°$ eingestellt. Es wird also angenommen, dass die Testperson zwischen zwei Kameraaufnahmen ihre Oberkörperorientierung um durchschnittlich 10° ändert.
Die Auflösung der Videosequenzen wurde halbiert, weil dadurch nur geringe Einbußen in der Systemleistung, dafür aber ein deutlicher Geschwindigkeitsgewinn zu verzeichnen waren.
Im Fall des SVM Ansatzes müssen für folgende Parameter Werte bestimmt werden:

1. Anzahl der Ringe eines Shape Contexts ($\#ring$)
2. Anzahl der Sektoren eines Shape Contexts ($\#wedge$)
3. Anzahl der Punkte mit denen eine Silhouette abgetastet wird (n)
4. Dimension eines HoSC-Merkmals (k)
5. Größe der Orientierungswinkelklassen (w)
6. der Kostenparameter der SVMs (C)
7. die breite des RBF-Kernels (γ)

Da die Modellbestimmung für den SVM Klassifikator der zeitaufwändigste Teil des Trainings ist, wurden die SVM Parameter in einem vorangehenden Schritt bestimmt. Dazu wurden zuerst einige Parameterwerte-Tupel $(\#ring, \#wedge, n, k, w)$ festgelegt und für die Transformation der Trainingsbilder zu HoSC-Merkmalsvektoren verwendet. Damit wurde für jedes Wertetupel der Bereich in dem die Werte für C und γ liegen sollten mit einer Gittersuche ermittelt. Um Overfitting zu vermeiden wurde für jedes Wertepaar (γ, c) die Klassifikationsrate mittels einer 10-fach Kreuzvaildierung evaluiert.

Anschließend wurde das Suchgitter auf den ermittelten Bereich eingeschränkt und die Suche wurde mit einem feineren Raster wiederholt. Das Ergebnis der Suche war, dass bei Werten von $C = 8$ und $\gamma = 2^{-15}$ die SVM unabhängig von der Parametrisierung der Methoden zur Merkmalscodierung eine gute Klassifikationsrate erzielt hat.
Um zu herauszufinden, wie die Parameter bei der Codierung mit einem HoSC am besten einzustellen sind wurde abermals eine Gittersuche verwendet. Als Maß für die Güte eines Parameterwerte-Tupels diente die Klassifikationsrate des Systems, die über alle Videosequenzen gemittelt wurde. Die Suche nach einem geeigneten Parametersatz für das Training des Nächster-Nachbar Schätzers wurde analog durchgeführt. Hierfür mussten jedoch nur die Shape-Context Parameter sowie die Grösse der Winkelklassen evaluiert werden.
Für beide Verfahren zur Schätzung der Oberkörperorientierung lagen jeweils die besten erzielten Ergebnisse, die bei der Verwendung diverser Parameterkombinationen ermittelt wurden, sehr nah beieinander. Daher wurden bei der Auswahl des Referenzparametersatzes neben der Schätzgenauigkeit noch zwei weitere Kriterien herangezogen. Weil die Geschwindigkeit des Gesamtsystems eine wichtige Rolle spielt, sollten die Werte für k und n möglichst klein sein, da diese Parameter den größten Einfluss auf die Rechenkomplexität der Merkmalsberechnung haben. Um eine gute Vergleichbarkeit beider Verfahren zu erreichen sollten außerdem ihre gemeinsamen Parameter möglichst die gleichen Werte haben. Dies führte zu einer Wahl von $\#ring = 5$, $\#wedge = 8$, $n = 50$ und $w = 30$ für beide Verfahren. Darüber hinaus wurde für den SVM Ansatz noch die Dimension eines HoSC-Merkmals zu $k = 50$ bestimmt.

3.3 Einfluss ausgewähler Parameter auf die Schätzleistung

In diesem Abschnitt wird untersucht, wie sich die Genauigkeit der Systemausgabe verhält, wenn einige Parameter variiert werden. Als Maß dafür wird die Rate der korrekt klassifizierten Orientierungswinkel des Oberkörpers herangezogen, weil das System in erster Linie zur Lösung eines Klassifikationsproblems eingesetzt wird. Daneben wird jedoch auch der mittlere Fehler der Orientierungsschätzung betrachtet, der entsteht, wenn die Systemausgabe als kontinuierlich angesehen wird.
Bei der Bestimmung des Referenzparametersatzes hat sich das System als recht unempfindlich erwiesen gegenüber der Wahl einiger Parameter, weshalb auf diese nicht im Detail eingegangen wird. So hatte die Dimension der

