Ein generisches System zur automatischen Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten

Zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

von der Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)
(Institut für Photogrammetrie und Fernerkundung)
genehmigte

Dissertation

von

Dipl.-Inform. Kai Jüngling

aus Adenau

Tag der mündlichen Prüfung: 24.01.2011

Referent: Prof. Dr.-Ing. Stefan Hinz
Korreferent: Prof. Dr. rer. nat. Maurus Tacke
Korreferent: Prof. Dr.-Ing. Christoph Stiller

Karlsruhe 2011

Kurzfassung

Eine wichtige Aufgabe im Bereich des maschinellen Sehens ist die *personenzentrierte Videoanalyse*. Diese findet in vielen Bereichen des heutigen Lebens, wie z.B. bei Fahrerassistenzsystemen, bei der Mensch-Maschine-Interaktion, militärischen Gefahrenerkennung und insbesondere auch visuellen Überwachung Anwendung.

Die Basis dieser personenzentrierten Analyse bildet die *Detektion und Verfolgung von Personen* in Videodaten. Diese ist Voraussetzung für alle folgenden Analyse- und Interpretationsschritte. Darüber hinaus ist auch die *Wiedererkennung von Personen* wichtiger Bestandteil vieler Anwendungen. So ist eine solche Wiedererkennung von Personen notwendig, wenn ein langer Zeitraum oder ein großer räumlicher Bereich betrachtet wird, da in diesem Fall Verbindungen zwischen den zeitlich oder räumlich nicht direkt zusammenhängenden Auftreten von Personen etabliert werden müssen. Ein typisches Beispiel hierfür ist die Überwachung großer öffentlicher Bereiche wie z.B. Flughäfen, bei der eine Vielzahl von Kameras vernetzt eingesetzt wird und typischerweise ein ausgedehnter Zeitraum relevant ist.

Aufgrund der Diversität der Anwendungsfälle für die Personendetektion, -verfolgung und -wiedererkennung ist es wünschenswert, ein generisches System zu entwickeln, das möglichst unabhängig von bestimmten Aspekten einzelner Anwendungsfälle und somit umfassend einsetzbar ist.

In dieser Arbeit wird ein solches System zur Personendetektion, -verfolgung und -wiedererkennung vorgestellt. Dieses System weist *Generizität* bzgl. verschiedener Aspekte auf. So ist das System unabhängig vom *Anwendungsszenario*, d.h. es werden keine Annahmen über die Anwendungsumgebung getroffen. So wird z.B. nicht vorausgesetzt, dass der Szenenhintergrund bekannt ist, oder dass weitere Informationen über die Szene vorliegen. Ebenso wird nicht angenommen, dass der aufzeichnende Sensor stationär ist, was bedeutet, dass das hier vorgestellte System auch bei bewegter Kamera einsetzbar ist. Gleichsam ist das System nicht auf eine bestimmte *Objektklasse* beschränkt, da außer Beispielen für das vollautomatische Training kein objektklassenspezifisches Wissen eingebracht wird. Darüber hinaus ist das System durch die ausschließliche Nutzung von auf Intensitätsgradienten basierenden lokalen Merkmalen weitestgehend *unabhängig vom verwendeten Sensor*. So ist das gesamte System sowohl im sichtbaren als auch im infraroten Spektralbereich anwendbar, da keine sensorspezifischen Merkmale wie Farbe oder Tiefe genutzt werden.

Die Systemgenerizität wird insbesondere durch ausschließliche Nutzung und Erweiterung des *Implicit Shape Model (ISM)* Ansatzes und *lokalen Bildmerkmalen* für alle drei Systemebenen erreicht. Diese sind dabei eng gekoppelt und verschmelzen zu einem integrierten Lösungsansatz. Für die *Personenverfolgung* wird eine Erweiterung des Implicit Shape Models vorgestellt, welche die Personendetektion und -verfolgung durch Kombination von bottom-up tracking-by-detection mit top-down modellbasierten Strategien vereinigt. Hierdurch wird eine Stabilisierung der Detektion sowie das automatische Tracking über Verdeckungssituationen erreicht. Ebenso werden separate Schritte und Heuristiken zur Datenassoziation, d.h. der Assoziation von Objekthypothesen über der Zeit, und Modellaktualisierung im Tracking überflüssig. Während der Verfolgung einer Person wird ein ISM-basiertes Identitätsmodell aufgebaut, welches zur *Wiedererkennung* der Person genutzt wird. Diese enge Kopplung von der Detektion bis zur Wiedererkennung macht das Gesamtsystem autark unter realen Bedingungen einsetzbar.

Abstract

An important area in computer vision is the *person-centered video analysis*. Applications cover many areas of today's life like driver assistance, human-machine-interaction, threat assessment in military context and specifically visual surveillance.

The basis of this person-centered analysis is *person detection and tracking* in video data. This is a precondition for all subsequent analysis or interpretation approaches. Moreover, *person reidentification* is a substantial component of many applications. Such a reidentification of persons is necessary in cases where a long time period or a large spatial area is considered. In these cases, connections between the occurrences of people that are not directly temporally or spatially connected are to be established. A typical example of this is the surveillance of large public spaces like airports where multiple networked cameras are utilised and a long time period is relevant.

Due to the diversity of application areas for person detection, tracking, and reidentification, it is desirable to develop a generic system that is most independent of certain aspects of application scenarios and thus universally applicable.

In this work, such a system for person detection, tracking and reidentification is introduced. This system is *generic regarding different aspects*. The system is independent of the *application scenario*, meaning that no assumptions on the application environment are made. For instance, it is not assumed that the scene background is known or other information regarding the scene is available. It is also not assumed that the recording sensor is stationary, which means the system introduced in this work is applicable in the case of a moving camera. Equally, the system is *not limited to certain object classes* since no object class specific knowledge other than a set of training samples is used. In addition, the system is mostly *independent of the used sensor* since no other than the intensity-gradient based local features are used. Thus, the overall system is applicable in the visible and the infrared spectral range since no features like color or depth are employed.

The system generality is specifically accomplished by the exclusive use of the *Implicit Shape Model* approach and *local image features* for all three system levels, whereby the levels are closely connected and merge in an integrated approach. For person tracking, an extension of the Implicit Shape Model, which combines bottom-up tracking-by-detection with top-down model-based strategies, is introduced. By that, a stabilisation of person detection and automatic tracking through short-term occlusion is accomplished. Likewise, separate steps and heuristics for data association, i.e the association of object hypotheses over time, and model update become redundant. During person tracking, an Implicit Shape Model based identity model, that is used for person reidentification, is established. By that tight coupling of all levels from detection to reidentification, the system is independently applicable under real conditions.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation und Anwendungsbereiche

Ein großes Anliegen des Menschen ist es, die eigenen Fähigkeiten auf Maschinen zu übertragen. Die Motivation einer solchen Nachbildung von menschlichen Fähigkeiten durch Maschinen ist offensichtlich – hierdurch ist es möglich, bestimmte Aufgaben, für die sonst menschliche Arbeitskraft eingesetzt werden müsste, von Maschinen übernehmen zu lassen. Wie schon durch die Bezeichnung deutlich wird, liegt eine solche Übertragung von menschlichen Fähigkeiten auch im Bereich des *maschinellen Sehens*[1] vor. Hier wird die menschliche Sehfähigkeit durch Sensoren nachgebildet bzw. erweitert[2]. Neben der automatischen Akquirierung von Sensordaten ist es natürlich auch wünschenswert, diese Daten automatisch auszuwerten, d.h. die Bild-, bzw. Videodaten hinsichtlich des Inhalts zu analysieren.

Die Anwendungsbereiche einer solchen *automatischen Videoanalyse* sind vielfältig. Neben der Nutzung in Bereichen wie der visuellen Navigation oder Sichtprüfung, in denen die beobachtete Szene selbst bzw. bestimmte Teilaspekte der Szene relevant sind, beschäftigt sich ein Großteil der Forschung im Bereich des maschinellen Sehens mit der *personenzentrierten Analyse*. In diesem Bereich stehen Personen und deren Verhalten im Mittelpunkt des Interesses.

Die Anwendungsbereiche für diese personenzentrierte Analyse sind breit gefächert und reichen von Fahrerassistenzsystemen und Mensch-Maschine-Interaktion über visuelle Überwachung bis zur Gefahrenerkennung im militärischen Kontext[3]. All diese Anwendungen haben gemeinsam, dass das Verhalten von Personen relevant ist. Die Aufgabe eines Videoanalysesystems ist es also, das Verhalten von Personen im jeweils relevanten Anwendungskontext zu erkennen und zu interpretieren.

Hierzu ist es unabhängig vom konkreten Anwendungskontext notwendig, eine stabile *Detektion*[4] und *Verfolgung (Tracking)*[5] von Personen in den Videodaten zu gewährleisten. D.h., dass die in der Szene sichtbaren Personen lokalisiert und während der Präsenz im Sichtbereich der Kamera verfolgt werden müssen. Die hierbei akquirierten Trajektorien können dann je nach Anwendungskontext Basis für weitere Verarbeitung sein oder direkt zur Analyse genutzt werden.

Eine solche direkte Nutzung der akquirierten Personentrajektorien[6] ist z.B. in Anwendungsbereichen

[1] Die Begriffe „maschinelles Sehen“ und „Computer Vision“ werden synonym genutzt

[2] Eine Erweiterung liegt z.B. bei Kameras für den infraroten Spektralbereich, Laser und Radar vor.

[3] Threat-Assessment.

[4] Auch Erkennung bzw. Objekterkennung. Hier wird der Begriff Detektion verwendet, um im Kontext von Personen eine eindeutige Abgrenzung zu „Erkennen bestimmter Personen“ sicherzustellen. Insbesondere ist hiermit die „Lokalisation“ von Personen gemeint.

[5] Die Begriffe Personen-, bzw. Objektverfolgung und Tracking werden in dieser Arbeit synonym benutzt.

[6] Das hier vorgestellte Verfahren liefert Bewegungstrajektorien in Bildkoordinaten. Je nach Anwendungskontext ist evtl. ein Transfer in ein globales Koordinatensystem notwendig.

wie der Fahrerassistenz gegeben, da hier im wesentlichen auf Gefahrensituationen die im Zusammenhang mit Fußgängern entstehen können hingewiesen werden soll. Hierzu reicht es aus, Fußgänger zu lokalisieren und bei Bewegung im Sichtbereich der Kamera zu verfolgen. Eine Interpretation, d.h. eine Prüfung, ob sich eine Person evtl. auf Kollisionskurs mit dem geführten Fahrzeug befindet, kann direkt auf Basis der durch die Personenverfolgung akquirierten Trajektorien durchgeführt werden. Ähnlich ist es bei der personenzentrierten Aktionserkennung, die im Rahmen der visuellen Überwachung oder militärischen Gefahrenerkennung durchgeführt wird. Hier reicht in manchen Fällen eine Betrachtung der Bewegung der Person aus, um z.B. festzustellen, dass eine Person einen bestimmten Bereich unerlaubt betreten hat. Oftmals wird hier allerdings auch weiteres Wissen über das Umfeld bzw. über Objekte im Umfeld der Person relevant.

In vielen Anwendungen reicht die reine Personenverfolgung bei zeitlich zusammenhängendem Auftreten einer Person im Sichtbereich einer einzelnen Kamera nicht aus, um die gewünschte Funktionalität zu bieten. Hier kann es notwendig werden, Verbindungen zwischen zeitlich oder räumlich nicht direkt verbundenen Auftreten einer Person zu etablieren. Dies ist dann der Fall, wenn ein ausgedehnter zeitlicher Abschnitt oder ein großer räumlicher Bereich betrachtet wird. In diesen Fällen wird eine *Wiedererkennung* von Personen notwendig, d.h. eine Person wird auf Basis ihrer Erscheinung als eine bereits zuvor gesehene Person identifiziert. Ein typisches Beispiel in dem die Personenwiedererkennung notwendig wird, ist die vernetzte Überwachung eines großen räumlichen Bereiches durch mehrere Kameras, wie sie z.B. an Flughäfen oder Bahnhöfen durchgeführt wird. In diesen Multi-Kamera-Szenarien reicht die separate Betrachtung der einzelnen Kameras nicht aus, um ein umfassendes Bild der Situation aufzubauen und somit eine Situationsanalyse durchführen zu können. Hierzu ist es notwendig, die Informationen der verschiedenen Kameras zusammenzuführen und in ihrer Gesamtheit zu interpretieren. Im Rahmen der personenzentrierten Analyse bedeutet dies, dass Verbindungen zwischen den Auftreten einer Person in unterschiedlichen Kameras etabliert werden müssen, sprich die „Tracks“ der Personen in den einzelnen Kameras müssen assoziiert werden, um die Trajektorie einer Person über mehrere Kameras hinweg bilden zu können. Durch diese *Personenwiedererkennung* ist es möglich, die Interpretation der Personenbewegung von einer auf mehrere Kameras und damit im besten Fall das gesamte überwachte Areal auszudehnen. Neben typischen sicherheitsrelevanten Anwendungen, wie z.B. der Feststellung auffälligen Verhaltens und der forensischen Analyse[7] gibt es für die Personenwiedererkennung auch andere Anwendungen. So kann die Wiedererkennung z.B. auch zur Suche vermisster Personen in einem größeren Areal genutzt werden.

Wie bei der Betrachtung eines großen räumlichen Bereichs durch mehrere Kameras, so ist es auch bei Betrachtung eines ausgedehnten zeitlichen Abschnitts nicht ausreichend, lediglich zeitlich direkt zusammenhängende Auftreten von Personen in Videosequenzen zu betrachten. Dies ist z.B. der Fall, wenn auffälliges Verhalten, welches in wiederholtem Aufsuchen eines bestimmten Orts bestehen kann, erkannt werden soll. Gleiches gilt z.B. für die Aufzeichnung von Kundenverhalten in Ladenlokalen oder allgemein für die Erstellung von Verhaltensprofilen. Ebenso kann eine Personenwiedererkennung in einer einzelnen mobilen Kamera schon bei Betrachtung eines kurzen Zeitraums notwendig werden. Hier können relevante Personen den Sichtbereich der Kamera bedingt durch die Kamerabewegung kurzzeitig verlassen und müssen beim Wiedereintreten wiedererkannt werden. Dies kommt besonders dann häufig vor, wenn eine Person von einem mobilen Sensor verfolgt wird, diese Person aber versucht, sich aus dem Sichtbereich des Sensors zu entfernen.

Die Anwendungsbereiche für eine *automatische Detektion, Verfolgung und Wiedererkennung von Personen* sind also vielfältig und breit gefächert. Aus diesem Grund ist es wünschenswert, ein System zu entwickeln, welches diese Aufgaben bearbeitet und in allen genannten Anwendungsbereichen einsetzbar ist.

[7] Eine „forensische Analyse“ beinhaltet z.B. die Nachverfolgung von Laufwegen nach bestimmten Ereignissen wie Anschlägen. Hierbei kann z.B. festgestellt werden, woher relevante Personen gekommen sind, wohin sie entkommen sind und in diesem Rahmen auch mit welchen anderen Personen bzw. Objekten sie Kontakt hatten.

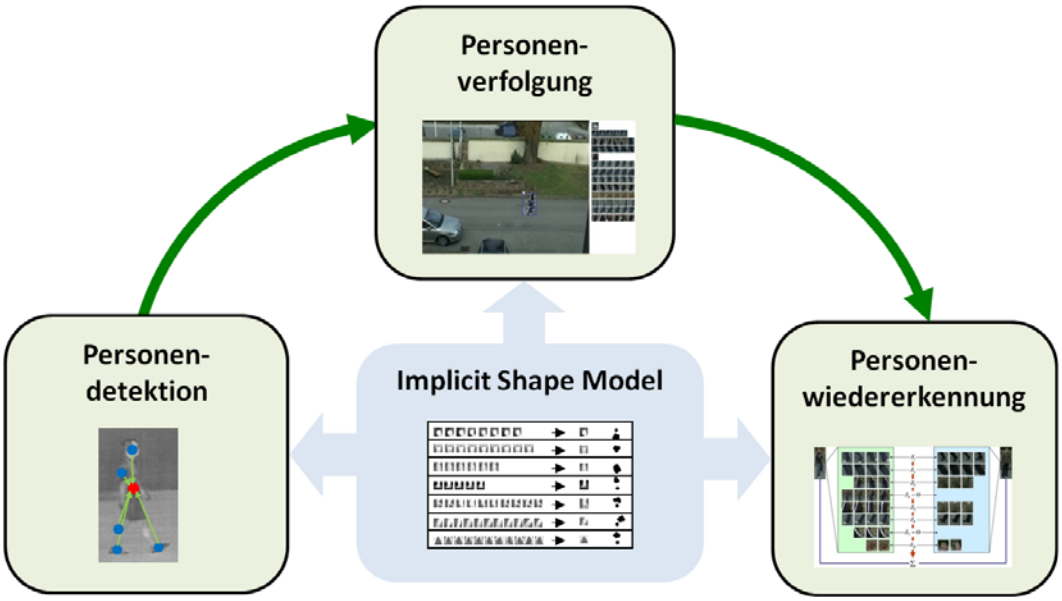


Abbildung 1.1: Skizzierung des Gesamtsystems. Die Teilbereiche Personendetektion-, verfolgung und -wiedererkennung bauen aufeinander auf. Insbesondere bildet das Implicit Shape Model sowie die hierfür in dieser Arbeit eingeführten Erweiterungen die Grundlage aller drei Teilbereiche.

1.2 Schwerpunkte und Beiträge der Arbeit

Aufgrund der Diversität der Anwendungsbereiche und den damit verbundenen unterschiedlichen Anforderungen ist es wünschenswert, ein System zu entwickeln, welches allgemein einsetzbar und nicht auf einen Anwendungsbereich beschränkt ist. Diese Arbeit widmet sich dieser Thematik und stellt einen generischen Ansatz zur automatischen *Detektion, Verfolgung und Wiedererkennung von Personen* in Videosequenzen vor[8].

Im Gegensatz zu bisherigen Arbeiten werden die drei Aufgabenbereiche als gemeinsame Problemstellung betrachtet und in einem integrierten Ansatz angegangen. Dazu wird ein für die Objektdetektion entwickelter Ansatz, das *Implicit Shape Model (ISM)*, zur Objektverfolgung erweitert. Hierbei wird ein neuer Ansatz zur Objektverfolgung, der Objektdetektion und -verfolgung integriert, vorgestellt. Durch die Kombination von bottom-up tracking-by-detection Methoden mit top-down modellbasiertem Tracking im Implicit Shape Model, werden zusätzliche Schritte und Heuristiken zur Datenassoziation und Modellaktualisierung bei der Personenverfolgung überflüssig. Insbesondere wird hierdurch gegenüber bestehenden Ansätzen ein automatisches Tracking bei kurzzeitigen Objektverdeckungen und eine Stabilisierung der Objektdetektion erreicht.

Während des Trackings werden Identitätsmodelle der Personen aufgebaut. Diese bestehen aus den lokalen Bildmerkmalen[9], die für Detektion und Tracking genutzt werden. Die Identitätsmodelle werden auch zur Wiedererkennung der Personen verwendet. Die Personenwiedererkennung ist also auch direkt in die Personenverfolgung integriert, und nutzt im Gegensatz zu den meisten bestehenden Ansätzen nicht lediglich die im Tracking akquirierte Position der Person, um Merkmale für die Wiedererkennung

[8] Das hier vorgestellte Verfahren zielt auf Szenarien ab, in denen die Personen anhand der Bilddaten als solche erkennbar sind. D.h. es werden hier keine aus großer Entfernung aufgezeichneten Luftaufnahmen oder Massenszenen von Personen (siehe [32, 78]) betrachtet. Das in dieser Arbeit vorgestellte System kann auf Ebene der Bildanalyse als komplementär zu solchen Systemen betrachtet werden.

[9] In dieser Arbeit werden SIFT (Scale Invariant Feature Transform) Merkmale genutzt.

zu berechnen. Diese Kopplung von Tracking und Wiedererkennung bedeutet insbesondere, dass das Gesamtsystem bis zur Wiedererkennung im Gegensatz zu den meisten bisher vorgestellten Ansätzen in realen Systemen einsetzbar ist, da auf keiner der Systemebenen von optimalen Ergebnissen der vorherigen Stufe ausgegangen wird oder sogar ganze Stufen durch manuelle Annotation ersetzt werden.

Wie in Abbildung 1.1 skizziert, stellt das Implicit Shape Model sowie die Erweiterungen die in dieser Arbeit hierfür eingeführt werden die Basis aller drei Teilbereiche des hier vorgestellten Systems dar. Durch die Tatsache, dass das gesamte System auf *einer* technischen Konzeption aufbaut, wird ein kompaktes, autarkes System entwickelt, welches leicht in andere Anwendungsbereiche transferierbar und auf andere Szenarien adaptierbar ist. Dies ist sehr wichtig, um die in Abschnitt 1.1 beschriebenen versatilen Anwendungsmöglichkeiten des Systems zu gewährleisten. Das hier vorgestellte System gewährleistet dies durch die ausschließliche Nutzung des ISM und lokaler Bildmerkmale. Ebenso wird kein anwendungsdomänen- oder szenariospezifisches Wissen, welches die Übertragbarkeit in andere Bereiche beschränken würde, in das System eingebracht. Konkret bedeutet dies, dass z.B. nicht angenommen wird, dass der aufzeichnende Sensor stationär oder der Szenenhintergrund bekannt ist. Dies macht das gesamte System bis hin zur Wiedererkennung bei bewegtem Sensor einsetzbar. Gleichsam werden keine Annahmen über spezielle Eigenschaften des genutzten Sensors gemacht, wie dies in vielen Systemen der Fall ist. So wird z.B. nicht vorausgesetzt, dass der Sensor Farbinformation liefert. Für alle Systemebenen werden lediglich die auf Intensitätsgradienten basierenden lokalen Bildmerkmale genutzt. Hierdurch wird gewährleistet, dass das System weitestgehend unabhängig von der Sensormodalität nutzbar und somit ohne jegliche Adaptionen sowohl im infraroten als auch im sichtbaren Spektralbereich einsetzbar ist. Das System wird im besonderen Hinblick auf die Verwendung im Kontext der personenzentrierten Analyse entwickelt, macht aber keine speziellen Annahmen über die erwartete Objektklasse. Somit ist das System auch für andere Objektklassen einsetzbar. Hierzu müssen lediglich die Trainingsbeispiele für die betreffende Klasse gewählt werden. Das Gesamtsystem ist somit generisch einsetzbar und weist Generizität bzgl. der Aspekte *Szenario*, *Objektklasse* und *Sensor* auf.