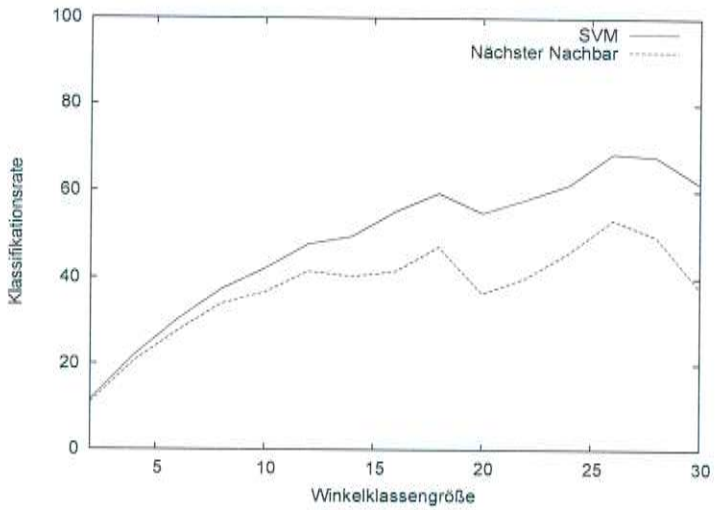


Abbildung 3.4: Abhängigkeit der Klassifikationsrate von der Winkelklassengröße: Um zu gewährleisten, dass die Klassifikationsrate eine gute Beschreibung der Systemleistung darstellt, muss die Klassengröße geeignet gewählt werden.

Shape-Contexts im untersuchten Bereich einen nur kaum messbaren Einfluss auf das Endergebnis beider Verfahren. Die Anzahl n der abgetasteten Silhouettenpunkte beeinflusst das System nur dahingehend, dass etwas bessere Resultate erzielt werden, wenn bei der Abtastung nicht alle Konturpunkte verwendet, sondern immer einige wenige (2-3) ausgelassen werden. Vermutlich hängt dies damit zusammen, dass dadurch Rauschen innerhalb der Konturinformationen etwas ausgeglichen wird. Aussagen darüber zu machen, wie die Leistung des SVM-Ansatzes von der Codebuchgröße k abhängt, ist schwierig, da der zur Codebuchbestimmung verwendete k-Mittelwerte Algorithmus indeterministisch ist. Dennoch scheinen kleinere Werte von k zu leicht besseren Ergebnissen zu führen.

3.3.1 Einfluss der Winkelklassengröße

Um Aussagen über die Leistung der implementierten Verfahren machen zu können ist eine geeignete Wahl der Winkelklassengröße w wichtig. Wird sie zu breit gewählt, dann können Schätzungen in die Bewertung des Verfahrens einfließen, die sich deutlich von der korrekten Oberkörperorientierung unterscheiden. Bei zu kleinen w werden dagegen Ergebnisse vernachlässigt, die sehr nah am wahren Ergebnis liegen. Darüber hinaus wirkt sich der Wert von w auf die Laufzeit aus, da damit die Anzahl der Klassen, die bei beiden Klassifikationsverfahren betrachtet werden müssen, implizit festgelegt wird.

In Abbildung 3.4 ist die Abhängigkeit der Klassifikationsrate von w aufgetragen. Eigentlich wäre es zu erwarten gewesen, dass mit zunehmendem Wert von w auch die Klassifikationsrate ansteigt. Der Grund warum dies in einigen Fällen nicht zutrifft hängt vermutlich damit zusammen, dass durch eine ungünstige Zuordnung der Trainingsdaten zu den entsprechenden Winkelklassen, eigentlich voneinander verschiedene Klassen im System eine ähnliche Repräsentation erhalten und daher leicht verwechselt werden können. In der Grafik wird auch deutlich, dass die Rate der korrekt klassifizierten Daten ab einer Winkelklassengröße von 18° nicht mehr so stark zunimmt, weshalb dieser Wert für w in den weiteren Experimenten verwendet wird.

3.3.2 Einfluss des Fusionsparameters

Ein weiterer Parameter, der einen großen Einfluss auf das Endergebnis der Schätzung hat, ist der Fusionsparameter σ. Über ihn lässt sich steuern, wie stark sich der Orientierungswinkel zwischen zwei Aufnahmen ändern darf. Durch eine geeignete Wahl werden Ausreißer vermieden und die Schätzung wird über die Zeit geglättet. Ein zu kleines σ sorgt jedoch dafür, dass das System zu träge reagiert um schnellen Orientierungsänderungen folgen zu können. Wird sein Wert dagegen zu hoch gewählt, dann geht der Vorteil einer Einbindung von zeitlichen Information in die Hypothesenfusion verloren. Der beschriebene Effekt kann in Abbildung 3.5 beobachtet werden. Darin ist auch ersichtlich dass bei den untersuchten Videosequenzen sich beste Ergebnisse erzielen lassen bei einem Wert von $\sigma = 12°$ (SVM Ansatz) bzw. $\sigma = 30°$ (Nächster-Nachbar Ansatz).
Allerdings muss angemerkt werden, dass der Einfluss dieses Parameters stark situationsabhängig ist. Führt die beobachtete Person generell nur sehr schnelle Bewegungen aus, so sollte σ dementsprechend hoch gewählt werden. Aus den gemachten Experimenten geht zusätzlich hervor, dass bei stark verrauschten Schätzungen höhere Werte von σ zu einem besseren Endergebnis führen.