Neben den wesentlichen Beiträgen dieser Arbeit – dem *generischen Gesamtsystem* und dem *integrierten Implicit Shape Model basierten Ansatz* für alle Verarbeitungsebenen, der die einzelnen Problemstellungen in einem Gesamtlösungsansatz ausschließlich auf Basis des ISM und SIFT-Merkmalen angeht, liegen die wissenschaftlichen Beiträge dieser Arbeit auch auf den einzelnen Ebenen:

Personendetektion

1. Erweiterung des ISM zur Hypothesendisjunktheit.
2. Nutzung des ISM zur hierarchischen Objektdetektion und Objektkomponentenklassifikation.

Personenverfolgung

1. Kopplung von Objektdetektion und -verfolgung durch Weiterentwicklung des ISM.
2. Kombination von tracking-by-detection mit modellbasiertem Tracking in neuem Ansatz.
3. Objektverfolgung bei starker Kamerabewegung durch objektspezifische Bewegungskompensation.

Personenwiedererkennung

1. Online-Aufbau von Personenmodellen zur Wiedererkennung.
2. Mehrstufiger Ansatz zur effizienten Personenwiedererkennung.
3. Personenwiedererkennung auf Basis von erweiterten Codebuchsignaturen.
4. Ansichtsbestimmung und -transformation von Personenmodellen zur Wiedererkennung.
5. Erscheinungsbasierte Personenwiedererkennung im infraroten Spektralbereich.

1.3 Gliederung

Die Einteilung der Arbeit folgt dem Aufbau des hier vorgestellten Systems:

Zunächst wird in Kapitel 2 ein Überblick über den aktuellen Stand der Forschung in den drei relevanten Teilbereichen dieser Arbeit gegeben. Zusätzlich erfolgt eine Einordnung und eine Diskussion der Beiträge dieser Arbeit.

In Kapitel 3 erfolgt die Einführung in das Implicit Shape Model, auf dem diese Arbeit wesentlich aufgebaut ist. Dazu erfolgt zunächst eine Einführung in lokale Bildmerkmale sowie eine detaillierte Vorstellung der hier verwendeten SIFT-Merkmale. Im weiteren werden die Grundlagen des zur Objektdetektion in Einzelbildern entwickelten Verfahrens sowie die Adaptionen und Erweiterungen, die für die Einzelbilddetektion vorgenommen wurden, eingeführt.

Kapitel 4 stellt den ersten Hauptblock dieser Arbeit – die Erweiterung des Implicit Shape Models zur Objektverfolgung – vor. Hier wird das generische Trackingverfahren vorgestellt, welches eine Kombination von modellbasiertem Tracking und tracking-by-detection einführt und in verschiedenen Anwendungsbereichen evaluiert.

In Kapitel 5 wird der zweite Hauptblock dieser Arbeit – die Implicit Shape Model basierte Personenwiedererkennung – vorgestellt.

In Kapitel 6 erfolgt eine abschließende experimentelle Validierung des Gesamtsystem. Kapitel 7 fasst die Arbeit zusammen und gibt einen Ausblick auf mögliche Erweiterungen und Folgearbeiten.

Kapitel 2

Stand der Forschung und Einordnung der Arbeit

2.1 Personendetektion und -verfolgung

Eine der grundlegendsten Aufgaben im Bereich des maschinellen Sehens ist die *Detektion* und *Verfolgung* von Objekten in Bildfolgen[1].

Viele der frühen [165, 75], sowie einige der heutigen Ansätze zur Objektdetektion und -verfolgung [21, 94, 100, 63], bauen darauf auf, die für die Auswertung relevanten Objekte als Vordergrundobjekte durch Hintergrundsubtraktion [151] (im infraroten Spektralbereich [48, 39], bzw. durch Fusion des infraroten und sichtbaren Spektrums [46, 112]), bzw. Bewegungssegmentierung zu detektieren. Diese Verfahren vereinfachen von der Realität, indem sie annehmen, dass die relevanten Objekte die einzigen Vordergrundobjekte in der Szene sind, oder, dass die relevanten Objekte durch nachgeschaltete Klassifikation determiniert werden können. Insbesondere kann in der Realität auch nicht immer davon ausgegangen werden, dass der Hintergrund der Szene bekannt ist. Ebenso beschränken diese Verfahren die Anwendbarkeit auf Szenarien mit statischen Kameras (wobei auch Bewegungssegmentierungsverfahren für bewegte Kameras existieren [25, 142]).

Im Gegensatz zu diesen rein datengetriebenen Ansätzen verfolgt ein anderer Zweig der Objektverfolgung den modellgetriebenen Ansatz. Hier wird ein Objekt auf Basis eines bekannten Musters (Templates) explizit in Bildern gesucht und dadurch verfolgt [12]. Das Template wird durch manuelle Auswahl im Bild oder durch explizite Modellierung definiert. Das Template kann dabei verschiedene Ausprägungen, von einfachen Pixelmasken bis hin zu komplexeren Beschreibungen, haben. Ein Hauptproblem dieser Ansätze ist die Aktualisierung des Modells (des Templates) bei der Objektverfolgung [119]. Ein Ansatz, der diese Art von Tracking durch Auswahl einer „Region of Interest (ROI)“ auf Basis von SIFT und Farbmerkmalen durch eine Mean-Shift Suche durchführt, wird in [181] vorgestellt.

Objektdetektion

Eine ähnliche Herangehensweise an die Problemstellung findet man bei aktuellen Objektdetektoren. Auch hier wird ein Modell zum Auffinden von Objekten in Bildern genutzt, allerdings ist hier nicht eine bestimmte Objektinstanz, sondern alle Instanzen einer bestimmten Objektklasse relevant. Die Aufgabenstellung für diese Objektdetektoren ist also, alle Instanzen einer Objektklasse in einem Eingabebild

[1] Aufgrund der großen Menge an Literatur allein im Bereich der Personenverfolgung kann hier kein umfassender Überblick gegeben werden. Hierfür sei auf [51, 174, 80] verwiesen.

zu detektieren.

In diesem Bereich der automatischen Objektdetektion wurden in jüngster Zeit große Fortschritte gemacht. Viele dieser heutigen Objekterkennungsverfahren [135, 146, 120, 178] bauen auf lokalen Bildmerkmalen auf. So geht der Fortschritt bei der Objekterkennung mit dem Fortschritt bei lokalen Bildmerkmalen einher.

Einer der ersten vielversprechenden lernbasierten Ansätze wurde von Viola und Jones in [160] vorgestellt. Dieser basiert auf *Haar-like features*[2] und einem Trainingsschritt, in dem durch den AdaBoost-Algorithmus [64] ein Klassifikator trainiert wird, der viele, auf einzelnen einfachen Merkmalen basierende Klassifikatoren in einer Kaskadenstruktur kombiniert. „Haar-like features“ sind in diesem Zusammenhang einfache Merkmale, die Pixeldifferenzen in rechteckigen Bildbereichen berechnen. Diese können durch Nutzung des Integralbildes[3] schnell berechnet werden. Durch Kombination vieler auf diesen einfachen Merkmalen basierender Klassifikatoren, sowie durch Auswahl der besten Klassifikatoren für die jeweiligen Objektklassen durch den AdaBoost-Algorithmus, führt dieser Ansatz zu einem stabilen, echtzeitfähigen Objektdetektionsverfahren. Der initiale Ansatz, der speziell auf die Gesichtsdetektion abzielt, wird in [161] speziell zur Nutzung zur Personendetektion über Einzelbilder hinaus erweitert. Hierbei werden zusätzlich zu den erscheinungsbasierten Merkmalen, Bewegungsmerkmale einbezogen, die auf aufeinanderfolgenden Bildern berechnet werden. Diese werden ebenfalls durch einfache Rechteckmerkmale dargestellt und zielen darauf ab, die Detektion bei bewegten Objekten durch diese Zusatzinformation zu verbessern.

Ein heute oft verwendeter Ansatz zur Personendetektion wurde von Dalal [43] 2005 entwickelt. Dieser baut auf dem *Histogram of oriented Gradients (HOG)* auf. Die sogenannten *HOG-Merkmale* stellen eine Beschreibung eines Bildausschnitts durch Gradientenorientierungen dar. Der wesentliche Unterschied zu den in dieser Arbeit verwendeten *SIFT-Merkmalen* [117] (siehe Abschnitt 3.1.1) besteht darin, dass HOG-Merkmale nicht auf Schlüsselpunkten aufbauen, d.h. nur in Bereichen von im Bild gefundenen Schlüsselpunkten berechnet werden, sondern diese in einer sogenannten *dichten Abtastung* berechnet werden. Dies bedeutet, dass im gesamten Eingabebild Merkmale in einem festen, vorgegebenen Raster berechnet werden. In Fall von HOG werden die Merkmale durch „Blocks“ definiert, die die Größe des Bildbereichs angeben, in dem ein Orientierungshistogramm gebildet wird. Die Blocks, und damit die HOG-Deskriptoren überschneiden sich in ihrem Ortsbereich. Zur Objektdetektion auf Basis von HOG-Merkmalen werden *Sliding-Window* Ansätze genutzt. Hierbei wird zur Detektion ein Klassifikatorfenster über das aus HOG-Merkmalen bestehende Bild geschoben. Das Seitenverhältnis muss dabei den erwarteten Objektausmaßen entsprechen. Um Skalierungsinvarianz zu gewährleisten, muss dies auf verschiedenen Skalierungsstufen durchgeführt werden. Die Klassifikation als Objekt oder nicht Objekt erfolgt auf Basis der in dem Fenster enthaltenen Merkmale. Als Klassifikator werden Kernel-Methoden [154], zumeist *Support Vector Machines (SVM)* genutzt. Seit der Einführung von HOG-Merkmalen und dem damit verbundenen Ansatz zur Objektdetektion hat es diverse Weiterentwicklungen dieses Ansatzes gegeben. Viele davon [98, 177, 140] zielen auf Verbesserungen bei der Geschwindigkeit des Algorithmus ab, wobei auch Weiterentwicklungen [35] des grundsätzlichen Ansatzes erfolgen. Eine ähnliche Erweiterung wie die in [161] für den Viola-Jones Detektor, wird in [44] für den HOG-Detektor vorgestellt. Dieser wird um eine zeitliche Komponente erweitert um die Bewegung in Videosequenzen einbeziehen zu können. Hierzu wird der optische Fluss auf Basis aufeinanderfolgender Bilder berechnet. Auf Basis des Flussbildes werden dann HOG-Merkmale berechnet, die wiederum zur Personendetektion genutzt werden.

Ein weiterer heute häufig verwendeter Ansatz zur Objektdetektion ist der, ein Objekt als *Zusammenschluss von Einzelkomponenten* zu betrachten. Die einzelnen Teile des Objekts können dann unabhängig voneinander detektiert werden und im Nachhinein zu dem Gesamtobjekt zusammengesetzt werden. Zur Beschreibung werden häufig deformierbare Teil-Modelle [62, 58] eingesetzt, die auch für

[2]Dies sind auf Haar Filtern aufbauende Merkmale, beziehen aber komplexere Funktionen ein und werden deshalb als „Haar-like“ bezeichnet.

[3]Das Integralbild ist ein Algorithmus zur schnellen Berechnung von Pixelsummen in einem rechteckigen Bildausschnitt.

artikulierte Objekte geeignet sind. In [56, 57] wird ein Ansatz vorgestellt, der diese Modelle zur Personendetektion nutzt. Zur Detektion der Einzelkomponenten wird jeweils ein HOG-Detektor eingesetzt. In [166] wird der Boosting-Ansatz zum Training von separaten Detektoren für einzelne Köperpartien auf Basis von „Edgelet“ Merkmalen[4] genutzt. Die Ergebnisse der Teildetektoren werden in einem Maximum-Likelihood-Modell vereinigt, was es ermöglicht, auch mögliche Verdeckungen zwischen Personen mit in die Modellannahme einzubeziehen. Weitere Beispiele für die Personendetektion durch die Detektion von Einzelkomponenten sind die Arbeiten von Mikolajczyk [125, 121] und Ferrari [60].

Bei *Implicit Shape Model (ISM)*[5] basierten Objektdetektionsansätzen ist diese Teil-Ganzes Struktur schon implizit im Modell enthalten. Dieses Verfahren zur Objektdetektion baut auf schlüsselpunktbasierten lokalen Merkmalen auf. Diese Merkmale werden nicht wie bei HOG in einem vorgegebenen Raster im Bild berechnet, sondern lediglich an bestimmten Bildpositionen, die durch *Schlüsselpunkte*[6], für die angenommen werden kann, dass sie unter verschiedenen Bildtransformationen stabil wiedergefunden werden können, gegeben sind. In einem Trainingsschritt wird auf Basis der in Trainingsdaten extrahierten lokalen Merkmale das Implicit Shape Model, das die Objektklasse durch die Ortsverteilung lokaler Merkmale beschreibt, generiert. Dieses ISM wird in Form eines Codebuchs, welches zur Detektion von Objekten in Eingabebildern genutzt wird, abgelegt. Zur Objektdektion wird ein kontinuierlicher Hough-Raum [13] durch die im Codebuch abgelegte Ortsverteilung der Merkmale aufgebaut. In diesem erfolgt die Suche nach Objekthypothesen durch eine Mean-Shift-Mode-Estimation [36].

Das Verfahren wurde initial von Bastian Leibe [106, 104, 107] vorgestellt und in [111] speziell zur Personendetektion in komplexen Szenen genutzt. Dort wird zusätzlich ein Ansatz vorgestellt, der einen Verifikationsschritt auf Basis einer Kombination von Chamfer Matching [14] und Segmentierung durchführt. Für das initiale Verfahren gibt es seit Einführung verschiedene Erweiterungen. In [105] wird ein Ansatz vorgestellt, wie die implizit im ISM vorhandene Information zu Objektteilen zur Klassifikation von semantischen Objektkomponenten genutzt werden kann. Insbesondere wird hier vorgestellt, wie semantische Information durch einen Lernansatz in das Modell eingebracht und zur Verbesserung der Objektdetektion in einem Verifikationsschritt genutzt werden kann.

In [147] wird eine Erweiterung des ISM vorgeschlagen, die spezifisches Vorwissen über artikulierte Objekte einbringt und hierdurch die Detektionsperformance verbessert. In [145] wird dieser Ansatz erweitert, indem eine Technik zum Aufbau von Instanzmodellen vorgestellt wird. Diese werden dazu genutzt, bestimmte Objektinstanzen in Eingabebildern auf Basis der vorher für dieses Objekt gesehenen Merkmalskonfiguration zu detektieren. In [102, 101] wird durch das *Principled Implicit Shape Model (PRISM)* eine neue theoretische Grundlage für die probabilistische Formulierung des Hough-Voting Ansatzes eingeführt. Die Neuformulierung der probabilistischen Grundlagen basiert auf einer Interpretation des Hough-Votings in Kombination mit linearer Sliding-Window Detektion. Hierdurch wird eine solide probabilistische Begründung des Hough-Voting Ansatzes gegeben. In [66] wird das Prinzip der Nutzung einer generalisierten Hough-Transformation weiterentwickelt. Hier werden die Codebücher zur Modellierung der Objektklassen durch sogenannte „Hough Forests“[7] ersetzt. In [155] wird eine Erweiterung des ISM für multiple Ansichten vorgestellt. Dazu wird das ISM mit dem Multi-View System von Ferrari [61] kombiniert. Hierbei werden zwar einzelne Codebücher für verschiedene Ansichten einer Objektklasse erstellt, diese werden aber in einem gemeinsamen Modell auf Ebene der einzelnen Codebucheinträge verbunden. Zur Detektion werden die Verbindungen zwischen den einzelnen Einträgen dazu genutzt, Codebuchaktivierungen zwischen unterschiedlichen Ansichten auszutauschen. Somit werden Votes von Merkmalen, die aus verschiedenen Trainingsansichten stammen zusammengeführt und ermöglichen es, Objekte in Zwischenansichten der Trainingsansichten korrekt zu detektieren.

[4] An der Objektsilhouette orientierte Merkmale.
[5] Das ISM bildet die Grundlage dieser Arbeit und wird daher in den nachfolgenden Kapiteln detailliert beschrieben.
[6] Keypoint, Interest point
[7] Hough-Forests sind vergleichbar mit Random-Forests, die aus einer Anzahl binärer Entscheidungsbäume [141] bestehen.

Objektverfolgung

Die immensen Fortschritte im Bereich der automatischen Objektdetektion haben dazu geführt, dass viele aktuelle Trackingansätze auf diesen dedizierten Objektdetektoren aufbauen. Die meisten dieser Ansätze sind tracking-by-detection Ansätze, die den Objektdetektor dazu nutzen, Objekte der relevanten Klassen(n), unabhängig voneinander in Einzelbildern zu detektieren und im Nachgang dazu das Tracking durch Spurbildung auf Basis der Einzeldetektionen durchzuführen. Hierbei kann zwischen solchen Ansätzen, die eine globale Optimierung der Spuren auf Basis der Gesamtinformation der Sequenz durchführen [5, 81, 110] und schritthaltenden Ansätzen, die lediglich die Information bis zum aktuellen Zeitpunkt zur Spurbildung ausnutzen [167, 29, 31, 52, 109], unterschieden werden.

Die meisten dieser Ansätze nutzen Farbe zur Datenassoziation bei der Track-Formierung. Hierzu wird häufig ein Farbhistogramm im Bereich der durch die Detektion definierten bounding box[8] berechnet. Einige Ansätze, insbesondere solche, die für den mobilen Bereich ausgelegt sind, nutzen weitere Sensoren wie Stereo-Kameras [52, 69, 129, 53, 67].

Andere Ansätze nutzen nicht nur Sensor-, sondern auch Objektklassenspezifika bei Detektion und Tracking aus. Dabei werden Information über den Aufbau von Objekten direkt zur Durchführung, bzw. Verbesserung des Trackings genutzt. Andriluka stellt hierzu einen auf dem ISM basierenden Ansatz [5] vor, der Wissen über den Gehzyklus einer Person ausnutzt, um die Position der Person vorherzusagen und die Detektion zu steuern. Voraussetzung dieser Vorgehensweise ist die Annotation der Trainingsdaten mit Körperteilsemantik. Auf Basis dieser Informationen wird Vorwissen über mögliche Artikulationen eingebunden und zur zeitlichen Modellierung des Gehzyklus genutzt. Die Modellierung findet dabei durch *hierarchical Gaussian process latent variable model (hGPLVM)* statt. In [6] wird dieses Modell erweitert um ein 3D-Tracking zu ermöglichen. Hierbei wird der ISM-basierte Detektor aus [5] durch einen auf *Pictorial Structures* [58, 57] basierten ersetzt.

Gammeter [67] stellt einen Ansatz zur mobilen Personenverfolgung vor, der die Artikulationen von Personen durch die in einem Trainingsschritt bestimmten Körperdynamiken auf Basis der durch ein Tracking auf Objektebene bestimmten Personensilhouetten rekonstruiert. Durch einen Artikulationstracker kann die Pose der Person rekonstruiert werden und gleichzeitig eine körperteilgenaue Prädiktion der Person zum Tracking erfolgen.

Die Herangehensweise von tracking-by-detection Ansätzen, die Detektion unabhängig vom Tracking durchzuführen ist nicht optimal, da somit bei der Detektion keine Informationen aus dem Tracking genutzt werden. Diese können (wie in dieser Arbeit gezeigt wird) aber die Detektion an sich verbessern. Einige Ansätze widmen sich der Thematik und verbinden Detektion und Tracking:

In [147, 145] wird ein Verfahren, welches primär zur Objektdetektion gedacht ist, vorgestellt. Hierbei wird das ISM auf unterschiedliche Arten erweitert. So wird in [147] eine Erweiterung vorgeschlagen, die spezifisches Vorwissen über artikulierte Objekte einbringt und hierdurch die Detektionsperformance verbessert. In [145] wird dieser Ansatz erweitert, indem eine Technik zum online Aufbau von Instanzmodellen vorgestellt wird. Diese werden dazu genutzt, bestimmte Objektinstanzen (bestimmte Personen) in Eingabebildern auf Basis der vorher für dieses Objekt gesehenen Merkmalskonfiguration zu detektieren. Es erfolgt kein Tracking der Personen, wobei angenommen wird, dass der Ansatz dazu geeignet ist, Personen über kurzzeitige Verdeckungen hinweg zu verfolgen.

Leibe stellt in [110] eine Tracking-Strategie vor, die Objektdetektion und Spurbildung zum Tracking als verbundenes Optimierungsproblem betrachtet, um den in der Realität vorhandenen Abhängigkeiten zwischen Tracking und Detektion Rechnung zu tragen. Die aus dem Tracking entstandenen Trajektorien werden dabei zur Steuerung des ISM-basierten Personendetektors in folgenden Frames verwendet. Die globale Optimierung von Detektion und Tracking findet durch Formulierung als kombiniertes „Quadratisch Boolsches Problem (QBP)“ statt, wodurch eine optimale Lösung unter Einbezug

[8] Die „bounding box“ definiert die Objektausmaße durch ein das Objekt umschließendes Rechteck. Im Fall der Objektdetektion definiert diese bounding box die erkannte Position und Fläche des Objekts.

beider Aspekte ermöglicht wird. Tatsächlich erfolgt hier keine direkte Kopplung von Detektion und Tracking, d.h. das Wissen aus dem Tracking wird nicht direkt in die Detektion eingebracht. In [109] wird dieser Ansatz zur Anwendung zum Tracking von Personen und Fahrzeugen von einer mobilen Kamera erweitert.

Wu stellt in [167] einen Objektkomponenten-basierten Ansatz zur Personendetektionen und Verfolgung vor. Dieser detektiert zunächst Körperteile auf Basis von Edgelets und kombiniert die Einzeldetektionen dann, um die Wahrscheinlichkeit für das Vorhandensein einer Person zu bestimmen. Das Tracking erfolgt auf Basis einer Kombination des Zusammenfügens der Einzelbilddetektionen und einer gezielten Mean-Shift Suche nach Personen. Hier wird tracking-by-detection mit modellbasiertem Tracking bestimmter Individuen also insofern verbunden, als dass die für die aktuelle Situation am besten geeignete Methode durch Heuristiken ausgewählt wird.

Personendetektion und -verfolgung im infraroten Spektralbereich

Viele der Fortschritte im Bereich der Objektdetektion im sichtbaren Spektralbereich können aufgrund der Beschaffenheit der Verfahren direkt in den infraroten Spektralbereich übernommen werden (siehe Zhang et al. [113]). Ein Beispiel hierfür sind HOG-basierte Ansätze [43], die in [152] zur Personendetektion in Infrarot genutzt wurden. Grundsätzlich besteht eine solche Transfermöglichkeit im Bereich der reinen Objektdetektion prinzipiell für alle Verfahren [160, 167, 107, 43, 57], die zur Merkmalsberechnung nur Grauwerte nutzen. Da es aufgrund der speziellen Gegebenheiten im infraroten Spektralbereich prinzipielle Unterschiede zum sichtbaren Spektralbereich gibt (siehe [54]), werden viele Methoden auch speziell für den infraroten Spektralbereich entwickelt: So nutzen Nanda und Davis [131] einen Template-basierten Ansatz zur Personendetektion in Infrarotdaten. Dieser wird von Davis und Keck [47] zu einem zweistufigen Ansatz erweitert, der in einem zweiten Schritt die durch Template-Matching bestimmten Personenkandidaten mit einem AdaBoost-Klassifikator verifiziert. Bertozzi et al. [26] detektieren Personen bei bewegter Kamera durch explizite Lokalisation von symmetrischen Objekten mit einer bestimmten Größe und einem bestimmten Seitenverhältnis. Fang et al. [169] stellen eine formunabhängige Methode zur Personendetektion vor, die eine Klassifikation auf Basis von Kontrastmerkmalen durchführt.

Im Bereich der Objektverfolgung sind die meisten Verfahren nicht einfach aus dem sichtbaren in den infraroten Spektralbereich übertragbar. Dies liegt daran, dass viele Verfahren Farbmerkmale zur Objektverfolgung nutzen und damit auf die Anwendbarkeit im sichtbaren Spektralbereich beschränkt sind. Allerdings gibt es auch Trackingverfahren, die sich auf den infraroten Spektralbereich konzentrieren und speziell für die Gegebenheiten hier entwickelt wurden:

So stellen Dai et al. [42] eine Methode zur Personenverfolgung im infraroten Spektralbereich vor. Hier wird ein generalisierter EM-Algorithmus verwendet, um eine Separierung des Vorder- vom Hintergrund vorzunehmen. Vordergrundobjekte werden hier als Personen deklariert. Ein einfacher Ansatz, der die Infrarotspezifika, nämlich dass eine Person im Gegensatz zum Hintergrund heller[9] erscheint, ausnutzt, wird von Yasuno et al. [173] vorgestellt. Personen werden hier durch die Lokalisation heller Regionen detektiert und dann durch ein Prädiktionsverfahren erster Ordnung verfolgt. Eine andere Möglichkeit besteht darin, die Schwierigkeiten im infraroten Spektralbereich durch Kombination mit Tracking im sichtbaren Bereich zu umgehen [46], [112]. Junfeng et al. [65] stellen einen dreistufigen Ansatz zur Echtzeit-Personenverfolgung in Infrarot vor, der auch bei bewegter Kamera anwendbar ist. Die drei Module *Region of Interest (ROI) Generierung*, *Objektklassifikation* und *Tracking* sind dabei in einer Kaskade angeordnet wobei jede Stufe komplementäre visuelle Merkmale verwendet. Ein Ansatz, der ebenfalls bei bewegter Kamera anwendbar ist wird von Xu und Fujimura [59] vorgestellt. Hier wird das Tracking auf einem SVM-basierten Personenklassifikator aufgebaut. Das Tracking wird durch eine Kombination von Kalman-Filter-Prädiktion mit Mean-Shift-Tracking durchgeführt.

[9] Oder dunkler, je nach Interpretation der Sensordaten.

2.2 Personenwiedererkennung

Die Aufgabe der Personenwiedererkennung besteht darin, eine Person, nachdem sie den Sichtbereich einer Kamera verlassen hat, beim Eintreten in den Sichtbereich der gleichen oder einer anderen Kamera in Kameranetzen als die gleiche Person wiederzuerkennen. Zur Bewältigung dieser Aufgabe sind prinzipiell alle Methoden geeignet, die eine Identitätsmodellierung einer Person anhand von Videodaten vornehmen.

Typische Verfahren in diesem Bereich sind die *biometrischen* Verfahren [164, 95], die insbesondere zur Identifizierung von Person, d.h. zur eindeutigen Bestimmung der Identität in großen Datenbasen ausgelegt sind. Die gebräuchlichsten Verfahren in diesem Bereich sind die Gesichtserkennung [168, 138, 34, 179, 171], die schon seit mehreren Jahrzehnten ein aktives Forschungsgebiet ist [157, 19], sowie die Iris Erkennung [45, 27, 149, 71], die besonders distinktiv ist [175]. Weitere biometrische Merkmale, die aber lediglich im absoluten Nahbereich nutzbar sind, sind *Netzhaut*, *Fingerabdruck*, *Handgeometrie* und *Handgefäßstruktur (Palmprint)*.

Gemeinsam ist all diesen Verfahren, dass sie zwar sehr distinktiv in ihrer Modellierung der Person sind, d.h. sie sind dazu geeignet, eine bestimmte Person in sehr großen Datenbeständen zu identifizieren, allerdings machen sie auch starke Annahmen über die vorhandenen Sensorinformationen. D.h., eine stabile Gesichtserkennung setzt voraus, dass das Gesicht mehr oder weniger frontal in ausreichend hoher Auflösung vorliegt. Ebenso muss zur Iriserkennung die Iris der Person in einem Maximalabstand zum Sensor sichtbar sein, was in der Praxis nur in kooperativen Umgebungen[10] gewährleistet werden kann.

Diese Verfahren sind somit nicht zur Wiedererkennung in typischen Überwachungsszenarien, in denen die Personen unter unbekannten Umgebungsbedingungen häufig nur in niedriger Auflösung zu sehen sind und in denen die Personen insbesondere nicht-kooperativ sind, nutzbar.

Ein biometrisches Verfahren, welches für diese Bedingungen ausgelegt ist, ist die *Gangerkennung (Gait-Recognition)*, bzw *Ganganalyse (Gait-Analysis)* [92, 143, 132, 115, 28]. Die Gangerkennung modelliert die Eigenarten des menschlichen Gangs und nutzt diese zur Identifikation von Personen. Da Ganganalyseverfahren auch in den hier zur experimentellen Validierung genutzten Datensätzen genutzt werden können, erfolgt in Kapitel 5 ein Vergleich mit den Ganganalyseverfahren von Tan et al. [153] und Wang et al. [162].

Typischerweise werden zur Wiedererkennung von Personen in Überwachungsszenarien aber Verfahren eingesetzt, welche die Erscheinung einer Person anhand visueller Merkmale modellieren.

Viele Verfahren bauen dabei auf *Farbe* als primären Merkmal zur Wiedererkennung von Personen auf. Hierbei wird die Farbverteilung auf dem Abbild der Person häufig in Form von Farbhistogrammen zur Wiedererkennung genutzt [137, 85, 136]. Einige weiterentwickelte Verfahren [40, 176] gehen einen Schritt über die Modellierung durch ein Histogramm hinaus und generieren distinktivere Modelle durch Hinzunahme von unter anderem der Position. Hierzu wird in [156] ein „color-position histogram“ vorgestellt.

Unabhängig davon, in welcher Form Farbe zur Wiedererkennung genutzt wird, bestehen in der Praxis zwei wesentliche Probleme. Zunächst ist die Annahme, dass Farbe tatsächlich ein distinktives Merkmal zur Personenunterscheidung ist, in der Praxis häufig nicht gegeben. So wird bei Betrachtung typischer Überwachungsszenarien [3] deutlich, dass die Kleidung von Personen, die eindeutig auch die Farbe einer Person definiert, in der Realität häufig sehr ähnlich ist und meistens in einem bestimmten Bereich liegt (z.B. tragen die meisten Leute im Winter eher dunkle Kleidung). In der Praxis ist die Farbe, insbesondere in Überwachungsszenarien, in denen Personen häufig nur in geringer Auflösung in den Videodaten auftreten und geringe Farbnuancen somit nicht zu unterscheiden sind, also kein wirklich distinktives Merkmal. Die zweite Schwierigkeit bei der Nutzung von Farbe ist, dass diese nicht

[10] Kooperative Umgebungen sind solche, bei denen der Mensch bei der Datenaufzeichnung mit dem System kooperiert, d.h. z.B. gezielt frontal in die Kamera schaut, wie dies bei Grenzkontrollen der Fall ist.

invariant gegenüber Umgebungsänderungen wie Beleuchtungsveränderungen ist. Dies ist insbesondere in Kameranetzen ein Problem, da in den von unterschiedlichen Kameras beobachteten Szenen meist auch unterschiedliche Beleuchtungsbedingungen vorliegen. Zusätzlich weisen unterschiedliche Kameras auch immer unterschiedliche Farbprofile auf, d.h. die gleiche reale Farbe wird mit unterschiedlichen Farbwerten in den Videodaten dargestellt. Hier müssen die Farbprofile der Kameras also ineinander überführt werden [37, 139], was insbesondere bei großen Kameranetzen aufgrund der Anzahl der Kameras aber problematisch ist.

In Kameranetzen besteht eine weitere Möglichkeit darin, über die Erscheinungsinformation einer Person hinaus auch Information über die Anordnung der Kameras zu nutzen [83, 84, 127]. Mit dieser Information kann die Menge der möglichen Korrespondenzen zwischen Personen in Kameranetzen reduziert werden (z.B. durch das Wissen, dass eine Person frühestens 5 Minuten nach dem Auftreten im Sichtbereich der einen Kamera im Sichtbereich der anderen auftreten kann).

Beim Vorhandensein überschneidender Kameraansichten können auch diese dazu genutzt werden, die Wiedererkennung zu verbessern. In [68] wurde dazu eine „Panorama Appearance Map“ aufgebaut.

Ein weiterer Ansatz ist, den Kontext von Personen in die Wiedererkennung mit einzubeziehen [180]. Dieser Ansatz nutzt den Gruppenkontext um die Wiedererkennung einzelner Personen in der Gruppe durch die Kontextbeschreibung zu verbessern. Hier werden SIFT-Merkmale zusammen mit RGB-Farbmerkmalen zur Wiedererkennung genutzt. Gray et al. geben in [72] einen Überblick über die verschiedenen erscheinungsbasierten Methoden zur Objektwiedererkennung und validieren diese für den Fall der Personenwiedererkennung.

Die meisten aktuellen Ansätze zur Personenwiedererkennung bauen auf lokalen bzw. lokalisierten Merkmalen auf. Basis dieser Ansätze ist es, eine Person nicht in ihrer Gesamtheit durch ein Modell zu beschreiben, sondern als Zusammenschluss mehrerer lokalisierter Modelle. Hierbei werden Farbbeschreibungen oft mit anderen Merkmalen, wie Textur- oder Formmerkmalen zusammengebracht.

In [99] wird eine solche Kombination von Farb- und Texturmerkmalen vorgestellt. In [70] erfolgt eine Kombination von Farbmerkmalen mit sogenannten „salient edgels“. Hierdurch wird die Invarianz bzgl. Beleuchtung und Kleidungsdynamik erhöht. Zur Suche nach Personen mit bestimmten Eigenschaften wird in [159] ein System vorgestellt, welches körperpartiespezifische Farbattribute mit sogenannten *soft biometrics* wie Haarfarbe, -länge, Bart/kein Bart, Brille, etc. zusammenbringt. In [17] wird ein sogenannter „Histogram Plus Epitome[11] Descriptor (HPE)“ zur Personenwiedererkennung vorgestellt. Dieser verbindet die globale chromatische Information in Form von HSV (Hue Saturation Value) Histogrammen mit wiederkehrenden lokalen Mustern in einem gemeinsamen Deskriptor. Auf einem ähnlichen Prinzip baut die Arbeit von Farenzena et al. [55] auf. Hier wird die chromatische Information zusammen mit lokalen Mustern hoher Entropie zur Modellierung einer Person genutzt. Dabei werden die jeweiligen Modelle unabhängig für bestimmte Körperregionen bestimmt, wobei eine Gewichtung anhand von Symmetrieeigenschaften erfolgt.

In [163] wird ein Ansatz vorgestellt, der die Erscheinungsinformation ebenfalls in einer Ortsverteilung im räumlichen Umfeld von Objektkomponenten, d.h. von Körperregionen von Personen modelliert. In [73] wird eine Beschreibung durch ein „Ensemble of Localized Features (ELF)“ vorgestellt. Hierbei wird kein bestimmtes Merkmal zur Unterscheidung von Personen vorgestellt, sondern ein auf AdaBoost basierendes maschinelles Lernverfahren, welches auf Basis eines Trainingsdatensatzes ein maximal diskriminatives Modell auf Basis von einfachen Merkmalen bestimmt. In [74] werden lokale SURF-Merkmale zur Wiedererkennung von Personen genutzt. Im Gegensatz zu den vorherigen Verfahren wird hier nicht nur ein einzelner Zeitpunkt betrachtet, sondern es werden Merkmale über Bildsequenzen hinweg in ein Modell integriert. Dieses nutzt keine Ortsinformationen, ist also in der Distinktivität beschränkt. Zur Klassifikation wird ein KD-Tree genutzt, der einen schnellen Vergleich von Personenmodellen erlaubt. Wie auch schon die vorherigen Ansätze baut dieser Ansatz auf der

[11] „Ein Bildepitom ist das Ergebnis der Aufteilung eines Bildes in eine Menge überlappender Bereiche, die die Essenz der Textur-, Form- und Erscheinungseigenschaften des Bildes enthalten [17].

Annahme auf, dass für die wiederzuerkennenden Personen Annotierungen in Form einer bounding box vorliegen. Obwohl diese Annahme schlüssig erscheint, da es zahlreiche Verfahren zur Personendetektion und -verfolgung gibt, welche solche bounding boxes als Ergebnis liefern, vernachlässigt sie einen wesentlichen Aspekt, nämlich dass diese Verfahren nicht immer 100% Erkennungsrate erreichen und Personen auch nicht in jeder Situation exakt lokalisieren können. Gerade in Überwachungsszenarien, in denen Personen oft nur in geringer Auflösung und unter Teilverdeckungen sichtbar sind, ist es nicht immer möglich exakte bounding boxes oder sogar pixelgenaue Segmentierungen der Personen zu erhalten und somit die von der Person eingenommene Bildfläche als bekannt vorauszusetzen. Unter realen Bedingungen würden in diesem Beispiel Teile des Hintergrunds oder anderer Personen in der Pixelmaske enthalten sein. Diese würden die Wiedererkennung erschweren, bzw. abhängig von dem Verfahren unmöglich machen. Somit kann nicht davon ausgegangen werden, dass die Verfahren unter den in echten Anwendungen vorliegenden komplexeren Bedingungen einsetzbar sind.

Einige wenige Verfahren zur Wiedererkennung treffen diese Annahmen nicht und führen eine Wiedererkennung unter realen Bedingungen hinsichtlich der Ergebnisse der vorherigen Stufen durch.

Bak schlägt in [10] einen Ansatz vor, der räumliche Kovarianzregionen von Körperpartien zur Wiedererkennung nutzt. Die Körperpartien werden dabei von einem HOG-basierten Detektor automatisch detektiert. Tests finden allerdings hier auch nur auf bereits im vorhinein ausgeschnittenen Personenbeispielen statt.

In [9] wird eine Wiedererkennungsmethode vorgestellt, die auf „Dominant Color Descriptors“ bzw. auf „Haar-like“ Merkmalen [160] aufbaut. Die Wiedererkennung wird auf Bildausschnitten getestet, die durch einen HOG-basierten Personendetektor akquiriert wurden. Obwohl hier ein automatischer Objektdetektor genutzt wurde, und die Wiedererkennung hierdurch mit realen Problemen wie Ungenauigkeiten in der Detektion zurecht kommen muss, ist das Verfahren nicht in einem realen System einzusetzen, da es auf einem offline-Trainingsschritt aufbaut in dem ein Adaboost-Klassifikator [160] trainiert wird.

Ein solcher offline Trainingsschritt ist ein weiteres Manko aktueller Verfahren. Obwohl dieser grundsätzlich auch in realen Systemen offline für Modelle in der Datenbasis erfolgen kann, ist eine direkte online Klassifikation einer aktuell verfolgten Person nicht möglich, da die Zeit zwischen den Auftreten einer Person in unterschiedlichen Kameras nicht unbedingt für das Trainieren eines Klassifikators ausreichen muss. Dies beschränkt die Anwendbarkeit in Überwachungssystemen auf bestimmte Teilaufgaben, wie z.B. zur nachträglichen Unterstützung eines Operator bei einer Anfrage. Es ist z.B. nicht möglich, eine solche Wiedererkennung zum schritthaltenden Tracking in Kameranetzen zu nutzen.

Zudem sind viele der Ansätze durch Nutzung von Farbe zur Wiedererkennung auch hinsichtlich der Nutzung unterschiedlicher Sensoren eingeschränkt. So wären diese Ansätze nicht bei Sensoren nutzbar, die keine Farbinformationen liefern.

2.3 Einordnung der Arbeit in den Stand der Forschung

In dieser Arbeit wird ein System vorgestellt, welches die drei Stufen *Personendetektion*, *Personenverfolgung* und *Personenwiedererkennung* in einem integrierten Ansatz in ein Gesamtsystem einbettet.

Im Gegensatz zu vorhandenen Systemen für die Einzelstufen, welche die jeweilige Problemstellung typischerweise isoliert betrachten und spezialisierte Lösungsansätze verfolgen, werden die drei Aufgaben im hier vorgestellten System als *integrierte* Problemstellung betrachtet und angegangen.

So wird im Gegensatz zu Standard tracking-by-detection Ansätzen [167, 29, 31, 52, 109], welche die Objektdetektion unabhängig von der Objektverfolgung durchführen und bei denen der Informationsfluss nur in einer Richtung, von der Detektion zum Tracking, stattfindet, Detektion und Tracking hier als kombiniertes Problem betrachtet. Zur Lösung des Problems wird eine echte Kombination von Detektion und Tracking durch Zusammenführen von *bottom-up* und *top-down* Strategien in einer Erweiterung

des Implicit Shape Models zum Tracking-ISM durchgeführt. Im Gegensatz zu den Arbeiten von Leibe [110, 109] wird nicht nur eine Steuerung der Detektion durch das Tracking vorgeschlagen, sondern es findet eine wirkliche Kombination auf Basis des ISM statt. Dabei wird ein integrierter Systemansatz vorgestellt, der Heuristiken, wie sie in [167] zur Auswahl der jeweiligen Methodik (top-down oder bottom-up) genutzt werden, überflüssig macht.

Hierdurch entstehen Vorteile auf beiden Ebenen: so wird die Objektdetektion hinsichtlich der Stabilität verbessert und es wird ermöglicht, Personen über Verdeckungen hinweg stabil zu verfolgen. Der hier vorgestellte Ansatz ist im Gegensatz zu Methoden wie [5, 81, 110] zudem online-fähig, da keine globale Optimierung auf Basis der gesamten Bildsequenz stattfindet. Wie die Auswertung in Abschnitt 4.5 zeigt, ist die Performance des Trackings trotzdem mindestens vergleichbar mit diesen Ansätzen, übertrifft diese sogar in einigen Fällen.

Gleichfalls erfolgt eine Verbindung von Tracking und Wiedererkennung. Auch diese Aufgaben können nicht separat betrachtet werden, da die Personenwiedererkennung darauf angewiesen ist, dass während des Trackings Modelle der Personen aufgebaut werden. Die meisten bestehenden Ansätze [74, 180, 10, 70, 55, 17] beachten dies nicht und gehen davon aus, dass der vorherige Schritt optimale Ergebnisse liefert[12]. Dies ist in der Realität natürlich kaum der Fall, so dass diese Ansätze nicht in realen Systemen einsetzbar sind. Im hier vorgestellten System werden die während des Trackings aufgebauten Merkmalsmodelle zur Wiedererkennung verwendet. Der Ansatz ist also auch an dieser Stelle voll integriert. Insbesondere werden diese Identitätsmodelle zur Wiedererkennung online erstellt, ohne einen offline-Trainingsschritt zu nutzen, wie dies bei vielen Ansätzen [9, 73] der Fall ist. Wie in der Auswertung in Abschnitt 6.3 gezeigt wird, erreicht die hier vorgestellte, rein SIFT-basierte ISM-Wiedererkennung trotzdem ähnliche bzw. bessere Performance als diese Ansätze.