3.3.3 Einfluss der Anzahl von Kameraansichten

Bei der Fusion spielt auch die Anzahl der Einzelschätzungen, die in die Endhypothese einfließen, eine Rolle. Eine genaue Untersuchung darüber ist jedoch schwierig, da deren Anzahl von der Bewertung der Oberkörpersilhouetten abhängt. Deshalb wurde statt dessen betrachtet, wie es sich auf die Systemleistung auswirkt, wenn die Endhypothese jeweils aus den am besten bewerteten 1 bis 4 Ansichten zusammengesetzt wird. In Abbildung 3.6 sind

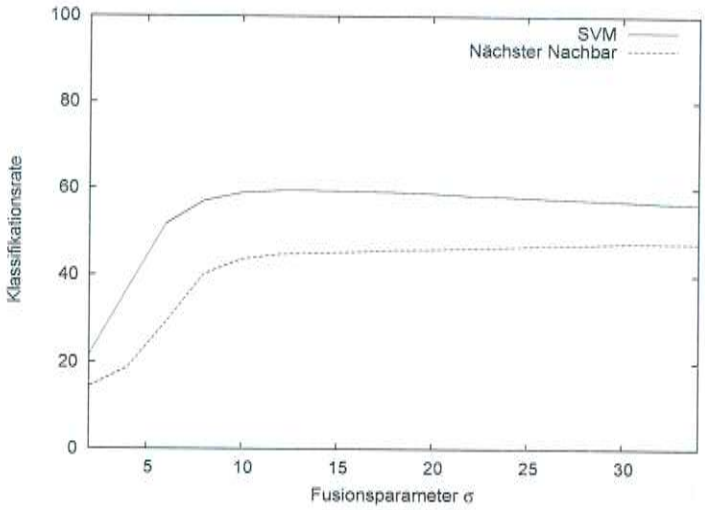
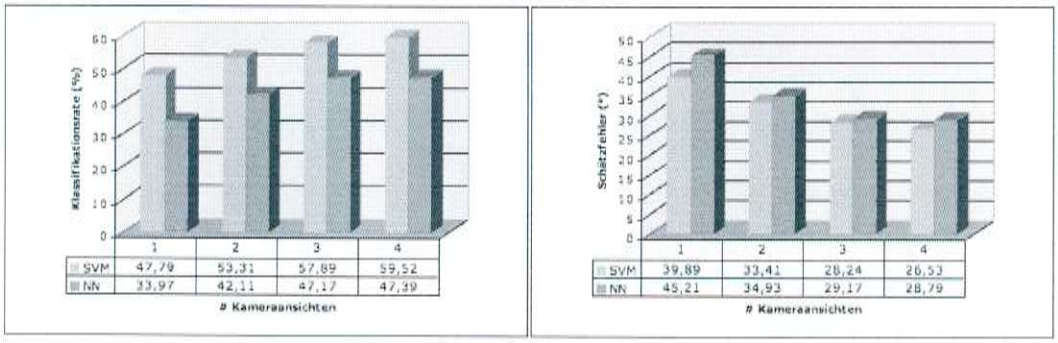


Abbildung 3.5: Leistung des Systems in Abhängigkeit vom Fusionsparamer σ.

die Ergebnisse dieser Experimente aufgeführt. Es zeigt sich, dass bei der Verwendung von mehreren Kameraansichten sich die Systemleistung verbessert. Eine mögliche Erklärung dafür ist, dass durch unklare Körperhaltungen oder Segmentierungsfehler verursachte Probleme besser aufgelöst werden können, wenn mehrere Ansichten vorliegen. Der geringe Anstieg der Genauigkeit bei der Hinzunahme von 4 statt 3 Ansichten, ist wahrscheinlich darin begründet, dass die zusätzliche Ansicht die schlechteste Segmentierungsqualität aufweist und daher in den meisten Fällen ohnehin verworfen wird.

3.4 Analyse der Ergebnisse

Wie aus Abbildung 3.7 hervorgeht, zeigen beide, der SVM Ansatz und der Nächster-Nachbar Ansatz, bei der Schätzung der Oberkörperorientierung ein ähnliches Verhalten. Die Ergebnisse fallen gemäß der in Abschnitt 3.1 beschriebenen Qualitätsklassen der segmentierten Videosequenzen aus. Die vergleichbar guten Ergebnisse bei den Sequenzen P3, P5, P7, P8, lassen sich somit erklären, dass in allen Kameraansichten weniger Segmentierungsfehler vorkommen.
Ein Beispiel für die segmentierten Ansichten der "schlechten" Sequenz P2 ist in Abbildung 3.8 gegeben. Enthalten sind auch die geschätzten Orientierungswinkel, die bloß auf der jeweiligen extrahierten Oberkörperkontur basieren. Die veringerte Erkennungsgenauigkeit des Systems bei dieser, und den übrigen problematischen Videosequenzen, liegen u.a. darin begründet, dass die