Zur Wiedererkennung selbst wird ein neuer mehrstufiger Ansatz vorgestellt, der Distinktivität mit Effizienz verbindet. Auf Stufe 1 werden ähnlich wie in [8, 113] für den Fall von Autos, Codebuchsignaturen zur Wiedererkennung genutzt, wobei diese hier während des Trackings aufgebaut werden. Auf Stufe 2 wird dieser sogenannte Bag-of-Features [150] Ansatz durch Kombination der ISM-Ortsverteilung mit der Codebuchsignatur erweitert. Hier wird also die ISM-Ausprägung zur Wiedererkennung genutzt. Abschließend werden auf Stufe 3 die SIFT-Deskriptoren verglichen, wobei auch hier das Codebuch zur Indizierung genutzt wird.

Erweiterungen bestehender Ansätzen finden auch auf Ebene der reinen Objektdetektion statt. So wird bei der Objektdetektion eine Methodik zur Klassifikation von Körperteilen vorgestellt, die im Gegensatz zu der in [105] ohne extensiven Lernschritt und Nutzung einer Segmentierung auskommt. Im weiteren wird ein Ansatz vorgestellt, der die Vorteile des ISM zur hierarchischen Objektdetektion ausnutzt. Darüberhinaus wird gegenüber dem Original-ISM [107] die Abhängigkeit von den Trainingsdaten reduziert und insbesondere ohne Zuhilfenahme einer Segmentierung eine kosistente Beschreibung, bei der die gleiche Bildregion nicht zwei unterschiedlichen Objekten zugeordnet werden kann, erstellt.

Durch die Tatsache, dass das Gesamtsystem keine lose Zusammenstellung unabhängiger Einzelmethoden ist, sondern auch die technische Umsetzung betreffend ein integriertes System darstellt, werden also Vorteile auf allen Einzelstufen erreicht. Ein weiterer wesentlicher Aspekt, der sich insbesondere durch die ausschließliche Nutzung des ISM zusammen mit SIFT-Merkmalen ergibt, ist, dass das Gesamtsystem in mehrfacher Hinsicht generisch ist. Dies ist von großer Bedeutung, da das System somit in unterschiedlichen Situationen und Anwendungen unabhängig nutzbar ist.

Die *Generizität* des Gesamtsystems bezieht sich dabei auf verschiedene Aspekte:

Objektgenerizität bedeutet, dass das System nicht auf eine bestimmte Objektklasse beschränkt ist. Dies wird durch Nutzung eines trainierbaren Objektdetektionsverfahrens erreicht. Hier wird, im Gegensatz zu Ansätzen wie [5, 67, 6], kein objektspezifisches Wissen, welches die Anwendbarkeit auf diese Objektklasse beschränkt, eingebracht. Das hier vorgestellte System wird anhand der

[12] Diese werden in Tests durch manuelle Annotationen ersetzt.

Objektklasse „Person“ vorgestellt, wobei an keiner Stelle im System Einschränkungen gemacht werden, welche die Anwendbarkeit auf diese Objektklasse beschränken würden. Die Objektklasse Person kann als Klasse des höchsten Schwierigkeitsgrads betrachtet werden, da es sich hierbei um ein stark artikuliertes Objekt handelt.

Szenariogenerizität heißt, dass das System keine Annahmen über das Anwendungsszenario macht. Es wird also kein Wissen über die Umgebung vorausgesetzt, wie dies in anderen Ansätzen der Fall ist. So wird bei vielen Ansätzen zur Objektdetektion und -verfolgung, aber auch bei Ansätzen zur Personenwiedererkennung vorausgesetzt, dass der Szenenhintergrund bekannt und/oder, bis auf Objekte der relevanten Klasse statisch ist. Dies impliziert, insbesondere für die Aufgabe der Objektverfolgung, dass der aufzeichnende Sensor stationär ist. Diese Annahme nutzen einige Ansätze [21, 94, 100, 63, 48, 39] zur Extraktion der Objekte, andere [110, 5, 81, 91] bauen zumindest bei der Modellierung der Objektdynamik (Bewegung) auf dieser Annahme auf. In beiden Fällen sind die Ansätze somit nicht bei bewegtem Sensor nutzbar. Allgemein schränken die Annahmen, welche für die Anwendungsumgebung durch Einbringen von Vorwissen getroffen werden, die Anwendbarkeit des Systems auf diese Umgebungen ein. Im hier vorgestellten System wird lediglich Vorwissen über die relevanten Objektklassen in Form von Beispielbildern dieser eingebracht. Ansonsten werden keinerlei Annahmen gemacht, was die Anwendbarkeit des Gesamtsystems in beliebigen Szenarien sicherstellt.

Sensorgenerizität bedeutet in diesem Kontext, dass das System weitestgehend unabhängig vom verwendeten Sensor ist. Dies wird durch die Nutzung von ausschließlich auf Intensitätsgradienten basierenden lokalen Bildmerkmalen erreicht. Die Sensorunabhängigkeit bezieht sich dabei auf den Typ der genutzten Kamera. In dieser Arbeit wird das Gesamtsystem zur Demonstration dieser Unabhängigkeit auf Daten aus dem sichtbaren sowie infraroten Spektralbereich ausgewertet. Es kann darüberhinaus angenommen werden, dass die Sensorunabhängigkeit nicht nur den Spektralbereich betrifft, sondern dass eine Verwendung des Verfahrens auch z.B. auf Sensoren wie Time-of-Flight Kameras, die Tiefeninformationen akquirieren, ohne größere Adaption möglich wäre.

Die meisten anderen Verfahren sind weit von dieser Sensorunabhängigkeit enfernt. Dies betrifft den Bereich der Objektverfolgung, bei dem in fast allen Verfahren [167, 29, 31, 52, 109] Farbe als Hauptmerkmal zur Datenassoziation im Tracking genutzt wird. Einige, auf Personenverfolgung bei mobilen Sensoren spezialisierte Ansätze [52, 69, 129, 53, 67], nutzen darüber hinaus spezielle Sensoren wie Stereokameras. Dieses bietet zweifelsohne Vorteile beim Personentracking. In bestimmten, insbesondere mobilen Szenarien ist die Nutzung zusätzlicher Sensoren oft auch die einzige Möglichkeit, eine stabile Personenverfolgung zu gewährleisten. Allerdings schränkt es die Anwendbarkeit des Systems auf Szenarien ein, in denen solche Sensoren zur Verfügung stehen.

Die Nutzung von Farbe verbietet nicht nur die Anwendbarkeit im infraroten Spektralbereich, sondern insbesondere auch bei panchromatischen Kameras, die gerade in Überwachungsszenarien aufgrund der höheren Lichtempfindlichkeit und aus der Entwicklungshistorie[13] noch in großem Ausmaß im Einsatz sind. Wie in Abschnitt 4.5 gezeigt wird, kann das hier vorgestellte System bei der Tracking-Performance trotz der Sensorunabhängigkeit mit spezialisierten Systemen, die Farbe nutzen, in solchen Szenarien, die Farbe bieten, konkurrieren.

Ebenso betrifft die nicht vorhandene Sensorgenerizität die meisten Verfahren im Bereich der Personenwiedererkennung. Auch hier verwenden die meisten Ansätze Farbmerkmale zur Wiedererkennung [180]. Neben der Tatsache, dass hierdurch die Anwendbarkeit eingeschränkt wird, ist dies, wie in Abschnitt 6.3 gezeigt wird, auch problematisch hinsichtlich der Wiedererkennung in Multi-Kamera-Netzen. Die Auswertung der Wiedererkennung in Abschnitt 6.3 zeigt, dass

[13] Panchromatische Kameras haben typischerweise eine höherere Lichtempfindlichkeit, d.h. sie liefern bei schlechteren Beleuchtungsverhältnissen als Farbkameras brauchbare Bilder. Durch den technischen Fortschritt sind heute auch Farbkameras in der Lage bei schlechteren Beleuchtungsbedingungen zu arbeiten. Allerdings sind viele der heute eingesetzten Überwachungskameras noch panchromatische Modelle.

das hier vorgestellte System, welches auch zur Wiedererkennung ausschließlich SIFT-Merkmale nutzt, auch hinsichtlich der Wiedererkennungsperformance mit solchen spezialisierten Systemen konkurrieren kann. Insbesondere wird in Abschnitt 5.4.1 gezeigt, dass mit dem hier vorgestellten Verfahren eine Wiedererkennung im infraroten Spektralbereich möglich ist.

Eine wichtige Eigenschaft des Gesamtsystems, die sich aus den genannten Generizitäten ergibt, ist die vollständig unabhängige Einsetzbarkeit des Systems. Dies ist in der Praxis sehr wichtig, da der Transfer eines solchen Systems in einen anderen Anwendungsbereich trotz der Generizitäten mit Aufwand, wie z.B. mit der Anpassung der Verfahren auf die Umgebungsbedingungen verbunden ist. Durch die Tatsache, dass das gesamte System auf dem ISM aufbaut, müssen bei Transfer in einen anderen Anwendungsbereich nur die anpassbaren Eigenschaften des ISM verändert werden. Es ist also nicht notwendig, locker zusammengefügte Teilverfahren in einem komplexen Prozess aufeinander abzustimmen. Insbesondere ist das hier vorgestellte System das einzige, welches Funktionalität in dieser Form mit Generizität vereinigt und somit als eines der wenigen bis hin zur Wiedererkennung in der Praxis einsetzbar ist[14].

[14]Im jetzigen Entwicklungsstadium ist für das Gesamtsystem noch keine Echtzeitumsetzung vorhanden. Eine solche Umsetzung sollte aber bei Betrachtung der Komplexität der Teilverfahren bei Nutzung aktueller Hardware und evtl. Parallelisierung möglich sein.

Kapitel 3

Personendetektion

Die Extraktion relevanter Objekte stellt die erste Stufe eines typischen Bildanalysesystems dar. Die vorliegende Arbeit konzentriert sich auf Personen als relevante Objektklasse, was bedeutet, dass im ersten Schritt eine Erkennung von Personen in den Bilddaten erfolgt.

Die Grundlage des in dieser Arbeit zur Personendetektion genutzten Verfahrens bildet das *Implicit Shape Model (ISM)*, welches von Leibe [107] (vgl. Jüngling und Arens [87]) als Verfahren zur Objektdetektion in Einzelbildern eingeführt wurde. Dieses Verfahren baut ausschließlich auf lokalen Bildmerkmalen auf und ist als trainierbares Verfahren für beliebige Objektklassen einsetzbar. Da die Objektdetektion direkt im Einzelbild erfolgt, ist es unabhängig von der Anwendungsumgebung, kann also z.B. auch bei bewegter Kamera und unbekanntem Szenenhintergrund eingesetzt werden. Durch die Nutzung von ausschließlich lokalen, auf Intensitätsgradienten basierenden Merkmalen ist es auch für unterschiedliche Sensoren, wie z.B. Kameras für den infraroten und sichtbaren Spektralbereich einsetzbar. Es erfüllt somit alle in Kapitel 1 definierten Anforderungen und ist zur Nutzung im Kontext dieser Arbeit geeignet.

In diesem Kapitel wird das unabhängig arbeitende Einzelbilddetektionsverfahren, welches eine Erweiterung und Adaptierung des in [107] beschriebenen Verfahren darstellt, eingeführt. Erweiterungen und Adaptierungen finden insbesondere in Hinblick auf den angestrebten Verwendungszweck über die Personendetektion hinaus statt. Wie bereits in Abschnitt 2.3 erwähnt, werden im Rahmen dieser Arbeit Detektion und Tracking kombiniert. Dazu wird der ISM-Objektdetektor in Kapitel 4 zum Tracking-ISM erweitert. Der Einzelbildobjektdetektor kommt im Gesamtsystem also nicht dediziert, sondern nur in der durch das Tracking erweiterten Form zum Einsatz. Da das Einzelbildverfahren aber die Grundlage für die in Kapitel 4 vorgestellten Erweiterungen zum Tracking darstellt, soll hier zunächst eine detaillierte Einführung in das prinzipielle Verfahren erfolgen.

Dazu wird in Abschnitt 3.1 zunächst eine Einführung in lokale Bildmerkmale gegeben, die eine wichtige Grundlage dieser Arbeit darstellen. Hierbei wird auf die Anforderungen eingegangen, die in dieser Arbeit an lokale Merkmale gestellt werden. Detailliert werden die in dieser Arbeit verwendeten SIFT-Merkmale in Abschnitt 3.1.1 vorgestellt.

Zur Objektdetektion wird im ersten Schritt, dem *Trainingsschritt* ein Codebuch zur Beschreibung der Objektklasse erstellt. Dieser Trainingsschritt wird in Abschnitt 3.2 beschrieben. Die eigentliche *Objektdetektion* wird in Abschnitt 3.3 eingeführt.

In den Abschnitten 3.4 und 3.5 wird gezeigt, wie das ISM durch einfache Erweiterungen über die reine Objektdetektion hinaus zur *Körperteilklassifikation*, bzw. zur *hierarchischen Objektdetektion* genutzt werden kann.

Eine experimentelle Validierung des vorgestellten Verfahrens erfolgt in Abschnitt 3.6. Da die Performance des ISM bei der Personendetektion bereits in anderen Arbeiten [111] umfassend validiert wurde,

und nicht beansprucht wird, dass die hier vorgestellten Änderungen des ISM eine wesentliche Performanceverbesserung zur Folge haben, wird hier nur eine grundlegende Performance-Charakterisierung vorgenommen, die im wesentlichen dazu dienen soll, die vorhandenen Defizite der reinen Objektdetektion aufzuzeigen.

In [102, 101] wurde gezeigt, dass die in [107] verwendete probabilistische Formulierung für das Hough-Voting Framework in dieser Form nicht gerechtfertigt ist[1]. In dieser Arbeit wird somit von einer solchen probabilistischen Formulierung abgesehen und eine allgemeinere Beschreibung verwendet.

3.1 Grundlagen: Lokale Bildmerkmale

Lokale Bildmerkmale bilden die Grundlage für den immensen Fortschritt, der in vielen Bereichen des maschinellen Sehens in den letzten Jahren erzielt wurde. Gerade beim Fortschritt im Bereich der automatischen Objektdetektion und hierbei insbesondere im Bereich der Personendetektion spielen sie eine tragende Rolle [120, 178, 146].

Lokale Bildmerkmale bilden ebenfalls die Basis des hier vorgestellten Ansatzes zur Personendetektion. Aus diesem Grund wird an dieser Stelle eine Einführung in diese Thematik gegeben.

Lokale Bildmerkmale sind Merkmale, die einen lokalen Bildbereich entweder um einen vorher determinierten *Schlüsselpunkt* oder in einem bestimmten Raster im ganzen Bild in Form einer *dichten Abtastung* beschreiben.

Im hier relevanten Kontext der Objektdetektion ist ein Beispiel für die Verwendung einer dichten Abtastung der HOG (Histogram of Oriented Gradients) Objektdetektor [43]. Das im Mittelpunkt dieser Arbeit stehende Implicit Shape Model verwendet im Gegensatz dazu eine auf Schlüsselpunkten basierende Beschreibung durch lokale Merkmale.

Der allgemeine Vorteil von lokalen Bildmerkmalen ist zum Einen, dass sie die Bildbeschreibung einer örtlich beschränkten Bildregion unabhängig vom globalen Kontext erlauben, insbesondere aber auch, dass die Beschreibung dieser Bildregion viele Invarianzen, wie typischerweise Translations-, Skalierungs- und Rotationsinvarianz beinhaltet. Dies bedeutet, dass die lokalen Merkmale unter diesen Bildtransformationen wiedergefunden und anhand der Deskriptoren zugeordnet werden können. Die Invarianz, z.B. im Fall der Skalierungsinvarianz, wird dabei teilweise durch die Art der Schlüsselpunktdetektion und/oder durch die Art der Merkmalsbeschreibung ermöglicht. Zur Beschreibung des Bildbereichs wird oft eine Intensitätsgradientenverteilung genutzt, wodurch neben den bereits genannten häufig noch weitere Invarianzen, wie z.B. Beleuchtungsinvarianz integriert werden.

In dieser Arbeit werden die von David Lowe in [117] eingeführten *SIFT (Scale Invariant Feature Transform)* Merkmale verwendet. SIFT-Merkmale sind eine Kombination von Schlüsselpunktdetektion auf Basis des *Difference of Gaussian (DoG)*, der eine Annäherung des *Laplacian of Gaussian (LoG)* darstellt und einer Merkmalsbeschreibung durch ein Gradientenorientierungshistogramm. Eine detaillierte Beschreibung der technischen Grundlagen dieser Merkmale erfolgt in Abschnitt 3.1.1.

Im Grundsatz sind in dieser Arbeit SIFT durch andere lokale Bildmerkmale wie z.B. SURF [15][2], oder eine Kombination von Harris-Laplace [123] oder Hessian-Laplace [126] als Schlüsselpunktdetektor mit z.B. Shape Context [20] als Deskriptor ersetzbar.

Als notwendig wird lediglich vorausgesetzt, dass diese Merkmale den folgenden Eigenschaften genügen: (i) Verschiebungsinvarianz, (ii) Skalierungsinvarianz, (iii) Beleuchtungsinvarianz.

[1] Die Marginalisierung (Summierung über alle Merkmale) zur Berechnung der Hypothesenstärke impliziert dass alle Merkmale mögliche Ausprägungen einer einzelnen Zufallsvariablen sind und damit nur ein Merkmal gleichzeitig beobachtet werden kann. Da in einem Bild mehrere Merkmale gleichzeitig beobachtet werden, sollte jedes Merkmal allerdings durch eine separate Zufallsvariable modelliert werden.

[2] Siehe hierzu die Arbeiten zur Personendetektion und -verfolgung von Jüngling und Arens [87, 86, 90, 88], in denen SURF verwendet wurde.

Da die lokalen Merkmale in dieser Arbeit für drei unterschiedliche Zwecke, nämlich die Detektion, Verfolgung und Wiedererkennung von Personen genutzt werden, ist es notwendig, dass der verwendete Merkmalstyp für alle drei Aufgaben gleichermaßen geeignet ist. Es reicht also nicht aus, wenn gute Generalisierungseigenschaften zur Objektdetektion vorhanden sind. Ebenso muss für die Verfolgung und insbesondere Wiedererkennung ein hohes Maß an Diskriminativität vorhanden sein.

Da die Anzahl der unterschiedlichen lokalen Merkmalstypen mit dem initialen Erfolg in den letzten Jahren immens zugenommen hat, gibt es in der heutigen Literatur eine Vielzahl von Merkmalstypen, die diese Eigenschaften erfüllen. Eine umfassende Übersicht über den gesamten Bereich soll hier nicht gegeben werden. Hierzu sei auf die Arbeiten [144, 126, 124, 158] verwiesen.

Diese Arbeiten bieten auch einen Vergleich der Merkmalseigenschaften. So wird in [124] ein umfangreicher Vergleich der Merkmalstypen bzgl. Eigenschaften wie Wiederholbarkeit und Distinktivität vorgestellt. Hierbei wird auch der Umfang der Invarianz z.B. bzgl. Veränderungen der Ansicht, Skalierung, und Beleuchtung betrachtet. Die Nutzbarkeit eines Merkmalstyps hängt immer vom Anwendungskontext, also von der Aufgabe für die ein Merkmal eingesetzt wird, ab. In [120] wird die Eignung verschiedener lokaler Merkmale für die Aufgabe der Objekterkennung betrachtet. Grundsätzlich zeigen die Vergleiche, dass es keinen einzelnen Merkmalstyp gibt, der den Anderen in allen Belangen überlegen ist. Allerdings liegt SIFT in allen Bereichen in der Spitzengruppe.

3.1.1 SIFT

Der SIFT-Algorithmus besteht aus vier aufeinander aufbauenden Stufen:

1. Detektion von scale-space Extrema.
2. Schlüsselpunktlokalisation.
3. Orientierungszuweisung.
4. Beschreibung durch den Deskriptor.

Diese werden in den folgenden Abschnitten beschrieben.

3.1.1.1 Detektion von scale-space Extrema

Zur Berechnung skalierungsinvarianter Merkmale wird der *scale-space (Skalenraum)* genutzt. Dieser dient der Darstellung eines Bildes auf Stufen unterschiedlichen Detaillierungsgrads. Hier geht die Darstellung auf höheren Skalierungsstufen mit der Verringerung der Auflösung bzw. äquivalent mit der Verringerung der Bilddetails, d.h. mit einer Informationsreduktion einher. Diese Informationsreduktion kann auf einem Eingabebild durch eine Tiefpassfilterung erzielt werden [49].

Zum Aufbau des scale-space wird das Eingabebild $I(x, y)$ mit einem Gaußfilter variabler Skalierung $G(x, y, \sigma)$ gefaltet:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y). \tag{3.1}$$

σ gibt dabei die Varianz des Gaußfilters an. $L(x, y, \sigma)$ ergibt somit eine Stufe, also ein Bild, im scale-space.

Zur Gewinnung skalierungsinvarianter Schlüsselpunkte in diesem scale-space, werden die Extrema der Faltung der *Difference of Gaussian (DoG)* Funktion mit dem Eingabebild verwendet [116].