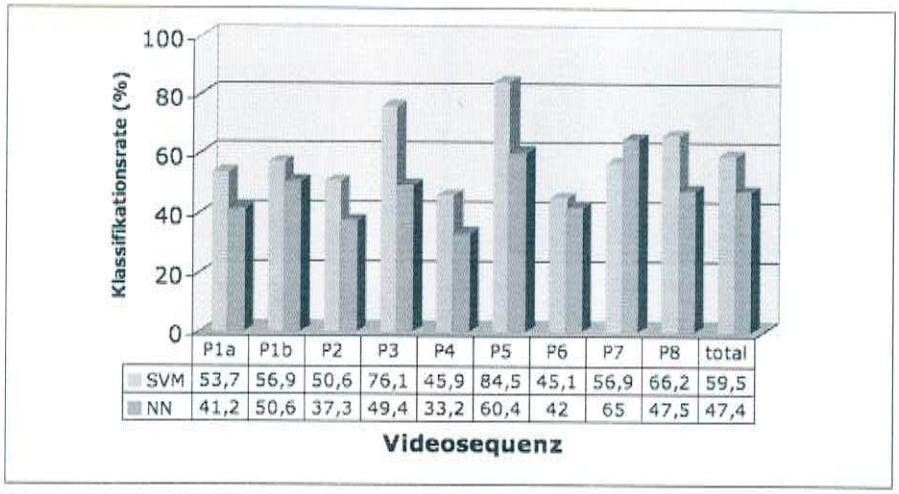


(a) Klassifikationsrate (b) Schätzfehler

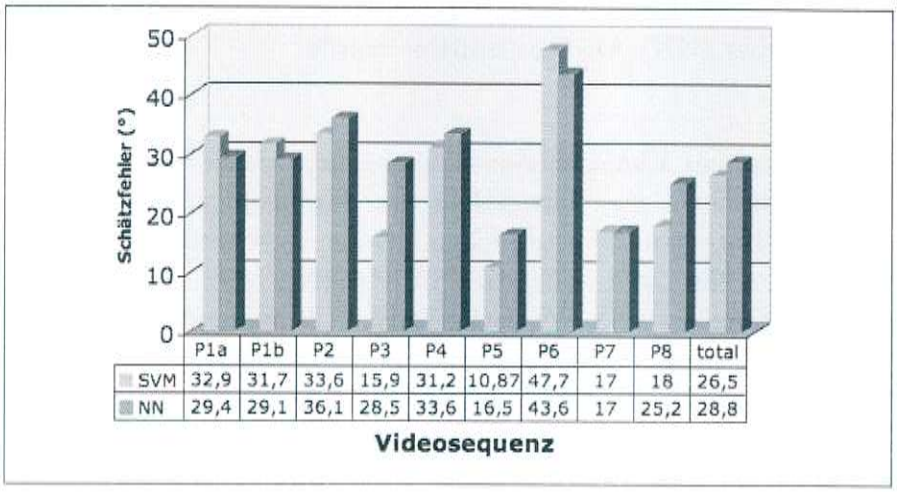
Abbildung 3.6: Schätzleistung des Systems in Abhängigkeit von der Anzahl verwendeter Kameraansichten. Die Werte beschreiben das durchschnittliche Ergebnis, welches bei allen Videosequenzen mit dem SVM- oder dem Nächster-Nachbar (NN) Ansatz ermittelt wurde.

gen problematischen Videosequenzen, liegen u.a. darin begründet, dass die extrahierte Silhouette der vierten Kamera durchgehend mit den Umrissen einer sich im Hintergrund befindlichen Person verschmilzt. Einerseits führt dies zum Fehlen von einer Schätzung im Fusionsschritt, wenn die fehlerhafte Silhouette korrekterweise verworfen wird. Auf der anderen Seite scheitert die Bewertungsfunktion manchmal daran die Kontur als fehlerbehaftet zu klassifizieren, wodurch die Schätzung der Orientierung basierend auf dieser Ansicht das Gesamtergebnis negativ beeinflussen kann.

Dadurch dass die Kleidungsfarbe der abgebildeten Testperson eine starke Ähnlichkeit aufweist mit einem sich hinter ihr befindlichen Objekt, werden in der dritten Kameraansicht Teile des Oberkörpers als Hintergrund klassifiziert. Dieser Effekt tritt zwar nur in dieser einen Videosequenz auf, aber auch in den meisten der anderen problematischen Sequenzen treten starke Segmentierungsfehler in mindestens einer der ersten drei Ansichten auf. Sie werden hervorgerufen durch das Verschmelzen des Personenumrisses mit Konturen von nahen Objekten, die im synthetischen Hintergrundmodell einen leicht veränderten Zustand haben als in der Testsequenz. Beispielsweise wird in den Aufnahmen von Person 1 ein Monitor als Vordergrundregion geschätzt, da in ihm ein anderes Bild dargestellt wird als es im Hintergrundmodell der Fall ist. Weil der Bildschirm in der Nähe des Kopfes der beobachteten Person positioniert ist, wird sein Umriss gemeinsam mit der Oberkörpersilhouette extrahiert. Solche Fehler treten zwar auch in den Videosequenzen auf, bei denen das System eine insgesamt gute Schätzleistung aufweist, aber sie wirken sich wahrscheinlich weniger negativ auf die Klassifikationsgenau-



	P1a	P1b	P2	P3	P4	P5	P6	P7	P8	total
SVM	53,7	56,9	50,6	76,1	45,9	84,5	45,1	56,9	66,2	59,5
NN	41,2	50,6	37,3	49,4	33,2	60,4	42	65	47,5	47,4

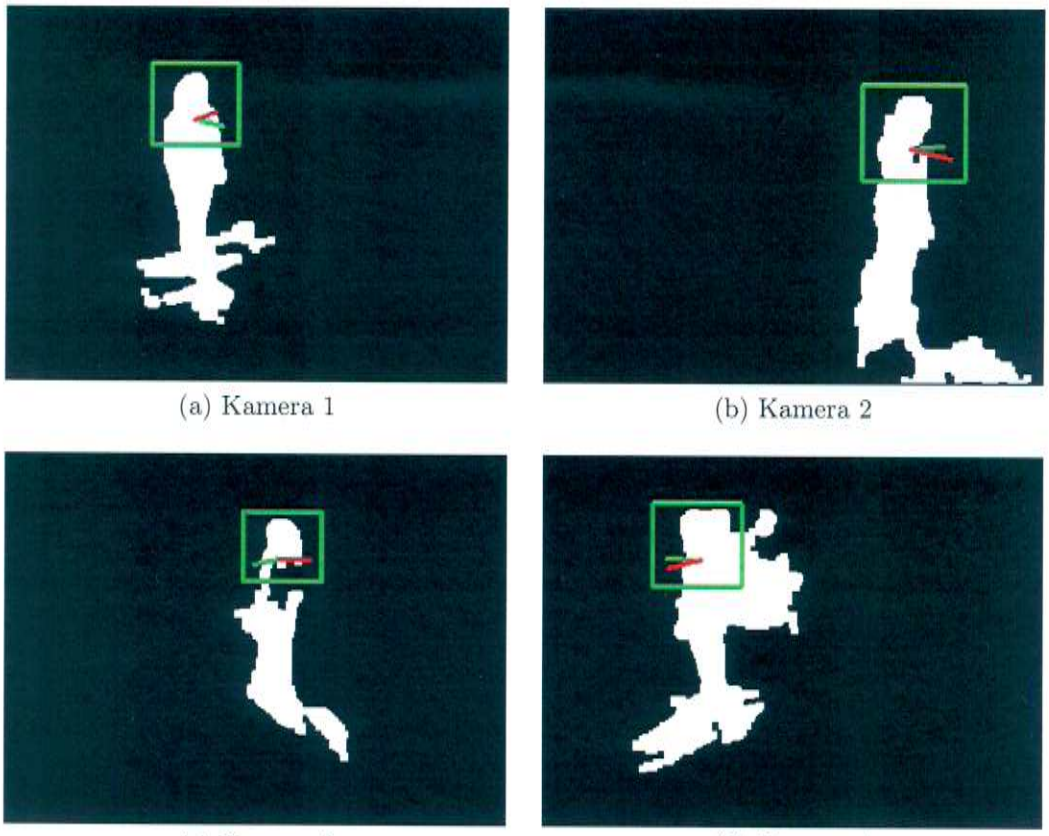


	P1a	P1b	P2	P3	P4	P5	P6	P7	P8	total
SVM	32,9	31,7	33,6	15,9	31,2	10,87	47,7	17	18	26,5
NN	29,4	29,1	36,1	28,5	33,6	16,5	43,6	17	25,2	28,8

Abbildung 3.7: Klassifikationsrate und Schätzfehler beider Verfahren auf allen Videosequenzen. Die Evaluation erfolgte nach dem leave-one-out Prinzip.

igkeit aus, aufgrund des Vorhandenseins einer klaren Oberkörpersilhouette in der vierten Kameraansicht. Damit ist eine weitere zuverlässige Hypothese vorhanden, die die übrigen korrekten Schätzungen stützt, wodurch kleinere Fehler ausgeglichen werden können.

Das vergleichbar schlechte Abschneiden des Nächster-Nachbar Ansatzes bei Person 3 hängt vermutlich damit zusammen, dass sich die Statur dieser Person deutlich von den übrigen unterscheidet. Damit sind nicht genügend Trainingsbeispiele vorhanden, um eine gute Schätzung für ihre Orientierung abgeben zu können. Dass der Orientierungswinkel dieser Person mit dem SVM Klassifikator trotzdem gut geschätzt wird, hat möglicherweise den Hintergrund, dass eine SVM besser generalisiert als ein Nächster-Nachbar Verfah-



(a) Kamera 1 (b) Kamera 2

(c) Kamera 3 (d) Kamera 4

Abbildung 3.8: Schätzung der Oberkörperorientierung in segmentierten Einzelbildern der Sequenz P2: Ein grünes Rechteck gibt eine akzeptierte Oberkörperregion an; die Linien visualisieren die auf der jeweiligen Ansicht basierende Schätzung (rot) und die wahre Orientierung (grün).

ren. Diese Vermutung wird gestützt durch die geringfügig schlechtere Leistung des Nächster-Nachbar Klassifikators auf fast allen Sequenzen.
Wie sich das System verhält, wenn man für die Orientierungsschätzung Silhouetten verwendet, welche anhand von Annotationen extrahiert wurden, ist in Tabelle 3.1 aufgeführt. Die Ergebnisse des SVM Ansatzes lassen vermuten, dass dieser empfindlich darauf reagiert, wenn die Suchfenster für den Oberkörper bei den klassifizierten Silhouetten nicht gleich ausgerichtet sind wie bei den Trainingsdaten. Die implementierte automatische Lokalisation und Extraktion der Oberkörperkonturen scheint dagegen keinen gravierenden Einfluss auf das Endergebnis des Nächster-Nachbar Ansatzes zu haben. Steht nicht die Klassifikationsrate im Vordergrund, sondern der mittlere Schätzfehler, den das System auf den Testsequenzen liefert, dann sollte die

	Klassifikationsrate	Schätzfehler
SVM	63,0%	25,6°
Nächster-Nachbar	46,3%	30,4°

Tabelle 3.1: Klassifikationsrate und mittlerer Schätzfehler des Systems, die sich ergeben, wenn die Oberkörperregionen nicht automatisch extrahiert werden, sondern auf Annotationen basieren.