Die DoG-Funktion $D(x, y, \sigma)$ ist eine Annäherung [114] des skalierungsnormalisierten *Laplacian of Gaussian* (LoG) $\sigma^2\nabla^2 G$ und kann durch Differenzbildung zweier benachbarter Stufen (die sich um einen Faktor k in der Skalierung unterscheiden) im scale-space effizient berechnet werden:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma). \tag{3.2}$$

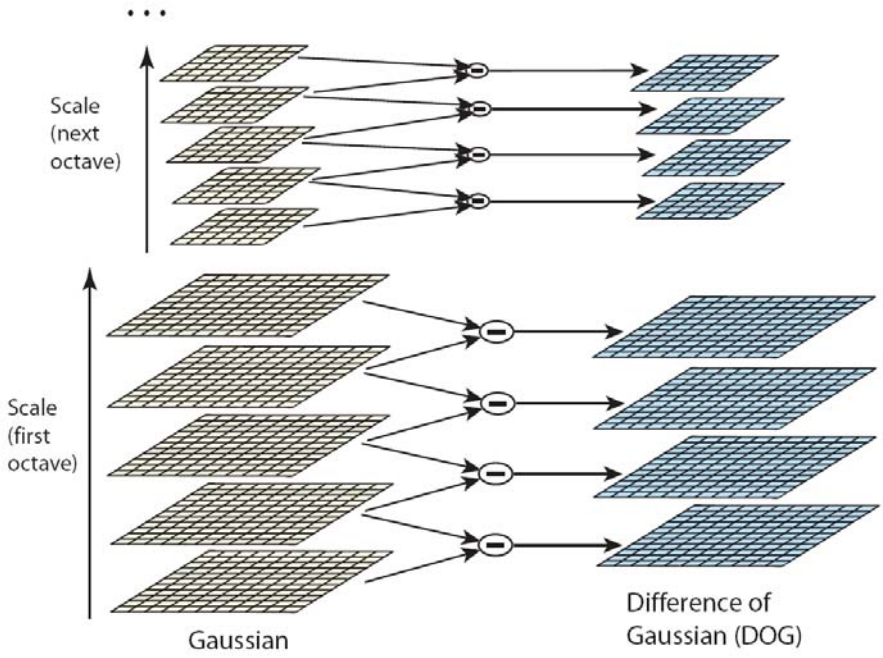


Abbildung 3.1: Aufbau des scale-space und Bilden der Difference-of-Gaussian zur Schlüsselpunktdetektion. (Abbildung aus [117].)

Diese Bilder müssen dabei nicht extra zur Berechnung der DoG-Funktion berechnet werden, sondern werden ohnehin zur Deskriptorberechnung benötigt. In [114] wurde von Lindeberg gezeigt, dass die Normalisierung mit σ^2 notwendig ist, um tatsächlich Skalierungsinvarianz herzustellen. Mikolajczyk et al. [122] zeigten, dass die Maxima und Minima des LoG im Vergleich mit anderen Schlüsselpunktfunktionen (Hessian, Harris) die stabilsten Bildmerkmale unter verschiedenen Transformationen generieren.

Der Zusammenhang zwischen D und $\sigma^2\nabla^2 G$ wird dabei aus der Wärmediffusionsgleichung abgeleitet:

$$\frac{\partial G}{\partial \sigma} = \sigma \nabla^2 G. \tag{3.3}$$

Hierbei kann $\frac{\partial G}{\partial \sigma}$ durch den Differenzenquotienten von zwei um den Faktor k getrennten Stufen des scale-space approximiert werden:

$$\frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, \sigma)}{k\sigma - \sigma}. \tag{3.4}$$

Womit sich der Zusammenhang

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G \tag{3.5}$$

ergibt.

In Abbildung 3.1 ist der Aufbau des scale-space visualisiert. Da beim Bilden des scale-space ein Ansteigen von σ mit einem Informationsverlust im Bild verbunden ist, kann der scale-space als Pyramide dargestellt werden. Das Bild wird also bei einer Verdopplung von σ in der Größe halbiert[3]. Der Vorteil

[3]Die Verdopplung des σ des Tiefpassfilters halbiert den Wert der höchsten Änderungsfrequenz des Bildes. D.h. es wird

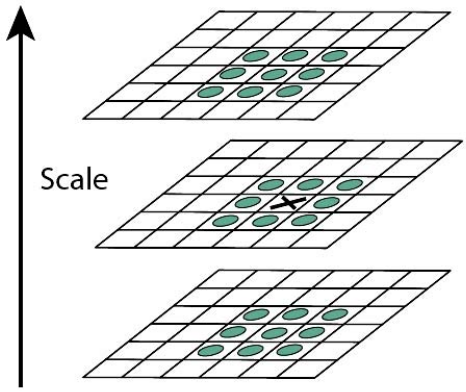


Abbildung 3.2: Maxima und Minima in der DoG werden durch Vergleich mit den 26 Nachbarn detektiert. (Abbildung aus [117].)

hierbei liegt in der Verringerung des Rechenaufwands durch Reduktion der Bildgröße bei gleichbleibendem Informationsgehalt. Für jede Pyramidenstufe (Octave, Verdopplung von σ) müssen dabei bei einem konstanten Abstand $k = 2^{\frac{1}{s}}$ des Glättungsgrads der Bilder, bei s Bildern innerhalb einer Octave, $s + 3$ Bilder generiert werden, um die jeweiligen DoG-Bilder zur Schlüsselpunktdetektion zu generieren.

3.1.1.2 Lokalisation von Schlüsselpunkten

Die Maxima und Minima der DoG-Funktion dienen als Ausgangspunkte der Schlüsselpunktdetektion. Zur Ermittlung der Maxima und Minima wird, wie in Abbildung 3.2 dargestellt, jeder Punkt in der DoG mit seinen 26 Nachbarn verglichen. Jeder Punkt, der bei diesem Vergleich als Maximum oder Minimum in der Nachbarschaft hervorgeht, wird als Ausgangspunkt zur genaueren Lokalisation der Bildposition und Skalierungsstufe des Schlüsselpunktes weiterverwendet. Dazu wird eine Interpolation durch die Taylorentwicklung (bis zu quadratischen Termen) von $D(x, y, \sigma)$ in diesem Punkt $\mathbf{x} = (x, y, \sigma)^T$ vorgenommen:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}. \tag{3.6}$$

Die Position des Extremums $\hat{\mathbf{x}}$ wird dabei durch Ableitung der Funktion nach $\mathbf{x}$ und Nullsetzen bestimmt:

$$\hat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2}\frac{\partial D}{\partial \mathbf{x}}. \tag{3.7}$$

Die Ableitungen von D werden dabei durch die Differenzen benachbarter Pixel approximiert. Man erhält somit ein lineares 3x3 Gleichungssystem das mit geringem Aufwand gelöst werden kann. Falls der Versatz $\hat{\mathbf{x}}$ größer als 0.5 in einer Dimension ist, so liegt das Extremum näher an einem Nachbarpunkt. In diesem Fall wird die Interpolation erneut mit dem Nachbarpunkt als Startpunkt durchgeführt. Anhand des Funktionswerts $D(\hat{\mathbf{x}})$ an der Position des Extremums können Extrema mit niedrigem

eine Neuabtastung mit halber Abtastrate ohne Informationsverlust möglich [33]. Bei SIFT wird für das Initialbild eine Glättung mit $\sigma = 1.6$ vorausgesetzt. Der Wert 1.6 wurde experimentell bestimmt. Dieser stellt einen Kompromiss zwischen Schlüsselpunktstabilität und Rechenaufwand (je größer σ desto höher ist der Rechenaufwand der Faltung) dar. Für Details und die genaue Umsetzung sei auf [117] verwiesen.

Kontrast ausgefiltert werden. Dazu wird Gleichung 3.7 in Gleichung 3.6 substituiert:

$$D(\hat{\mathbf{x}}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\hat{\mathbf{x}}. \tag{3.8}$$

Die Wahl des Schwellwertes für $|D(\hat{\mathbf{x}})|$ bestimmt, bis zu welchem Kontrast Merkmale für Schlüsselpunkte berechnet werden. Je niedriger der Schwellwert, desto mehr Merkmale werden also aus dem Eingabebild extrahiert. Dies ist in dieser Arbeit insbesondere bei der Anwendung der Algorithmen auf Daten aus den infraroten Spektralbereich relevant. Hier kommt es vor (siehe z.B. CASIA C Datensatz [1]), dass Personen nur in sehr geringem Kontrast zum Hintergrund auftreten. In diesem Fall muss der Schwellwert für $|D(\hat{\mathbf{x}})|$ also im Vergleich zu Szenarien, in welchen ein guter Kontrast von Personen zum Hintergrund zu erwarten ist, sehr niedrig gewählt werden.

Ein weiteres Kriterium, das für die Auswahl der Schlüsselpunkte herangezogen werden kann ist die Krümmung (Edgeness). Die DoG-Funktion erzeugt viele Extrema an Kanten. Da ein Extremum entlang einer Kante aber nicht gut zu lokalisieren ist, ist es wünschenswert, Extrema, die auf Kanten, aber nicht auf Ecken liegen, auszufiltern. Da ein solches Extremum eine starke Krümmung in Richtung der Kante, aber eine schwache Krümmung senkrecht dazu, entlang der Kante, aufweist, kann die Krümmung dazu genutzt werden, solche Extrema auszufiltern. Dazu wird die 2x2-Hesse-Matrix $\mathbf{H}$, welche die zweiten Richtungsableitungen und damit die Krümmungen enthält, berechnet:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \tag{3.9}$$

Die Eigenwerte λ_1 und λ_2 von $\mathbf{H}$ sind proportional zu den Hauptkrümmungen, wobei die Eigenwerte nicht explizit berechnet werden müssen, da lediglich das Verhältnis der Werte relevant ist. Entsprechend [76] wird zur Berechnung des Verhältnis die Spur und die Determinante von $\mathbf{H}$ genutzt:

$$trace(\mathbf{H}) = D_{xx} + D_{yy} = \lambda_1 + \lambda_2, \tag{3.10}$$

$$det(\mathbf{H}) = D_{xx}D_{yy} - D_{xy}^2 = \lambda_1\lambda_2. \tag{3.11}$$

Im Fall einer negativen Determinante haben die Krümmungen unterschiedliche Vorzeichen und das Extremum wird zurückgewiesen. Sei r das Verhältnis von größerem Eigenwert λ_1 und λ_2, mit $\lambda_1 = r\lambda_2$. Dann ist:

$$\frac{trace(\mathbf{H})^2}{det(\mathbf{H})} = \frac{(\lambda_1 + \lambda_2)^2}{\lambda_1\lambda_2} = \frac{(r+1)^2}{r}. \tag{3.12}$$

Der Wert von $\frac{(r+1)^2}{r}$ ist minimal wenn beide Eigenwerte gleich sind und wächst mit r. Daher reicht zur Überprüfung, dass die Hauptkrümmungsrate unter einem Schwellwert r liegt, die folgenden Prüfung:

$$\frac{trace(\mathbf{H})^2}{det(\mathbf{H})} < \frac{(r+1)^2}{r}. \tag{3.13}$$

3.1.1.3 Orientierungszuweisung

Um Rotationsinvarianz des Merkmalsdeskriptors zu erreichen, wird diesem eine Hauptorientierung zugewiesen. Diese wird anhand der Gradienten in der Region um den Schlüsselpunkt bestimmt. Der Deskriptor, der ebenfalls ausschließlich auf den Gradienten im Bereich um den Schlüsselpunkt aufbaut, kann dann in Abhängigkeit (also relativ zu) dieser Orientierung berechnet werden.

Zur Berechnung der Orientierung wird das Bild L gewählt, dessen (durch σ gegebene) Skalierung am nächsten an der Skalierung des Schlüsselpunktes liegt. Für jedes Pixel mit Position (x, y) dieses Bildes werden Gradientenstärke $m(x, y)$ und Gradientenorientierung $\theta(x, y)$ bestimmt:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}, \tag{3.14}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right). \tag{3.15}$$

Auf Basis der Gradienten in einer Region um den Schlüsselpunkt wird ein Gradientenorientierungshistogramm mit 36 Fächern (bins) berechnet. Bei der Berechnung werden die Gradienten mit der Gradientenstärke sowie zusätzlich mit einem im Schlüsselpunkt zentrierten Gauß-Fenster mit σ von 1.5 mal der Skalierungsstufe des Schlüsselpunktes gewichtet. Die zusätzliche Gewichtung mit einem Gaußkernel wird durchgeführt, um Gradienten, die weiter vom Schlüsselpunkt entfernt sind, also in den Außenbereichen der Region liegen, im Einfluss auf die Orientierungsberechnung abzuschwächen. Hierdurch soll eine starke Veränderung der Hauptorientierung bei kleinen Veränderung der Schlüsselpunktposition vermieden werden. Maxima im Gradientenorientierungshistogramm geben die dominanten Orientierungen der Deskriptorregion an. Für das stärkste Maximum sowie alle anderen Maxima, die in einem Bereich von 80% der Stärke des stärksten Maximums liegen, werden Deskriptoren mit der jeweiligen Orientierung berechnet. Für einen Schlüsselpunkt werden also potentiell mehrere SIFT-Merkmale berechnet.

Im Kontext dieser Arbeit wird die Rotationsinvarianz des Deskriptors nicht ausgenutzt, da sich in der Praxis gezeigt hat, dass sich diese negativ auf die Qualität des Detektionsverfahrens auswirkt. Dies ist der Fall, da die Orientierung des SIFT-Deskriptors relativ instabil ist und sich kleine Änderungen in der der Orientierungsbestimmung zugrunde liegenden Region in einer Änderung der Hauptorientierung auswirken können. Da eine Übereinstimmung zweier Deskriptoren, welche die gleiche Bildregion unter geringen Änderungen beschreiben, voraussetzt, dass die Orientierung bei beiden Merkmalen korrekt (also gleich in Bezug auf die zu beschreibende Bildregion) bestimmt wird, sind Instabilitäten bei der Orientierungsbestimmung in der hier angestrebten Anwendung nicht tolerierbar. Eine Möglichkeit die Instabilitäten in der Praxis zu beheben besteht darin, über Orientierungsmaxima, deren Stärke innerhalb von 80% der Stärke des stärksten Maximums liegen hinaus, für jeden gefundenen Schlüsselpunkt mehrere Merkmale mit jeweils unterschiedlichen Orientierungen zu berechnen. Ein Nachteil hierbei ist, dass die Menge der Merkmale, die aus einem Bild extrahiert wird, künstlich vergrößert, und insbesondere auch verwässert wird. Dies hat in den Anwendungen, die auf der Merkmalsmenge aufbauen, natürlich eine erhöhte Laufzeit so wie evtl. Ungenauigkeiten, die sich aus der verwässerten Merkmalsmenge ergeben, zur Folge.

Insbesondere verringert sich durch die Rotationsinvarianz auch die Distinktivität eines Deskriptors. So ist bei der Auswahl des Merkmalsdeskriptors immer ein Kompromiss zwischen Distinktivität und Invarianz zu schließen. Da Distinktivität im Kontext dieser Arbeit eine übergeordnete Rolle spielt und durch Rotationsinvarianz auch Unsicherheiten in das den zentralen Punkt dieser Arbeit bildende Implicit Shape Model eingefügt werden, wird auf die Nutzung der Rotationsinvarianz in diesem Kontext verzichtet. Dazu wird die Orientierung des Schlüsselpunkt auf 0 gesetzt.

3.1.1.4 Deskriptor

Die Deskriptorberechnung ist in Abbildung 3.3 schematisch dargestellt. Zunächst werden die Bildgradienten in einer Region um den Schlüsselpunkt berechnet. Hierbei wird der Bildbereich vor Berechnung der Gradienten, mit einem von der Skalierungsstufe des Schlüsselpunkts abhängigen Gaußfilter geglättet. Um Rotationsinvarianz des Deskriptors zu gewährleisten, werden die Gradienten um die Orientierung des Schlüsselpunkts rotiert (hier mit Orientierung = 0 ist dieser Schritt nicht notwendig). Um zu vermeiden, dass schon kleine Positionsänderungen des Deskriptors bei einer in der Praxis selten optimalen Schlüsselpunktlokalisation, zu starken Veränderungen des Deskriptors selbst führen, erfolgt zur Deskriptorberechnung eine Gewichtung (siehe Abbildung 3.3 links) der Gradienten mit einem Gaußfilter. Da hierdurch die Gradienten an den Rändern des Deskriptors geringeren Einfluss

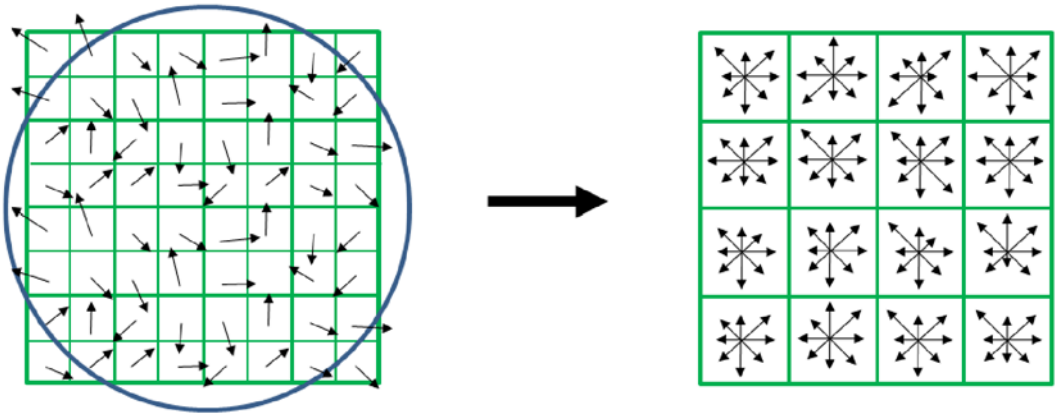


Abbildung 3.3: Zur Berechnung des SIFT-Deskriptors werden die Bildgradienten an jedem Pixel in einer Region um den Schlüsselpunkt berechnet.

im Deskriptor haben, schlagen sich kleine Positionsänderungen nicht mehr in so starken Deskriptoränderungen nieder.

Zur Berechnung des Deskriptors erfolgt eine Einteilung in ein Ortshistogramm mit 4x4 Subregionen. In jeder dieser 16 Regionen wird unabhängig von den anderen Regionen ein Gradientenorientierungshistogramm berechnet. Hierbei erfolgt eine Gewichtung der Gradientenorientierung mit der Gradientenstärke. Das Orientierungshistogramm hat dabei eine Fachbreite von 45°, wodurch sich 8 Fächer pro Histogramm ergeben. Der resultierende SIFT-Deskriptor hat also eine Größe von 4x4x8=128.

Um zu vermeiden, dass sich leichte Veränderungen der Gradienten zu stark auf den Deskriptor auswirken, werden die Histogrammfächer trilinear interpoliert. D.h. ein Gradient wird nicht fest einem einzelnen Fach zugeordnet, sondern auch gewichtet in die angrenzenden Fächer verteilt. Die Gewichtung erfolgt dabei pro Dimension mit $1 - d$, wobei d die Distanz des Gradienten vom zentralen Wert des jeweiligen Fachs ist. Bei einem Orientierungshistogramm mit 10° Fächern und einem Gradienten mit einer Orientierung von 28° würde dies bedeuten, dass dieser mit einer Gewichtung von $1-(\frac{|28-25|}{10})$ dem $20 - 30°$ Fach und mit $1 - (\frac{|28-35|}{10})$ dem $30 - 40°$ Fach zugeordnet würde. Um nicht nur Veränderungen der Gradientenorientierung, sondern auch der Position zu behandeln, wird die Interpolation auch für das Positionshistogramm durchgeführt. Durch Normierung des Deskriptorvektors auf Einheitslänge wird der Deskriptor invariant gegenüber gleichmäßigen Beleuchtungsveränderungen in der Deskriptorregion. Um auch eine weitgehende Invarianz gegen nicht gleichmäßige, also nichtlineare Beleuchtungsveränderungen zu integrieren, wird der Einfluss von starken Gradienten im Deskriptor beschränkt. Dies geschieht unter der Annahme, dass nicht gleichmäßige Beleuchtungsveränderungen den Betrag einiger Gradienten im Deskriptorbereich stark ansteigen lassen, wobei die Gradientenorientierung unbeeinflusst bleibt. Um den Deskriptor so weit wie möglich invariant gegenüber diesen Veränderungen zu halten, wird der Wert eines einzelnen Eintrags im normalisierten Deskriptor auf maximal 0.2 beschränkt. Für den Vergleich von Deskriptoren bedeutet dies, dass hier der absolute Wert von großen Gradienten an Bedeutung verliert.

Wenn sich in der folgenden Arbeit auf *Merkmal* bzw. *Bildmerkmal* bezogen wird, ist damit immer das gesamte SIFT-Merkmal, bestehend aus dem Schlüsselpunkt mit Bildposition und Skalierung gemeint. Unter *Deskriptor* wird, falls nicht spezieller Bezug genommen wird, das in diesem Fall 128-dimensionale Gradientenorientierungshistogramm, welches die Bildregion um den Schlüsselpunkt beschreibt, verstanden.

3.2 Training

In der vorliegenden Arbeit werden zur Objektdetektion, -verfolgung und -wiedererkennung erlernte Objektmodelle genutzt. Diese werden in einem Trainingsschritt generiert. Genauer werden in diesem Trainingsschritt Codebücher für die relevanten Objektklassen erstellt. Als Eingabe dieses Trainingsschritts ist Trainingsmaterial für die jeweilige Objektklasse notwendig. Dies sind zum einen Bilder, welche Objekte der relevanten Klasse enthalten, sowie eine Annotation der Trainingsdaten, die angibt, an welcher Bildposition sich die relevanten Objekte befinden. Die Annotation wird zur Selektion der für die Objektklasse relevanten Merkmale genutzt. An Stelle einer Annotation in Form eines das Objekt umschließenden Rechtecks (bounding box), kann auch eine Referenzsegmentierung (Figure/Ground Segmentierung) dieser Trainingsbilder zur genaueren Merkmalsauswahl genutzt werden.

Der *erste Schritt* im Training ist die *Extraktion von lokalen Bildmerkmalen* aus den Trainingsbildern. In dieser Arbeit werden die in Abschnitt 3.1.1 beschriebenen SIFT-Merkmale genutzt. Die Selektion der relevanten Merkmale wird anhand der Annotation durchgeführt. Es ergibt sich somit eine Menge Π_{img} von Merkmalen. Ein Merkmal $\pi_{img} \in \Pi_{img}$ ist dabei durch die Bildposition $(\pi_{img,x}, \pi_{img,y})$, die Skalierungsstufe $\pi_{img,s}$ und den Merkmalsdeskriptor $\pi_{img,\zeta}$ bestimmt.