	Schätzfehler bei	
	$w = 4°$	$w = 18°$
SVM	20,6°	26,5°
Nächster-Nachbar	22,9°	28,6°

Tabelle 3.2: Mittlerer Schätzfehler der implementierten Verfahren. Um Quantisierungsfehler zu vermindern sollte für die Winkelklassengröße w ein kleiner Wert festgelegt werden.

Größe einer Winkelklasse möglichst klein gewählt werden. Dadurch wird der Fehler, der durch die Quantisierung des kontinuierlichen Orientierungswinkels in diskrete Klassen entsteht, gering gehalten. In Tabelle 3.2 sind die Schätzfehler aufgeführt, die bei einer Winkelklassengröße von 4° im Mittel auf allen Testsequenzen erhalten wurden. Besonders hier zeigt sich, dass die Verwendung des Nächster-Nachbar Schätzers nur zu vernachlässigbaren Leistungseinbußen im Vergleich zum SVM Ansatz führt.
Die Stärke der Nächster-Nachbar Orientierungsschätzung liegt jedoch in ihrer Laufzeit, wie an dieser Stelle erläutert werden soll. Die entwickelten Verfahren wurde in C++ implementiert und auf einem 3 GHz System getestet. Es wurden folgende externen Bibliotheken verwendet:

1. *Open Source Computer Vision Library* (OpenCV)[1] - eine Sammlung vieler Funktionen, die die Erstellung von Anwendungen aus dem Bereich der digitalen Bildverarbeitung vereinfachen

2. *LIBSVM*[2] - ein Rahmenwerk zur Verwendung gängiger SVM Verfahren

3. *C Clustering Library*[3] - eine Bibliothek zur unüberwachten Ballung von Daten

[1] http://sourceforge.net/projects/opencvlibrary/
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[3] http://bonsai.ims.u-tokyo.ac.jp~mdehoon/software/cluster/

4. *Background Subtraction Library*[4] - eine Implementierung der adaptiven Hintergrundsegmentierung nach [35]

Mit den in Abschnitt 3.2 und 3.3 ermittelten Systemparametern konnte bei der Orienterungsschätzung mit SVMs eine Laufzeit von $34ms$ pro Ansicht ermittelt werden. Davon wurden allein $22ms$ für die Transformation der extrahierten Silhoueteninformtionen zu einem HoSC-Merkmal benötigt. Dagegen betrug die Berechnung der Merkmale für den Nächster-Nachbar Klassifikator nur $2ms$ und die Klassifikation zusätzliche $22ms$. Durch die Vordergrundsegmentierung kamen bei beiden Verfahren noch $38ms$ pro Kameransicht zur Gesamtlaufzeit des Systems hinzu.
Auch die Zeit, die für das Trainieren des Nächster-Nachbar Verfahrens benötigt wird, ist mit ca. $30min$ deutlich geringer, als die $15 - 20h$ beim SVM Ansatz. Besonders zeitaufwändig haben sich hierbei die indeterministische Bestimmung eines Codebuches mit dem k-Mittelwerte Algorithmus, sowie die Modellsuche für den SVM Klassifikator erwiesen. Die wesentlich geringere Trainingszeit des Nächster-Nachbar Ansatzes hat den Vorteil, dass bei diesem Verfahren viel einfacher ein Parametertuning durchgeführt werden kann.
Insgesamt haben sich beide im Rahmen dieser Arbeit entwickelte Verfahren zur Schätzung der Oberkörperorientierung als ähnlich gut erwiesen. Der Nächster-Nachbar Ansatz wies zwar eine deutlich schlechtere Klassifikationsrate auf, was jedoch relativiert wird, wenn der mittlere Schätzfehler betrachtet wird. In diesem Fall fällt der Unterschied zum SVM Ansatz viel geringer aus. Dennoch wird der Nachteil einer etwas schlechteren Leistung des Nächster-Nachbar Ansatzes durch eine deutlich bessere Laufzeit ausgeglichen.

[4] http://staff.science.uva.nl/~zivkovic/DOWNLOAD.html

Kapitel 4

Zusammenfassung und Ausblick

In dieser Arbeit wurde ein System entwickelt, welches die Oberkörperorientierung einer sich in einem Intelligenten Raum befindlichen Person schätzt. Zur Sensorausstattung des Raums gehören vier Kameras, die unterschiedliche Ansichten auf die Person liefern. Innerhalb deren Aufnahmen wird mit einer adaptiven Vordergrundsegmentierung die Silhouette der beobachteten Person lokalisiert. Die Position der Person geht aus Annotationen hervor.
Der Segmentierungsalgorithmus wurde derart angepasst, dass bloß Regionen bei der Adaption des Hintergrundmodells berücksichtigt werden, die nicht zum Personenumriß gehören. Dadurch soll verhindert werden, dass Personen, die über eine längere Zeit beinahe statisch sind als Hintergrund klassifiziert werden. Innerhalb der detektierten Vordergrundregionen wird die Oberkörperkontur der beobachteten Person in jeder Kameraansicht separat extrahiert. Dies geschieht auf Basis einer Triangulation der Kopfposition und anschließender Projektion einer um den Kopf positionierten Box auf die jeweilige Kameraansicht. Um fehlerhaft segmentierte Silhouetten für die Orientierungsschätzung auszuschließen, wird die Qualität der extrahierten Oberkörperregion bewertet. Nur wenn diese das definierte Gütemaß erfüllen, werden sie zur Hypothesenbildung für den Orientierungswinkel herangezogen.
Es werden zwei Klassifikationsverfahren vorgestellt, mit denen eine Schätzung der Oberkörperorientierung durchgeführt werden kann. Der SVM Ansatz basiert auf einer bestehenden Arbeit von [1] und verwendet HoSC-Merkmale zur Beschreibung der Silhouetteninformation. Weil sich die Berechnung der HoSC als rechenintensiv erwiesen hat, wurde ein weiterer Ansatz zur Orientierungsschätzung entwickelt. Dabei wird die Oberkörperregion mit einer Menge von Shape-Contexts codiert. Deren Orientierung wird mit einem Nächster-Nachbar Verfahren vorausgesagt. Die Hypothesenbildung wird für jede Kameraansicht separat vorgenommen und diese Einzelhypothesen wer-

den mit einer Bayes-Filter Fusion zur endgültigen Schätzung der Oberkörperorientierung zusammengeführt.
Zur Evaluation beider Verfahren wurde ein Satz von Referenzparametern ermittelt mit denen das entwickelte System auf einer Reihe von Videosequenzen durchschnittlich eine gute Leistung aufweist. Für den SVM Ansatz wurde bei einer Winkelklassengröße von 18° insgesamt ein mittlerer Schätzfehler von $26,5°$ gemessen. Der Fehler des Nächster-Nachbar Verfahrens lag zwar mit $28,8°$ etwas darüber, dafür ist dieses Schätzverfahren insgesamt um ca. 30% schneller als der SVM Ansatz.
Beide Verfahren haben sich als robust gegenüber vereinzelten Segmentierungsfehlern herausgestellt. Trotzdem wäre eine Verbesserung der Segmentierungqualität wünschenswert, da dies die Genauigkeit der Schätzung anheben würde. Denkbar wäre dazu eine Kombination verschiedener Methoden. So könnten Kanten- und Farbinformationen als Merkmale für die Segmentierung herangezogen werden. Auch eine Auflösung gegenseitiger Verdeckungen von Personen sollte realisiert werden. Dies ist notwendig, damit das System in realistischen Szenarios, in denen sich mehrere Personen unkontrolliert bewegen können, eine gute Schätzleistung aufweisen kann.
Die Geschwindigkeit und Lesitung des Systems könnte durch das Einbinden einer Vorderkopf-Hinterkopf Klassifikation erhöht werden. Damit würde der Bereich der möglichen Orientierungswinkel um die Hälfte reduziert werden. Die Laufzeit des SVM Ansatzes könnte außerdem verbessert werden, wenn andere Merkmale niedriger Dimension verwendet werden, die sich schneller als HoSC-Merkmale berechnen lassen.

Literaturverzeichnis

[1] AGARWAL, A. ; TRIGGS, B. : 3D human pose from silhouettes by relevance vector regression. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 2, 2004, S. 882–888

[2] AGARWAL, A. : *Machine Learning for Image Based Motion Capture*, Institut National Polytechique de Grenoble, Diss., 2006

[3] BAYES, T. : An essay toward solving a problem in the doctrine of chance. In: *Philosophical Transactions of the Royal Society* 53 (1763), S. 370–418

[4] BERNARDIN, K. ; GEHRING, T. ; STIEFELHAGEN, R. : Multi- and Single View Multiperson Tracking for Smart Room Environments. In: *CLEAR Evaluation Workshop* (2006)

[5] BOLONGIE, S. : Shape Matching and Object Recognition Using Shape Contexts. In: *IEEE Transactions on Pattern Analysis and Machine Intellignece* 24 (2002), Nr. 24, S. 509–522

[6] BOSER, B. E. ; GUYON, I. ; VAPNIK, V. : A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, S. 144–152

[7] BRAND, M. : Shadow Puppetry. In: 1237-1244 (Hrsg.): *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999

[8] BRUNELLI, R. ; MICH, O. : On the use of histograms for image retrieval. In: *Proceedings of the IEEE International Conference on Multimedia Computing and Systems* Bd. 2, 1999, S. 143–147

[9] BUTLER, D. ; SIRDHARAN, S. ; BOVE JR., V. M.: Real-time adaptive background Segmentation. In: *Proceedings of the IEEE International Conference on Acustics, Speech & Signal Processing*, 2003, S. 341–344