Diese Merkmale werden im *zweiten Schritt* zur *Erstellung des Codebuchs* für die zu trainierende Objektklasse genutzt. Dazu werden die Merkmale anhand der Deskriptorähnlichkeiten gruppiert, um Deskriptorprototypen der Codebuchklassen zu erhalten. Für das Beispiel der Klasse Person würde dies im Optimalfall z.B. bedeuten, dass alle Deskriptoren, die den linken Fuß einer Person darstellen in einer Gruppe zusammengefasst werden, alle Merkmale die den rechten Fuß darstellen, in einer weiteren, usw..

Die Gruppierung der Merkmale zum Erstellen des Codebuchs wird durch Anwendung eines *Ballungsverfahrens (Clustering)* [108] durchgeführt, wobei das Ziel darin besteht, möglichst kompakte Ballungen (Cluster) zu erhalten, die in sich konsistent bzgl. der repräsentierten Struktur sind. Dies ist von großer Bedeutung, da als Repräsentanten im Codebuch die Clustermittelpunkte genutzt werden und zu ausgedehnte Cluster somit in großen Ungenauigkeiten innerhalb des Codebuchs resultieren würden. Dieses Ziel muss mit der Motivation des Clusterings, nämlich eine Verringerung der Rechenzeit bei der Objektdetektion und eine Generalisierung der repräsentierten Struktur durch Zusammenfassung von Merkmalen zu erhalten, in Einklang gebracht werden. Es ist also nicht ausreichend, kompakte Cluster zu generieren, sondern diese Cluster sollen auch eine möglichst große Menge an Deskriptoren, und somit eine für die Objektklasse signifikante Struktur beinhalten. Insgesamt ist also ein Kompromiss zwischen Genauigkeit und Allgemeingültigkeit der Repräsentation zu finden.

Zum Clustering wird der RNN-Algorithmus[4] für agglomeratives Clustering gewählt. Dieser bietet im Gegensatz zu partitionierenden Verfahren wie dem K-Means-Algorithmus [77] den Vorteil, dass kein a-priori Wissen über die Anzahl der Cluster notwendig ist und die gewünschte Kompaktheit der Cluster direkt über einen Schwellwert einstellbar ist. Da die Zeitkomplexität im Trainingsschritt ein zu vernachlässigendes Kriterium ist, wird der RNN-Algorithmus im Gegensatz zu [107], wo ein *average-link* Cluster-Kriterium benutzt wird, mit einem *complete-link* Kriterium verwendet. Somit ist eine genaue Festlegung des maximal erlaubten Abstands innerhalb eines Clusters möglich, wodurch eine minimale Ähnlichkeit zwischen Prototyp und jedem Deskriptor in einem Cluster in jedem Fall gewährleistet ist. Ergebnis des Clusterings ist eine Anzahl von Clustern, deren Beschreibung durch die Clustermittelpunkte die Prototypen des initialen Codebuchs für die trainierte Objektklasse ergibt.

Im nächsten Schritt wird nun auf Basis des initalen Codebuchs ein *implizites Modell der Objektform*, das *Implicit Shape Model (ISM)*, erstellt. Dieses besteht aus dem initialen Codebuch C und einer Ortsverteilung P_C, die den räumlichen Bezug der Merkmale zum Objektzentrum angibt. Diese wird unabhängig für jeden Codebuchprototypen angegeben und nicht parametrisch dargestellt, um eine Generalisierung durch eine parametrische Beschreibung, wie z.B. eine Gaußverteilung, zu vermeiden.

[4] RNN: Reciprocal Nearest Neighbor

Zur Erstellung der Repräsentation werden die im ersten Schritt auf den Trainingsbildern extrahierten Merkmale ein weiteres Mal betrachtet. Diese werden nun mit den Codebucheinträgen verglichen, wobei wie beim Erstellen der Cluster als Abstandsmaß die *Sum of Squared Differences (SSD)* genutzt wird. Der Abstand $\delta(\vec{x}, \vec{\zeta})$ zwischen einem Codebucheintrag mit dem Repräsentanten $\vec{x}$ und einem Merkmal mit Deskriptor $\vec{\zeta}$ berechnet sich dabei durch

$$\delta(\vec{x}, \vec{\zeta}) = \sum_{i=1}^{I} (x_i - \zeta_i)^2, \tag{3.16}$$

wobei I die Deskriptordimension ist. Hier wird die SSD verwendet, da diese bei einem vorgegebenen Schwellwert für den Maximalabstand effizient berechnet werden kann. So muss die Berechnung nur in wenigen Fällen, in denen der Abstand zwischen den Deskriptoren unter dem vorgegeben Schwellwert liegt, für den kompletten Deskriptor durchgeführt werden. Ansonsten kann die Abstandsberechnung abgebrochen werden sobald der Schwellwert überschritten wird.

Für ein Merkmal werden alle Codebucheinträge, deren Ähnlichkeit zum Merkmal über dem bereits beim Clustering genutzten Schwellwert τ_ρ^{train} liegt, aktiviert. Für jeden aktivierten Codebucheintrag wird der Versatz (offset) des aktivierenden Merkmals zum Zentrum des Objekts, dem das Merkmal zugeordnet wurde, der Ortsverteilung des Codebucheintrags hinzugefügt. Innerhalb der Verteilung erfolgt dabei eine Gewichtung der Einträge mit der Ähnlichkeit von Merkmal und Codebucheintrag und wird ausgehend von einer Gewichtung von 0 beim Schwellwert τ_ρ^{train} ansteigend bis zur Gewichtung von 1 beim Abstand von 0 modelliert:

$$p(\pi_{\vec{\zeta}}|C_{\vec{x}}) = \max\left(\frac{\delta(\vec{x}, \vec{\zeta}) - \tau_\rho^{train}}{-\tau_\rho^{train}}, 0\right). \tag{3.17}$$

Die Ortsverteilung für einen Codebucheintrag besteht somit aus einer Menge von gewichteten Koordinatenoffsets, welche die möglichen Positionen des Objektzentrums relativ zur Position der durch den Codebuchprototypen repräsentierten Merkmale angeben. Die Erstellung des ISM ist in Abbildung 3.4 schematisch dargestellt.

Das Ergebnis des Trainingsschritts ist ein Codebuch einer bestimmten Objektklasse, das als Basis für die Detektion von Objekten dieser Klasse genutzt werden kann. Eine visuelle Veranschaulichung eines Ausschnitts eines solchen Codebuch für das Beispiel der Klasse *Person* ist ebenfalls in Abbildung 3.4 dargestellt. Hier sind in den Zeilen verschiedene Codebucheinträge zu erkennen. Pro Eintrag sind einige diesem Eintrag zugeordnete Merkmale durch die zugehörigen Bildausschnitte dargestellt. Das Mittel dieser Bildausschnitte ergibt dann das visuelle Äquivalent zum Repräsentanten eines Codebucheintrags. Besonders relevant ist die Verteilung der Merkmale innerhalb eines einzelnen Codebucheintrags, da diese implizit das Objektmodell darstellt. Diese Verteilung kann eine eindeutige Ballung sein, falls die zugeordneten Merkmale den gleichen Teil des Objekts darstellen und dieser Objektteil immer mit dem gleichen Versatz zum Objektzentrum wahrgenommen wird. Dies ist allerdings nicht immer der Fall. Aufgrund vorhandener visueller Ähnlichkeiten von unterschiedlichen Objektteilen können weit gefächerte Ortsverteilungen entstehen. Diese Verteilungen sind aufgrund der Unsicherheiten, die hierdurch bei der Objektdetektion entstehen, nicht optimal. Solche Codebucheinträge können durch eine Verringerung des Schwellwerts τ_ρ^{train} für den maximalen Abstand von Deskriptoren innerhalb einer Klasse vermieden werden. Da hierdurch allerdings eine größere Anzahl an Clustern entsteht, kann dies eine zu große Spezialisierung auf die Trainingsdaten, welche bei gewünschter Detektion von Individuen einer Objektklasse die nicht im Trainingsmaterial enthalten sind, zu Problemen führen kann, bedeuten. Bei der Wahl des Schwellwerts τ_ρ^{train} ist also ein Kompromiss zwischen Generizität und Exaktheit der Repräsentation zu finden.

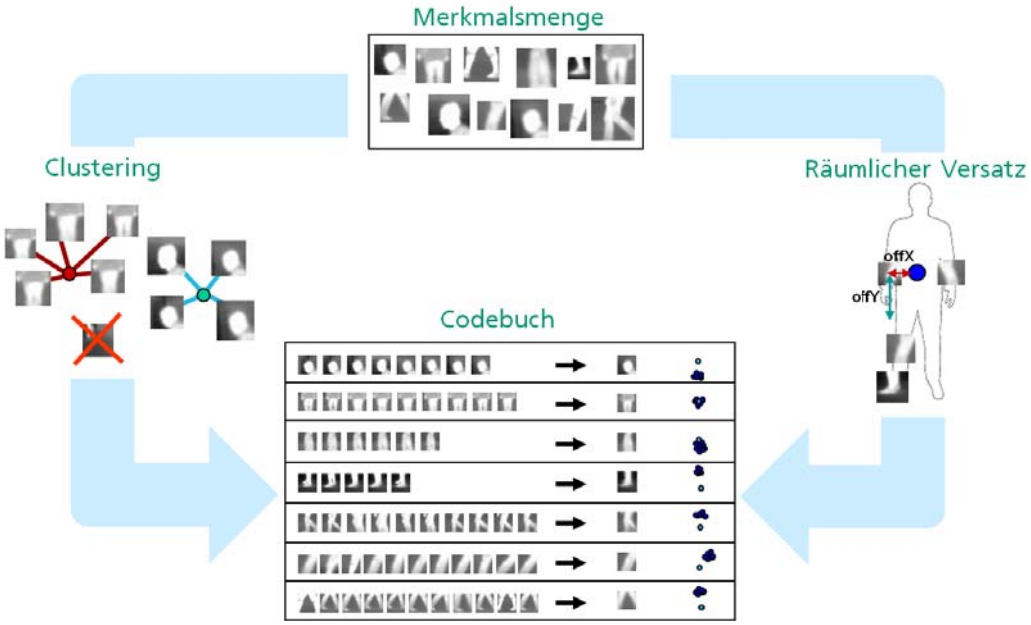


Abbildung 3.4: ISM-Trainingsschritt. Durch Clustering der Trainingsmerkmale wird im ersten Schritt ein initiales Codebuch mit Prototypen erstellt (linker Pfad). Im zweiten Schritt wird das ISM durch Zuordnung der Trainingsmerkmale zu Prototypen erstellt (rechter Pfad). Der Objektzentrumsversatz bildet dabei die räumliche Verteilung. Das resultierende Codebuch enthält eine Menge von Prototypen in Form von SIFT-Deskriptoren sowie eine Ortsverteilung der möglichen Objektzentrumspositionen.

Generieren von Templates für Objektkomponenten

Ziel dieser Arbeit ist unter anderem, ein Verfahren zu entwickeln, das über die einfache Objektdetektion hinaus die Möglichkeit bietet, Objektkomponenten zu erkennen und zu unterscheiden. Um dies zu ermöglichen, ist die Einbindung von Wissen über den Aufbau von Objekten notwendig, das über eine visuelle Beschreibung hinaus geht. Diese Einbindung kann direkt durch die Trainingsannotation erfolgen. Hierzu muss die Annotation über die Position des Objekts in den Trainingsdaten hinaus die Positionen bestimmter Objektteile enthalten. Dies können im Fall von Personen z.B. die Positionen von Kopf, Händen und Füßen sein. Die aus den Trainingsbilder extrahierten Deskriptoren können mit dieser Annotation versehen werden. Dazu werden Merkmale, die innerhalb eines Abstands $\Gamma_{dist,\beta}$ zur Position einer annotierten Objektkomponente liegen und deren Flächenabweichung zur Objektkomponente unter einem Schwellwert Γ_{area} liegt mit der Komponentenklasse aus der Annotation bezeichnet. Die Schwellwerte sind dabei abhängig vom Komponententyp.

Alternativ zu dieser Vorgehensweise ist es für den Fall, dass kein auf Objektkomponentenebene annotiertes Trainingsmaterial vorhanden ist, möglich, eine nachträgliche Annotation auf Ebene der Zuordnungen im Codebuch durchzuführen. Eine Voraussetzung hierfür ist es, dass eine einzelne Codebuch-Gruppierung auch eindeutig einer Komponentenklasse zuzuordnen ist. Dies ist selbst bei sehr kompakten Gruppen in der Praxis oftmals nicht der Fall, so dass die Extraktion der Komponenteninformation aus einer vorhandenen Annotation der zu bevorzugende Weg ist.

Falls eine solche Annotation nur für einen Teil der Trainingsdaten vorhanden ist, kann ein automatisches Annotationsverfahren auf Basis dieser unvollständigen Grundwahrheitsinformation durchgeführt werden [105].

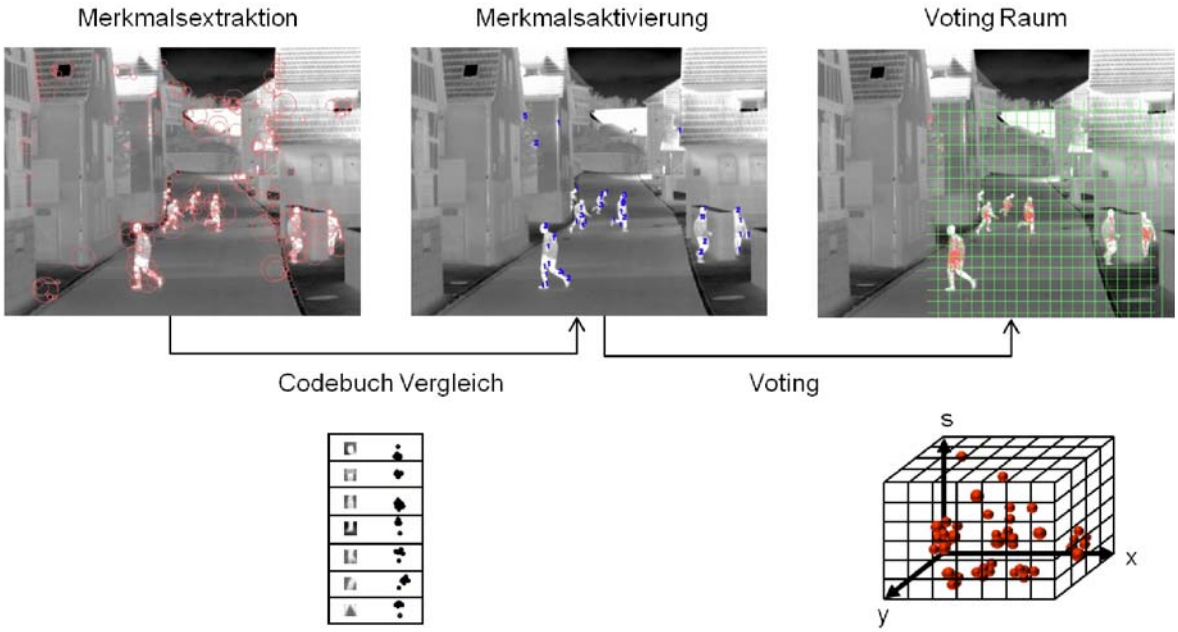


Abbildung 3.5: Ablauf Objektdetektion: Zunächst werden SIFT-Merkmale im Eingabebild extrahiert. Diese werden mit dem Codebuch verglichen, wobei für jedes Bildmerkmal alle Codebucheinträge aktiviert werden, deren Ähnlichkeit zum Bildmerkmal über einem Schwellwert τ_ρ^{det} liegen. Die Ortsverteilung der aktivierten Einträge stimmt ausgehend von der Position der aktivierenden Bildmerkmale für Objektzentrumspositionen in einem Hough-Voting Raum ab.

3.3 Detektion

Zur Detektion von Objekten in Eingabebildern wird das in Abschnitt 3.2 erstellte Codebuch verwendet.

3.3.1 Merkmalsbasierte Objektdetektion

Zur Detektion von Objekten in Eingabebildern erfolgt wiederum eine *Merkmalsextraktion* in diesen Eingabebildern. Die daraus hervorgehenden Deskriptoren werden zur Detektion von Objekten einer bestimmten Klasse mit dem im Trainingsschritt erstellten Codebuch dieser Klasse verglichen. Hierbei werden pro Bildmerkmal alle Codebucheinträge, deren Ähnlichkeit zum Bildmerkmal über einem vordefinierten Schwellwert τ_ρ^{det} liegt, aktiviert. Die Ähnlichkeitsbestimmung erfolgt dabei wie im Trainingsschritt zwischen Deskriptor und Repräsentant eines Codebucheintrags nach Formel 3.17 und wird für ein Bildmerkmal π_{img} und den i-ten Codebucheintag C_i mit $p(C_i|\pi_{img})$ bezeichnet.

Abbildung 3.5 zeigt die Ergebnisse des Vergleichs von im Eingabebild extrahierten Merkmalen mit dem Codebuch. Die Zahlen im mittleren Bild zeigen die Anzahl der aktivierten Codebucheinträge an. Jeder der aktivierten Codebucheinträge enthält eine Anzahl an Zentrumsversätzen, welche die räumliche Verteilung des durch den Repräsentanten beschriebenen visuellen Merkmals im Trainingsmaterial wiedergeben. Die Versätze wurden im Trainingsschritt mit einer Zuordnungsgewichtung versehen, die auf der Ähnlichkeit des Merkmals mit dem Repräsentanten beruht (siehe Formel 3.17). Sie geben mögliche Positionen des Objektzentrums relativ zu dem Merkmal an, das der Codebucheintrag repräsentiert. Zur Objektdetektion werden nun die Versätze aller aktivierten Prototypen in den Bildraum projiziert und stimmen somit für mögliche Positionen des Objektzentrums (siehe Abbildung 3.5, Bild rechts).

Hierbei wird der jeweilige Versatz in Abhängigkeit der Skalierungsstufen des Trainings- und Bildmerkmals adaptiert. Ein Versatz $\epsilon = \{\epsilon_x, \epsilon_y, \epsilon_s\}$, der von einer Übereinstimmung mit einem Bildmerkmal $\pi_{img} = \{\pi_{img,x}, \pi_{img,y}, \pi_{img,s}\}$ aktiviert wurde, stimmt dabei für die Bildposition

$$\begin{aligned} V_x &= \pi_{img,x} - \epsilon_x \left(\frac{\pi_{img,s}}{\epsilon_s} \right), \\ V_y &= \pi_{img,y} - \epsilon_y \left(\frac{\pi_{img,s}}{\epsilon_s} \right), \\ V_s &= \frac{\pi_{img,s}}{\epsilon_s}. \end{aligned} \tag{3.18}$$

Die Aussagekraft dieser Stimme[5] hängt dabei von der Ähnlichkeit $p(C_i|\pi_{img})$ zwischen Bildmerkmal π_{img} und Codebucheintrag C_i und der Zuordnungsstärke des Ortsverteilungseintrags im Codebuch $p(V_{\vec{k}}|C_i)$ ab. $p(C_i|\pi_{img})$ und $p(V_{\vec{k}}|C_i)$ sind jeweils durch die Deskriptorähnlichkeit nach Formel 3.17 bestimmt. Das Gewicht $V^w_{\vec{m}}$ einer Vote für Position $\vec{m}$ wird somit durch

$$V^w_{\vec{m}} = p(C_i|\pi_{img}) p(V_{\vec{k}}|C_i) \tag{3.19}$$

bestimmt.

Da den verschiedenen Codebucheinträgen im Trainingsschritt nicht unbedingt die gleiche Anzahl an Merkmalen zugeordnet werden, enthalten die Ortsverteilungen der Codebucheinträge unterschiedlich viele Einträge und geben somit auch unterschiedlich viele Votes ab. Da beim Voting allerdings keine Abhängigkeit von der absoluten Anzahl an zugeordneten Merkmalen entstehen soll, die lediglich von der Menge und Beschaffenheit des Trainingsmaterials abhängig ist, muss eine Normalisierung der Gewichtungen $p(V_{\vec{k}}|C_i)$ vorgenommen werden. Eine einfache Normalisierung der Summe der Stimmengewichte eines Codebucheintrags auf eins, wie sie in [107] durchgeführt wird, würde die Problematik der ungleichen Gewichtung von Codebucheinträgen allerdings lediglich in die andere Richtung, also hin zu einer stärkeren Gewichtung von Einträgen mit wenig zugeordneten Merkmalen verschieben. In diesem Fall würde eine Stimme, die als einzelne in einem Codebucheintrag vorkommt, ebenso stark gewichtet wie die Summe aller Stimmen eines anderen Codebucheintrags, selbst wenn dieser Eintrag eine wesentlich größere Anzahl an Stimmen enthält. Dies hätte eine in keiner Beziehung zur Aussagekraft stehende Stärke einer einzelnen Stimme zur Folge. Bei der Suche nach Objekthypothesen könnte dies unter Umständen dazu führen, dass Codebucheinträge mit wenigen zugeordneten Merkmalen aufgrund ihrer hohen Gewichtung alleine dazu ausreichen, eine Objekthypothese zu generieren. Die gleiche Problematik gilt für eine Normalisierung der Ähnlichkeiten $p(C_i|\pi_{img})$. Diese dürfen nicht abhängig von der absoluten Anzahl der für ein Merkmal aktivierten Codebucheinträge sein. Um diese Problematik bereits bei der Objektdetektion abzufangen, wird in dieser Arbeit im Gegensatz zum Original-ISM die Anzahl der genutzten Stimmen eines einzelnen Bildmerkmals für eine Objekthypothese auf eine beschränkt. Diese Vorgehensweise entspricht bei genauer Betrachtung auch eher der eigentlichen Motivation des Verfahrens, da ein einzelnes Bildmerkmal auch lediglich einmal Evidenz in eine Objekthypothese einbringen kann.

Nachdem durch Projektion der Ortsverteilungen der Voting-Raum, der einen kontinuierlichen Hough-Raum (vgl. [13] und [79]) darstellt, aufgebaut wurde, wird in diesem Raum nach Objekthypothesen gesucht. Diese Suche läuft in zwei Schritten ab:

Zunächst wird ein dreidimensionales *Hough-Akkumulator-Array* über dem kontinuierlichen Hough-Raum aufgebaut. Der Raum wird also diskretisiert. Die Quantisierung hängt dabei von der erwarteten Objektgröße ab [107]. Die Votes werden nun gewichtet in die jeweiligen Felder eingeteilt, wobei hier pro Bildmerkmal und Feld nur die jeweils stärkste Stimme gewertet wird. Eine Visualisierung dieses Voting-Raums ist in Abbildung 3.5 (rechts) zu sehen.

[5] Im Zusammenhang mit dem Hough-Voting-Raum werden diese Stimmen fortan als „Votes“ bezeichnet.

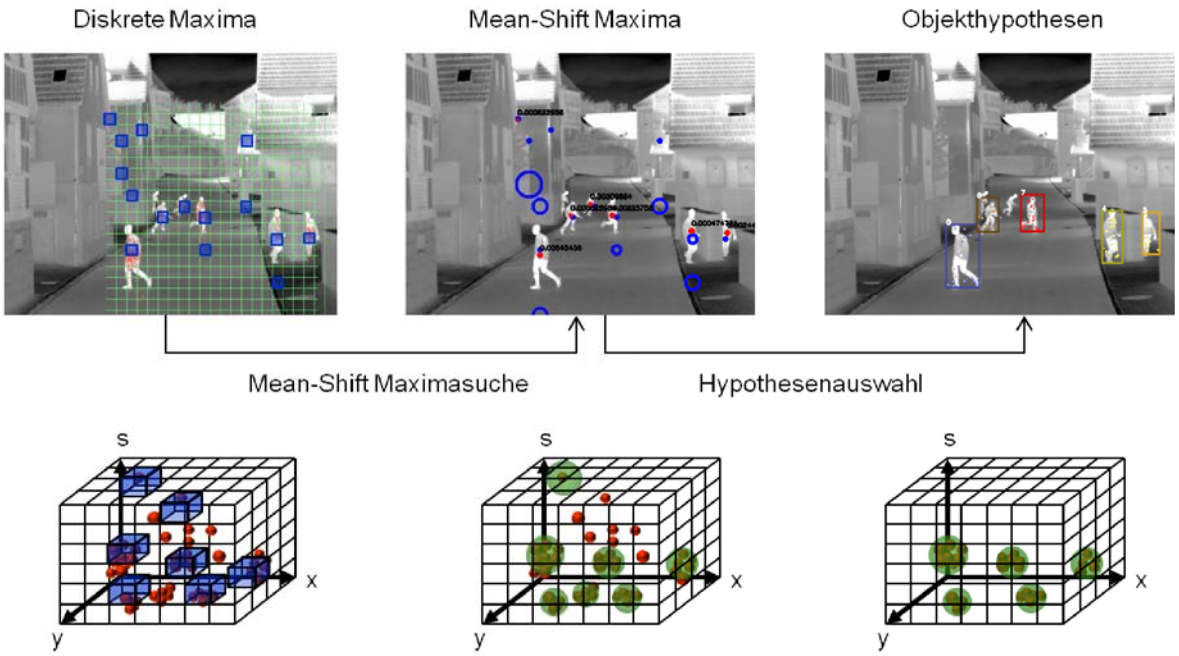


Abbildung 3.6: Ablauf Objektdetektion: Der Hough-Voting Raum wird zur Detektion initialer Maxima diskretisiert. Maxima im diskreten Raum bilden die Ausgangspunkte für die genaue Bestimmung der Maximaposition durch eine Mean-Shift-Mode-Estimation mit skalierungsadaptivem Kernel. Als Objekthypothesen werden alle Maxima übernommen, die einer Mindestanforderung an die Maximastärke genügen.

Im nächsten Schritt wird nun nach lokalen Maxima in diesem diskretisierten Hough-Raum gesucht (siehe Abbildung 3.6 (links)). Durch Anwendung eines Schwellwerts auf die Stärke der gefundenen Maxima (die Stärke ergibt sich aus der Summe der beinhalteten Vote-Stärken) werden zu schwache Maxima an dieser Stelle ausgefiltert. Die verbleibenden lokalen Maxima dienen als initiale Startpunkte für die genaue Maximabestimmung im nächsten Schritt. Diese wird durch eine *Mean-Shift-Mode-Estimation* [36, 38] im kontinuierlichen Hough-Raum durchgeführt (siehe Abbildung 3.6 (Mitte)). Dabei wird ein Kernel verwendet, dessen Größe sich nach der erwarteten Objektgröße $\Xi = (\Xi_x, \Xi_y)$ richtet. Zusätzlich wird die Kernelgröße in Abhängigkeit der Skalierungsstufe adaptiert, um den Größenunterschieden der Objekte auf den unterschiedlichen Skalierungsstufen Rechnung zu tragen. Die erwartete Objektgröße wird dabei im Trainingsschritt für die Referenz-Skalierungsstufe $s = 1.2$ bestimmt und für die anderen Skalierungsstufen anhand des Basiswerts berechnet.

Das Resultat dieser Schritte ist in Abbildung 3.6 zu sehen. Die blau dargestellten Kreise stellen die initialen Maxima im Voting-Raum dar. Die Größe ist hierbei ein Maß für die Skalierungsstufe des Maximums. Die roten Kreise geben die angepassten Positionen von Hypothesen nach der Mean-Shift-Mode-Estimation an. Diese sind mit der jeweiligen Stärke der Hypothese annotiert. Diese Stärke ergibt sich aus der Summe aller zu dieser Hypothese beitragenden Votes:

$$\mu_\gamma = \sum_{i=1}^{I} V_i^w. \tag{3.20}$$

Wobei zusätzlich mit dem Volumen des Kernels normalisiert wird, um den unterschiedlichen Skalierungsgrößen der Personen im Bild Rechnung zu tragen. Anhand dieser Hypothesenstärke können

Abbildung 3.7: Bild links: Überschneidungen von Detektionen auf Merkmalsebene. Die zwei roten Kreise zeigen zwei von einer Person generierten Maxima die als Objekthypothesen übernommen werden, die aber von teilweise gleichen Bildmerkmalen verursacht wurden. Bild rechts: die stärkste Detektion wurde ausgewählt. Nach neuer Zuverlässigkeitsbewertung der zweiten Detektion nach Entfernen der überschneidenden Merkmale, ist die Stärke dieser Detektion zu gering, um eine Objekthypothese aufrecht zu erhalten.

wiederum zu schwache Objekthypothesen ausgefiltert werden.

Jede der verbleibenden Objekthypothesen $\gamma \in \Gamma$ besteht aus der ihr zugeordneten Merkmalsmenge Π_γ, der Hypothesenstärke μ_γ und der Bildposition des Objektzentrums $\vec{\lambda^\gamma} = (\lambda_x^\gamma, \lambda_y^\gamma, \lambda_s^\gamma)$ (x-, y-Position in Bildkoordinaten sowie die Skalierungsstufe). Die Merkmalsmenge Π_γ bildet sich dabei aus allen Merkmalen, die zur Hypothese γ beitragen. Ein $\pi_\gamma \in \Pi_\gamma$ enthält das zugehörige Bildmerkmal π_{img}, das Offset ϵ, sowie die nach Formel 3.19 berechnete Gewichtung V^w der Bildmerkmal-Codebucheintrag Kombination.

3.3.2 Auflösen von Detektionsüberschneidungen

Bei der bisherigen Modellierung wurden die Objekthypothesen unabhängig voneinander betrachtet. In der Praxis befindet sich jedoch selten nur ein einzelnes relevantes Objekt in der Szene, sprich die Bildanalyse muss multiple Objekte gleichzeitig einbeziehen. Hierbei kann es zur Interaktion der Sensorsignale der Objekte kommen. Auf Merkmalsebene bedeutet dies, dass ein Merkmal potentiell an mehreren Objekthypothesen beteiligt sein kann. Da dies allerdings der Logik widerspricht, dass ein bestimmter Schlüsselpunkt nur auf einem einzelnen Objekt liegen kann, ist dies nicht wünschenswert. Insbesondere würde dies auch bedeuten, dass keine Eindeutigkeit in der Beziehung Bildmerkmal-Objekthypothese bestehen würde. Dieser Zustand ist für die auf der Objektdetektion aufbauende Objektverfolgung allerdings nicht tragbar, da hier Eindeutigkeit bei der Zuordnung von Merkmalen zu Objekthypothesen erforderlich ist.

Durch die in Abschnitt 3.3 beschriebenen Veränderungen am ursprünglichen Detektionsansatz, nämlich jedes Bildmerkmal nur einmal innerhalb einer Detektion abstimmen zu lassen, konnte die Abhängigkeit von Spezifika der Trainingsdaten minimiert werden. Da an dieser Stelle allerdings alle Detektionen unabhängig voneinander betrachtet wurden, konnte noch nicht gewährleistet werden, dass alle Detektionen bzgl. ihrer Merkmalsmengen disjunkt sind. Da solche Überschneidungen zwischen den Merkmalsmengen von Detektionen aber nicht erwünscht sind (ein visuelles Merkmal wird im Normalfall nur von einem Objekt generiert) und in der Praxis zu doppelten Hypothesen führen, d.h. eine Person im Bild generiert mehrere Detektionen, soll hier Eindeutigkeit hergestellt werden. Dazu werden

alle Detektionen betrachtet und auf Überschneidungen auf Merkmalsebene geprüft. Detektionen, die sich in mindestens einem Merkmal überschneiden, bilden eine Gruppe[6], wobei diese Gruppen transitiv erweitert werden. Für jede Gruppe wird dann separat Disjunktheit auf Merkmalsebene hergestellt.

In jedem Durchlauf wird die Detektion mit der höchsten Zuverlässigkeit ausgewählt, aus der Gruppe entfernt und der Menge der validen Detektionen hinzugefügt. Alle Merkmale dieser Detektion, die ebenfalls in anderen Detektionen vorkommen, werden aus diesen entfernt. Für alle verbleibenden Detektionen wird die Zuverlässigkeit auf Basis der jetzt noch enthaltenen Merkmale neu berechnet. Sollte die neue Zuverlässigkeit unter dem Schwellwert für valide Detektionen liegen (siehe Abschnitt 3.3), wird die Detektion komplett entfernt. Diese Prozedur wird so lange fortgeführt, bis keine der Gruppen mehr Detektionen enthält. Ergebnis dieses Nachbearbeitungsschritts ist eine Menge von validen Detektionen, deren Merkmalsmengen disjunkt sind.

Neben der Tatsache, dass durch die Herstellung der Disjunktheit, wie in Abbildung 3.7 zu sehen ist, Falschalarme in Form von doppelten Hypothesen minimiert werden und hierdurch ein konsistentes Bild der Umgebung bei Anwesenheit mehrerer auf Signalebene interagierender Personen erzeugt wird, ist die Disjunktheit essentiell in den folgenden Verarbeitungsstufen „Personenverfolgung“ und „Personenwiedererkennung“, da hier eine eindeutige Zuordnung von Bildmerkmalen zu Objekthypothesen notwendig ist.

3.4 Detektion von Objektkomponenten

Um ein umfassenderes und detailreicheres Bild der aktuellen Situation aufzubauen, reicht es oftmals nicht aus, eine Person durch eine bounding box in der Bildebene darzustellen. Will man z.B. Handlungen von Personen erkennen, so ist eine Modellierung auf Objektebene nicht mehr ausreichend und es wird eine detaillierte Beschreibung der Personen notwendig. Eine solche Beschreibung kann z.B. die Beschreibung durch die Positionen von einzelnen Körperteilen der Person sein. Aufbauend auf einer solchen Beschreibung kann eine detaillierte Interpretation der Situation, bzw. zunächst eine 3D Rekonstruktion einer Person [96] erfolgen.

Die Klassifikation von Objektkomponenten wird durch die in dieser Arbeit eingeführten Erweiterungen des ISM möglich. Dies bezieht sich insbesondere auf die herbeigeführte Eindeutigkeit der Beziehung zwischen Bild- und Hypothesenmerkmal. Im Gegensatz zum in [105] vorgestellten Ansatz zur Klassifikation von Objektkomponenten zielt der hier vorgestellte Ansatz nicht auf die Verbesserung der Detektion, sondern auf die Klassifikation von Bildmerkmalen als bestimmte semantische Objektteile ab.

Die Detektion von einzelnen Objektkomponenten einer vorhandenen Hypothese ist durch die Kombination von Annotation der Trainingsbeispiele auf Komponentenebene und der Objektdetektion durch ein merkmalsbasiertes Verfahren implizit bereits vollzogen. So kann eine Annotation von Merkmalen der Trainingsmenge durch die Übertragung in das Codebuch auch bei der Detektion genutzt werden. Hier können die Annotationen der Ortsverteilung (durch Votes repräsentiert), die zu einer Objekthypothese beitragen, wie in Abbildung 3.8 durch die grüne Verbindung dargestellt, direkt auf die korrespondierenden Bildmerkmale übertragen werden. Da nicht grundsätzlich alle Codebuch-Merkmale mit einer Annotation versehen sein müssen, können an dieser Stelle auch die in Kapitel 3.2 beschriebenen *Komponentenmuster* zur Annotation von Merkmalen mit einer Objektkomponentenklasse genutzt werden. So können Merkmale, die im ersten Schritt nicht mit einer Semantik versehen wurden, durch Vergleich mit den Komponentenmustern (Abbildung 3.8, rote Verbindung) einer Klasse zugeordnet werden.

An dieser Stelle sind noch weitere Verfahren zum Auffinden noch nicht detektierter Objektkomponenten möglich. So könnten auf Basis der Trainingsdaten automatisch Abhängigkeiten zwischen verschiedenen Objektkomponenten gelernt und so fehlende Teile abhängig von den detektierten Komponenten

[6] Dies ist keine Gruppe im mathematischen Sinn.

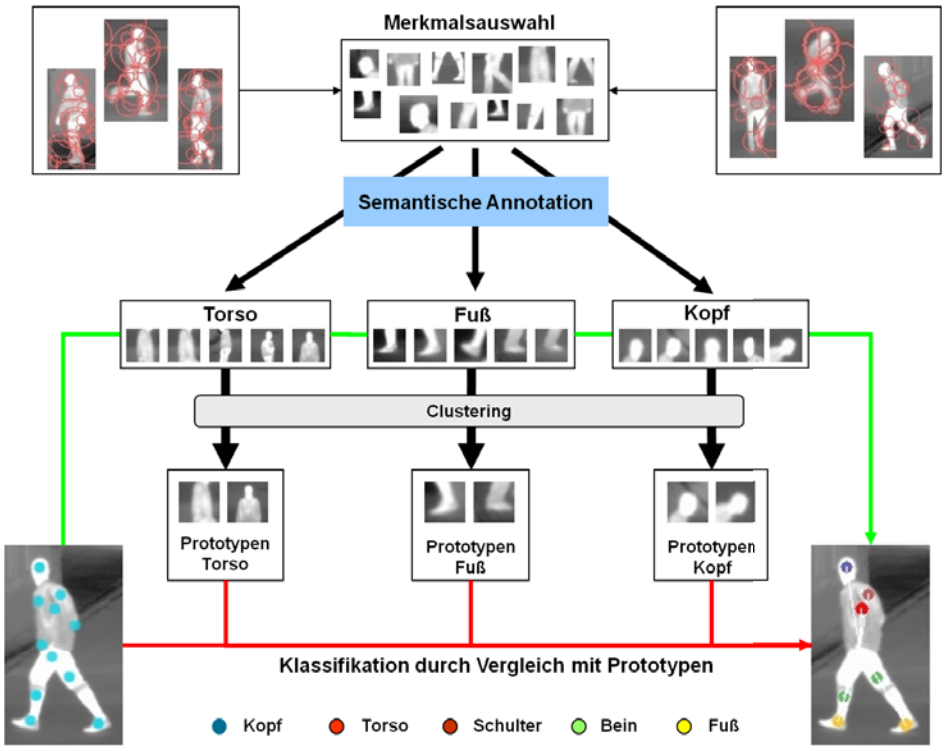


Abbildung 3.8: Ablauf Körperteilklassifikation.

gezielt im Bild gesucht werden. Diese Abhängigkeiten können bei Doppeldeutigkeiten (z.B. zwei Köpfe in einer Hypothese) auch zur Validierung und zur Auswahl der wahrscheinlichsten Komponentenkonfiguration genutzt werden (vgl. auch [87, 88] sowie Abschnitt 3.6.2).

3.5 Hierarchische Objektdetektion

In diesem Abschnitt wird eine weitere Anwendung der ISM-basierten Objektdetektion vorgestellt, welche prinzipielle Vorteile des ISM-Ansatzes gegenüber Klassifikatoren wie den Sliding-Window-Ansätzen oder dem Viola-Jones Ansatz erschließt.

Zur wirklichen Analyse einer Szene oder einer bestimmten Situation ist es oftmals notwendig, nicht nur Objekte einer einzigen relevanten Klasse, wie z.B. Personen oder Autos, zu detektieren, sondern in der Praxis sind häufig viele unterschiedliche Objektklassen relevant. Sind die relevanten Objekte räumlich separiert, so kann eine Objektdetektion der einzelnen Objekte direkt mit mehreren, jeweils auf Trainingsdaten der relevanten Klasse trainierten Detektoren stattfinden. Schwierigkeiten entstehen hier allerdings, wenn die Objekte nicht räumlich getrennt sind, sondern eine Teil-von-Beziehung eingehen. Ein typisches Beispiel, welches häufig in der Praxis auftritt, sind z.B. Personen auf Fahrzeugen. Klassifikatoren wie z.B. die Sliding-Window-Ansätze haben, nachdem eine Bildregion z.B. als Auto klassifiziert wurde, keine Möglichkeit darauf zu schließen, welche Teile der Bildregion tatsächlich an der Klassifikation beteiligt waren, also welche Bereiche der Bildregion tatsächlich auf ein Auto rückschließen lassen. Im Gegensatz dazu bietet das ISM-basierte Detektionsverfahren Informationen darüber, welche Bildmerkmale am Zustandekommen einer Objekthypothese beteiligt waren. Insbesondere mit den in Abschnitt 3.3.2 vorgestellten Erweiterungen besteht Eindeutigkeit, welche Merkmale

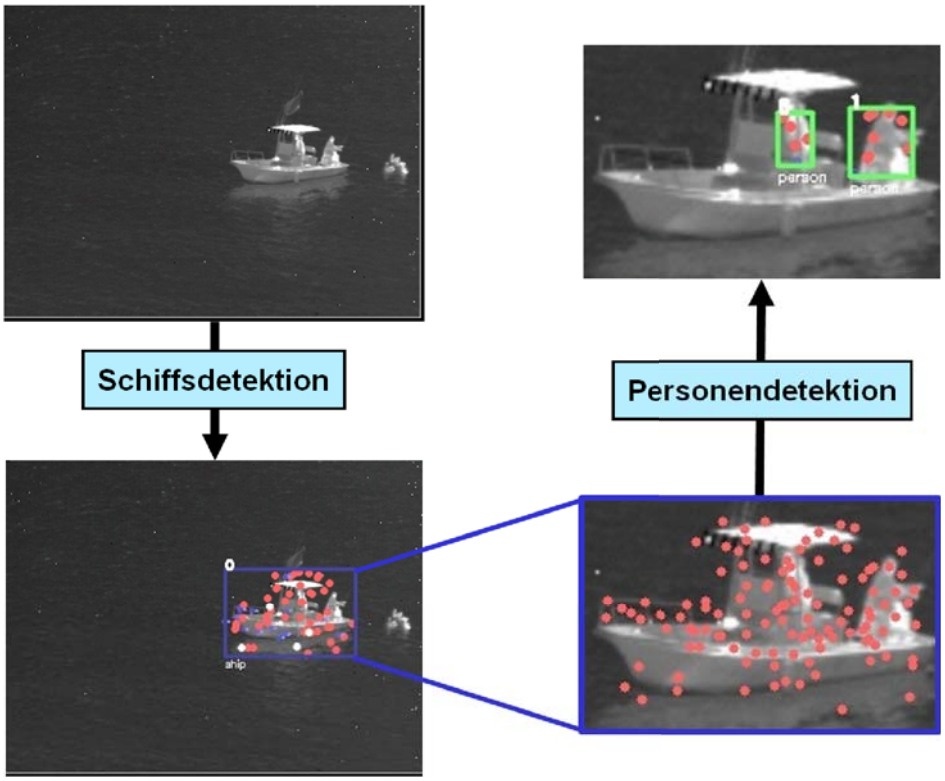


Abbildung 3.9: Ablauf der hierarchischen Objektdetektion am Beispiel „Personen auf Schiff“.

für welche Objekthypothese gestimmt haben.

Bildregionen, die zwar in der das Objekt umschließenden bounding box liegen, die aber nicht an der Objekthypothese beteiligt waren können somit objektspezifische Eigenheiten sein die nicht im generalisierten Objektmodell enthalten sind, oder aber zu einem anderen Objekt gehören. In Fällen, in denen zu erwarten ist, dass ein Objekt ein anderes subsumiert, wie z.B. bei Fahrzeugen, die typischerweise (noch) nicht ohne Fahrzeugführer unterwegs sind, kann diese Information über Merkmale, die nicht an der ursprünglichen Hypothese beteiligt waren, zusammen mit der Erwartung einer bestimmten Objektklasse im Kontext des Hauptobjekts zur Lokalisierung eines subsumierten Objekts genutzt werden.

Ein solches Beispiel ist in Abbildung 3.9 zu sehen. Hier sind zwei Personen auf einem Schiff anwesend. Nach Detektion des Schiffes können diese Personen in einem zweiten Detektionsschritt, der auf Basis der Bildmerkmale im Bereich des Schiffes durchgeführt wird, detektiert werden. Wichtig ist, dass beim zweiten Detektionsschritt nur die Merkmale betrachtet werden, die nicht an der Objekthypothese des Schiffes selbst beteiligt waren. Somit wird sichergestellt, dass das Bild der Umgebung konsistent bleibt. In der Praxis wird hierdurch auch vermieden, dass Teile des Schiffes Falschhypothesen der Klasse Person generieren.