[10] CHANG, C.-C. ; LIN, C.-J. : *LIBSVM: a library for support vector machines*, 2001. – Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[11] CUCCHIARA, R. ; GRANA, C. ; NERI, G. ; POCCARDI, M. ; PRATI, A. : The Sakobot System for Moving Object Detection and Tracking. In: *Proceedings of the ACM International Conference on Multimedia*, 2002, S. 223–226

[12] DELAMARRE, Q. ; FAUGERAS, O. : 3D articulated models and multi-view tracking with silhouettes. In: *Proceedings of the IEEE International Conference on Computer Vision* Bd. 2, 1999, S. 716–721

[13] DEUTSCHER, J. ; BLAKE, A. ; REID, I. : Articulated body motion capture by annealed particle filtering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Bd. 2, 2000, S. 126–133

[14] HARITAOGLU, I. ; HARWOOD, D. ; DAVIS, L. S.: W^4: Real-Time Surveillance of People and their Activities. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* Bd. 22, 2000, S. 809–830

[15] HORPRASERT, T. ; HARWOOD, D. ; DAVIS, L. S.: A Robust Background Subtraction and Shadow Detection. In: *Proceedings of the IEEE Asian Conference on Computer Vision*, 2000

[16] IWASAWA, S. ; OHYA, J. ; TAKAHASHI, K. ; SAKAGUCHI, T. ; EBIHARA, K. ; MORISHIMA, S. : Human body postures from trinocular Camera Images. In: *Proceedings of the 21st International Conference on Automatic Face and Gesture Recognition*, 2000, S. 326–331

[17] KNOOP, S. ; VACEK, S. ; DILLMANN, R. : Sensor fusion for 3D human body tracking with an articulated 3D body model. In: *Proceedings 2006 IEEE International Conference on Robotics and Automation*, 2006, S. 1686–1691

[18] KYUNGNAM, K. ; CHAILDABHONGSE, T. H. ; HARWOOD, D. ; DAVIS, L. S.: Real-time foreground-background segmentation using codebook model. In: *Real-Time Imaging* 11 (2005), Nr. 2

[19] LANZ, O. ; BRUNELLI, R. : An Appearance-based Particle filter for Visual Tracking in Smart Rooms. In: *Proceedings of the CLEAR'07 workshop*, 2007

[20] MIKIC, I. ; TRIVEDI, M. ; HUNTER, E. ; COSMAN, P. : Articulated body posture estimation from mulit-camera voxel data. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 1, 2001, S. 455–460

[21] MITTAL, A. ; ZHAO, L. ; DAVIS, L. S.: Human body pose estimation using silhouette shape analysis. In: *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003, S. 263–270

[22] MORI, G. ; BOLONGIE, S. ; MALIK, J. : Eficient Shape Matching Using Shape Contexts. In: *IEEE Transactions on Pattern Analysis and Machine Intellignece* 27 (2005), Nr. 11

[23] POPPE, R. ; POEL, M. : Comparison of silhouette shape descriptors for example-based human pose recovery. In: *7th International Conference on Automatic Face and Gesture Recognition* (2006), S. 541–546

[24] ROSALES, R. ; SCLAROFF, S. : Inferring Body Pose without Tracking Body Parts. In: *Proceedings of the IEEE Conference on Compuer Vision and Pattern Recognition* Bd. 2, 2000, S. 721–727

[25] SARLE, W. S.: *Neural Network FAQ.* `ftp://ftp.sas.com/pub/neural/FAQ.html`, 2002

[26] SCHÖLKOPF, B. ; SMOLA, A. J.: *Learning with Kernels.* MIT Press, 2002

[27] SIGGELKOW, S. : *Feature histograms for content-based image retrieval*, Albert-Ludwigs-Universität Freiburg, Fakultät für Angewandte Wissenschaften, Diss., 2002

[28] STAUFFER, C. ; GRIMSON, W. : Adaptive background mixture models for real-time tracking. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 2, 1999

[29] SUN, Y. ; BRAY, M. ; THAYANATHAN, A. ; TORR, B. : Regression-based human motion capture from voxel data. In: *British Machine Vision Conference*, 2006

[30] TIAN, Y.-L. ; LU, M. ; HAMPAPUR, A. : Robust and efficient foreground analysis for real-time video surveillance. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 1, 2005, S. 1182–1187

[31] VOIT, M. ; STIEFELHAGEN, R. : A Bayesian Approach for Multi-view Head Pose Estimation. In: *EEE Int. Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006

[32] WREN, C. R. ; AZARBAYEJANI, A. ; DARRELL, T. ; PENTLAND, A. P.: Pfinder: Real-Time Tracking of the Human Body. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997), Nr. 7

[33] WU, T.-F. ; LIN, C.-J. ; WENG, R. C.: Probability estimates for multi-class classification by pairwise coupling. In: *Journal of Machine Learning Research* 5 (2004), S. 975–1005

[34] ZIEGLER, J. ; NICKEL, K. ; STIEFELHAGEN, R. : Tracking of the Articulated Upper Body on Multi-View Stereo Image Sequences. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Bd. 1, 2006, S. 774–781

[35] ZIVKOVIC, Z. ; HEIJDEN, F. van d.: Efficient adaptive density estimation per image pixel for the task of background subtraction. In: *Pattern Recognition Letters* 27 (2006), Nr. 7, S. 773–780