In manchen Fällen ist es hilfreich, die Detektion der nächsten Hierarchiestufe mit einer anders parametrisierten Merkmalsextraktion durchzuführen. Dies kann z.B. der Fall sein, wenn auf dieser Stufe eine sehr geringe Objektgröße erwartet wird. In diesem Fall könnte die Merkmalsextraktion zur Extraktion von Merkmalen geringerer Größe parametrisiert werden. Wichtig ist, dass in diesem Fall ein zusätzlicher Vergleichsschritt notwendig wird, um sicherzustellen, dass die auf der letzten Stufe an einer Objekthypothese beteiligten Merkmale nicht in die Detektion auf dieser Stufe einbezogen werden.

Abbildung 3.10: Beispielergebnisse der Körperteilklassifikation. Dargestellt sind Kopf(blau), Torso(rot), Schulter(hellrot), Bein(grün) und Fuß(gelb).

3.6 Auswertung

Die Auswertung der reinen Personendetektion nach dem in diesem Kapitel vorgestellten Schema erfolgt, um die prinzipielle Funktion des Objektdetektionsansatzes zu demonstrieren, auf drei Bildfolgen unterschiedlichen Schwierigkeitsgrads. Eine umfangreichere und detaillierte Evaluierung der Detektionsperformance erfolgt bei der Auswertung der Personenverfolgung in Kapitel 4.

3.6.1 Trainingsdaten

Ein essentieller Punkt bei der Auswertung eines Objektdetektors ist die Wahl der Trainingsdaten. Der hier vorgestellte Personendetektor wird mit 97 Beispielen von Personen (8 verschiedene Personen), die von unterschiedlichen Blickpunkten sichtbar sind und mit unterschiedlichen Originalauflösungen aufgezeichnet wurden, trainiert. Die Trainingsdaten sind mit einer Referenzsegmentierung, die zur Auswahl der für die Objektklasse relevanten Bildmerkmale genutzt wird, annotiert. Zusätzlich werden die Bildmerkmale mit Körperteilsemantik annotiert.

3.6.2 Körperteilklassifikation

Die Körperteilklassifikation erfolgt nach dem in Abschnitt 3.4 beschriebenen Schema. Beispielergebnisse der Körperteilklassifikation sind in Abbildung 3.10 zu sehen. Die Menge der betrachteten Körperteile umfasst hier: *Kopf, Torso, Schulter, Bein und Fuß.* Wie zu sehen ist, ist es nicht in allen Fällen möglich die Körperteile korrekt zu erkennen. Dies kann zum Einen daran liegen, dass an den jeweiligen Bildpositionen kein Schlüsselpunkt zur Berechnung eines SIFT-Merkmals gefunden wurde. Zum Anderen kann es sein, dass ein dort gefundenes Merkmal nicht in die Personenhypothese integriert wurde oder die Deskriptorähnlichkeit zu den in den Trainingsdaten vorhandenen Körperteilen zu gering ist. Hierdurch sind die Einzelbildklassifikationen häufig unvollständig, bzw. in manchen Fällen sind Doppeldeutigkeiten vorhanden. Um die Klassifikationsergebnisse für eine Interpretation auf höherer Ebene (z.B. zur Aktionserkennung) nutzen zu können, müssen diese Probleme gelöst werden. Ein vielversprechender Ansatz hierfür wäre eine zeitliche Betrachtung der Körperteile über das Einzelbild hinaus sowie das Einbringen von a-priori Wissen (z.B. wieviele Köpfe hat eine Mensch).

3.6.3 Bewertungskriterien

Zur Auswertung der Personendetektion werden die in den Testsequenzen sichtbaren Personen mit bounding boxes annotiert. Als notwendiges Kriterium für die Annotation wird eine Sichtbarkeit von mindestens ca. 75% der Person gewählt.

Zur Bewertung der Detektionsperformance werden der *recall* und die *Falschalarme pro Bild*, welches Standardmaße im Bereich der Objektlokalisation sind, genutzt:

$$recall = \frac{\#\text{Korrekte Detektionen}}{\#\text{Objekte in der Grundwahrheit}}. \tag{3.21}$$

Zur Bewertung einer Detektion als korrekte Detektion oder als Falschalarm werden an dieser Stelle zwei unterschiedliche Kriterien angewendet. Das *bounding box Kriterium (BBI)* bewertet eine Detektion als korrekte Detektion falls sich das Detektionszentrum innerhalb der bounding box der Grundwahrheit befindet.

Das *Überschneidungskriterium (OL)* bewertet Detektionen anhand der Überschneidung der bounding boxes der Grundwahrheit und der Objekthypothese. Die Überschneidung wird dabei durch das *Intersection over Union* Kriterium (vgl. Jaccard-Index [82]) bewertet:

$$\text{Überscheidung} = \frac{\text{Fläche}(B_p \cap B_{gt})}{\text{Fläche}(B_p \cup B_{gt})}. \tag{3.22}$$

Das erste Kriterium wird an dieser Stelle gezielt gewählt, um die reine Detektionsperformance unabhängig von der geometrischen Genauigkeit zu bewerten. Speziell im hier vorliegenden Fall, bei dem die bounding box einer Detektion auf Basis der für eine Objekthypothese abstimmenden Merkmale berechnet wird, würde eine Detektion, die nur den Oberkörper einer Person enthält, bei Nutzung des Überschneidungskriteriums evtl. als Falschalarm gewertet. Um aber auch die Akkuratheit einer Detektion zu bewerten, wird das Überschneidungskriterium mit unterschiedlichen Anforderungen an den Überschneidungsgrad ausgewertet.

In jedem Fall wird pro Grundwahrheitsobjekt nur eine einzelne Detektion, welche das Kriterium erfüllt, als korrekte Detektion gewertet. Alle anderen Detektion werden, sofern sie das Kriterium nicht für ein anderes Grundwahrheitsobjekt erfüllen, als Falschalarme gewertet. Sofern mehrere Detektion das Kriterium erfüllen, wird diejenige Detektion zur Auswertung herangezogen, die das jeweilige Kriterium am besten erfüllt (geringster Abstand des Objektzentrums, bzw. größte Überschneidung).

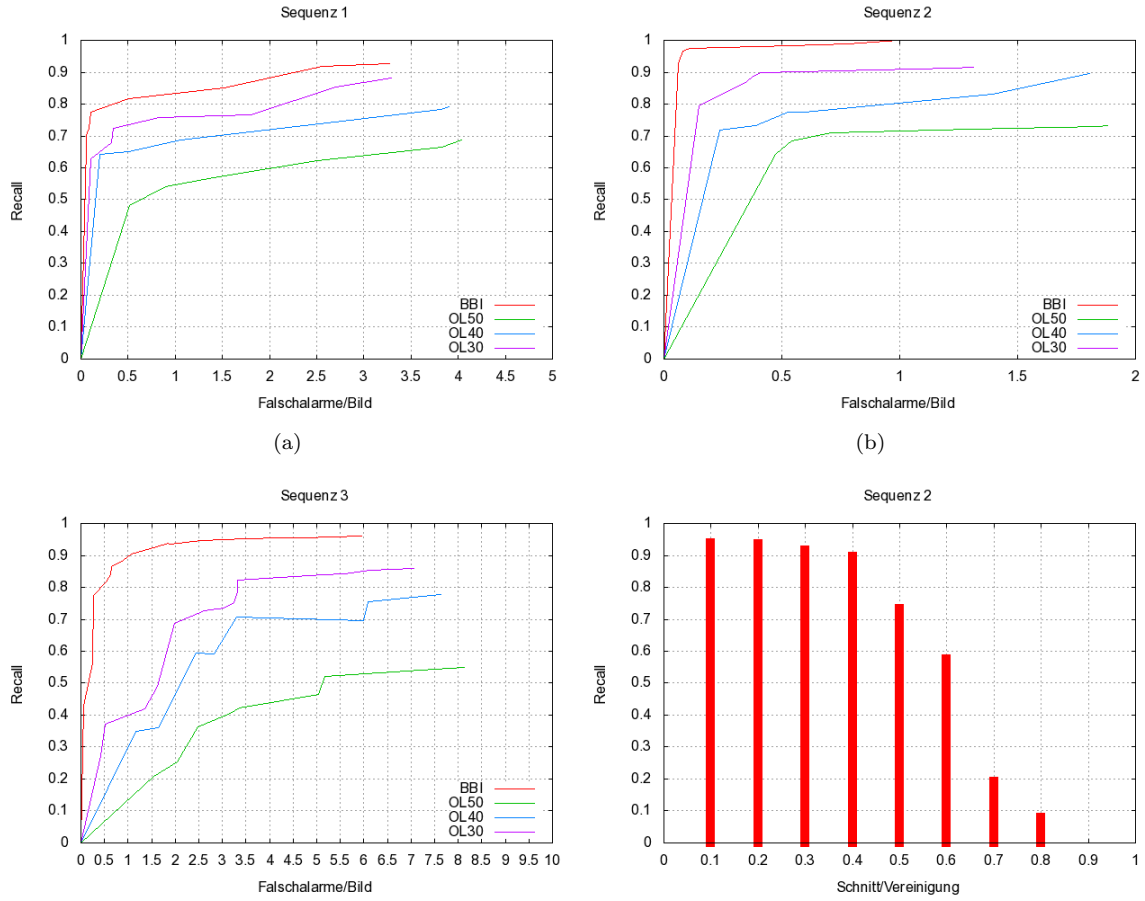


Abbildung 3.11: ROC-Kurven (Receiver-Operator-Characteristics) für (a) Sequenz 1, (b) Sequenz (2), (c) Sequenz 3. Es ist jeweils der recall in Abhängigkeit der Falschalarme pro Bild dargestellt. Jedes Diagramm zeigt vier Graphen für die unterschiedlichen Auswertungskriterien. BBI: Bounding box Kriterium. OL30/40/50: Schnitt/Vereinigung Kriterium mit 30, 40 und 50% Überschneidungsforderung. (d) Auswertung der Detektionsperformance (Recall) von Sequenz 2 bei verschiedenen Überschneidungsforderungen.

3.6.4 Quantitative Auswertung

Die quantitative Auswertung wird in drei Sequenzen vorgenommen, die unterschiedliche Anforderungen an die Objektdetektion stellen.

Sequenz 1 enthält insgesamt 301 Auftreten von 6 unterschiedlichen Personen, die in einer ähnlichen Größe zu sehen sind. Die Personen laufen von rechts nach links durch den Sichtbereich der Kamera, wobei Teilverdeckungen zwischen Personen auftreten. Die Sequenz wurde von einer bewegten Kamera mit einer Auflösung von 640x480 aufgezeichnet. Die Auswertung dieser Sequenz in Abbildung 3.11 (a) zeigt Kurven für den recall bei verschiedenen Auswertungskriterien (OLx: Bounding box Überschneidung mit einer Überschneidungsanforderung von mindestens x%; BBI: Bounding box Kriterium) in Abhängigkeit der Falschalarme pro Bild. Beispieldetektionen dieser Sequenz sind in der ersten Reihe von Abbildung 3.12 zu sehen.

Abbildung 3.12: Beispielergebnisse der Personendetektion. Reihe1: Sequenz 1, Reihe 2: Sequenz 2, Reihe 3: Sequenz 3.

Sequenz 2 aus dem OTCBVS Datensatz [46] enthält 763 Auftreten von 2 unterschiedlichen Personen. Die Personen sind in einer ähnlichen, hier sehr geringen Größe von ca. 10x25 Pixeln, zu sehen. Die beiden Personen bewegen sich unabhängig voneinander in der Szene so dass es zu keinerlei Verdeckungen der Personen kommt. Die Bildfolge wurde von einer statischen Kamera mit einer Auflösung von 320x240 aufgezeichnet. Die Auswertung in Abbildung 3.11 (b) zeigt, dass sich die Detektionsperformance an sich (BBI Kurve) schon bei einer niedrigen Falschalarmrate auf einem hohen Niveau befindet und sich diese auch bei höheren Falschalarmraten nicht mehr wesentlich verbessert. Die Detektionsperformance an sich ist in dieser Sequenz also sehr gut, wobei die Detektionsgenauigkeit, wie an den Kurven der Überschneidungskriterien zu sehen ist, im Vergleich zur guten Detektionsperformance eher gering ist. Dies ist auch in der detaillierten Auswertung für unterschiedliche Überschneidungskriterien in Abbildung 3.11 (d) zu sehen. Hier zeigt sich der Abfall der Detektionsperformance mit ansteigendem Überschneidungsgrad. Dies zeigt aber wiederum auch, dass die Detektionsqualität an sich gut ist, aber eher ungenau in Form der bounding box. Die Verbesserung dieser Genauigkeit spielt auch eine Rolle bei der im nächsten Kapitel beschriebenen Erweiterung der Detektion zum Tracking. Beispieldetektionen dieser Sequenz sind in Reihe zwei von Abbildung 3.12 zu sehen.

Sequenz 3 enthält 1471 Auftreten von 8 unterschiedlichen Personen und wurde mit einer Auflösung von 640x480 von einer bewegten Kamera aufgenommen. Diese Sequenz bringt die größten Herausforderun-

gen für die Personendetektion mit sich, da hier Personen auf sehr unterschiedlichen Skalierungsstufen mit einer Größe von 10x30 bis zu 90x220 Pixeln auftreten. Wie in den Beispielbildern in der unteren Reihe von Abbildung 3.12 zu sehen ist, nehmen einige Personen im Hintergrund der Szene nur wenige Bildpixel ein, wohingegen Personen im Vordergrund einen großen Anteil der gesamten Bildfläche einnehmen. Zusätzlich kommen hier erhebliche Verdeckungen vor, die durch das Kreuzen der Personenlaufwege zustande kommen. Die Performancekurven in Abbildung 3.11 (c) zeigen, dass die reine Detektionsperformance auch hier, mit einem recall von mehr als 90% bei weniger als 1.5 Falschalarmen pro Bild, gut ist. Allerdings zeigt sich wie bei der Auswertung von Sequenz 2 auch hier, dass die Detektionsgenauigkeit eher gering ist. Bei Anwendung des 50% Überschneidungskriteriums fällt die Performance stark, auf 50% recall bei 5 Falschalarmen pro Bild, ab. Diese Ungenauigkeiten sind auch in den Beispieldetektionen in der unteren Reihe von Abbildung 3.12 zu sehen.

Insgesamt ergibt die Auswertung also, dass die Detektion von Personen mit dem hier vorgestellten Ansatz prinzipiell möglich ist, dass diese aber hinsichtlich der zeitlichen Stabilität und insbesondere auch der Genauigkeit noch Verbesserungspotential bietet.

Kapitel 4

Personenverfolgung

Um Anwendungen höherer Komplexitätsstufen wie z.B. eine Situationsanalyse zu ermöglichen, bedarf es eines umfassenden Bildes der aktuellen Situation. Wie bereits in Abschnitt 3.4 erwähnt, reicht hier eine einfache Modellierung einer Person durch eine bounding box oftmals nicht aus. Auch bei einer detaillierten Modellierung einer Person durch einzelne Körperteile bietet die Einzelbilddetektion allerdings nur eine Momentaufnahme. Diese ist nur in den seltesten Fällen ausreichend um eine Interpretation, wie z.B. eine Aktionserkennung, durchzuführen, da alle Handlungsabläufe immer auch eine zeitliche Ausdehnung haben. Aus diesem Grund ist es notwendig, eine zeitliche Analyse der relevanten Weltausschnitte durchzuführen. Bei der personenzentrierten Analyse bedeutet dies eine Modellierung und schritthaltende Aktualisierung des (Bild-)Zustands einer Person im System. Konkret heißt dies, dass die aus dem Bild extrahierbaren Eigenschaften einer Person im System mitgeführt und aktualisiert werden. Die Modellierung bezieht sich bei dieser *Personenverfolgung (Tracking)* typischerweise auf die Position und Erscheinung der Person im Bild. Hierbei wird die Erscheinung einer Person häufig durch Farbe, Form und Textur modelliert.

Ein wichtiger Aspekt der Personenverfolgung ist die korrekte Beibehaltung von Objektidentitäten über der Zeit. In der Praxis bedeutet dies, dass Observationen einer Person auch immer der gleichen Objektidentität im Trackingsystem zugewiesen werden. Dies ist von immenser Bedeutung, da nur bei konsistenter Wahrung der Objektidentität über der Zeit auch korrekte Bewegungstrajektorien extrahiert werden können. Diese wiederum sind Voraussetzung für jede Art der weiteren Interpretation.

Die Aufgabe der Identitätswahrung beim Tracking entspricht in der Praxis dem Problem der Datenassoziation. Die Datenassoziation bezieht sich dabei auf die Assoziation von neuen Bilddaten mit im System bereits vorhandenen Daten auf Basis der für eine Person modellierten Eigenschaften. Beim Tracking gibt es hierbei zwei grundlegend verschiedene Ansätze.

Im *modellbasierten Tracking* wird ausgehend von einem existierenden Modell, das z.B. ein Farbhistogramm einer Bildregion oder ein Template in anderer Form sein kann, *top-down* nach einem Objekt im Bild gesucht. Diese Herangehensweise setzt voraus, dass das Objekt vorher bekannt ist, oder dass es durch andere Verfahren, wie z.B. einen dedizierten Objektdetektor, initial akquiriert wurde. Hierbei findet die Datenassoziation also durch direkte Suche eines bestimmten Objekts in den Daten statt. Im Gegensatz dazu basieren *tracking-by-detection* Ansätze darauf, dass ein Prozess *bottom-up* Informationen über mögliche Objekte im Bild liefert. Im einfachsten Fall kann dies ein Änderungsdetektionsverfahren sein, welches die Positionen von Vordergrundsegmenten liefert. Aufgrund des jüngsten Fortschritts im Bereich der trainierbaren Objektdetektionsverfahren werden allerdings immer häufiger auch diese Verfahren eingesetzt. Unabhängig von den speziellen Eigenschaften liefert das Verfahren pro Zeitpunkt, d.h. pro Bild, eine Anzahl von Objekthypothesen. Zum Tracking muss im Nachhinein eine Datenassoziation durchgeführt werden. Hierzu werden häufig auf Basis der bounding box, oder falls vorhanden einer Segmentierung, Merkmale berechnet, die dann als Observation zur Datenassoziation

genutzt werden.

Beide Ansätze haben Nachteile. So fließt bei tracking-by-detection Verfahren kein Wissen über die Historie und bekannten Eigenschaften der zu verfolgenden Objektindividuen in die Detektion der Objekte ein. Beim modellbasierten Tracking muss zunächst ein Modell des zu verfolgenden Objekts vorhanden sein (z.B. durch manuelle Auswahl einer Bildregion) und dieses Modell muss so beschaffen sein, dass es auch bei umgebungsbedingten Veränderungen der Objekterscheinung im Bild, noch zur Beschreibung des Objekts, und damit zum Tracking geeignet ist.

In diesem Kapitel wird ein neues Verfahren vorgestellt, welches diese beiden Ansätze kombiniert, weiterentwickelt und bei dem Nachteile der einzelnen Ansätze durch Kombination behoben werden. Dazu wird eine Weiterentwicklung des in Kapitel 3 vorgestellten und in [107] eingeführten Implicit Shape Model (ISM) vorgestellt, die das für die Objektdetektion entwickelte Modell zur Nutzung für die Objektverfolgung ausbaut und dabei die Objektdetektion und -verfolgung als integriertes Problem angeht. Hierdurch wird zum Einen eine Verbesserung der Detektionsqualität (vgl. Abschnitt 3.6.4) hinsichtlich der Genauigkeit der Detektion in Form von bounding boxes und der zeitlichen Konsistenz der Detektion erreicht. Insbesondere wird aber auch eine Herangehensweise vorgestellt, welche die Herausforderungen bei der Objektverfolgung, nämlich die Identitätswahrung und die schritthaltende Aktualisierung des Objektmodells zur Behandlung von umgebungsinduzierten und objekteigenen Erscheinungsveränderungen des Objekts, durch einen integrierten Ansatz ohne Nutzung spezieller Heuristiken löst. Ein weiterer Vorteil, der sich hierdurch im Vergleich zu Verfahren, die Detektion und Tracking getrennt voneinander durchführen, ergibt, besteht in der impliziten Behandlung von Verdeckungssituationen.

Bei der Erweiterung des ISM zum Tracking wird die ausschließliche Verwendung von SIFT-Merkmalen ohne Nutzung von Farbe beibehalten, was die in Abschnitt 2.3 beschriebene Sensorgenerizität sicherstellt. Insbesondere erfolgt auch hier keine Einbindung von objektspezifischem Wissen oder a-priori Wissen über die Szene, wodurch auch die Objekt- und Szenariogenerizität gewährleistet sind.

In Abschnitt 4.1 wird das Trackingverfahren vorgestellt. Hierbei wird in Abschnitt 4.1.1 zunächst auf den essentiellen Punkt – die Datenassoziation auf Basis von SIFT-Merkmalen – eingegangen. Abschnitt 4.1.2 führt die Erweiterung des ISM zum Tracking-ISM ein und Abschnitt 4.1.3 beschreibt die Durchführung des Trackings im Hough-Voting-Raum. In Abschnitt 4.2 wird der Aufbau und die Rolle von Identitätsmodellen beim Tracking beschrieben.

Die beim Tracking verwendete Dynamikmodellierung wird in Abschnitt 4.3 eingeführt. Hierbei wird in Abschnitt 4.3.1 zunächst eine einfache Kalman-Filter basierte Modellierung beschrieben und in Abschnitt 4.3.2 ein neues Modell zur integrierten Bewegungskompensation eingeführt, welches ein korrektes Tracking auch bei stark bewegter Kamera gewährleistet. Die Auswertung des ISM-Trackings in verschiedenen Bildfolgen aus dem infraroten und sichtbaren Spektralbereich sowie der Performancevergleich mit anderen Verfahren erfolgt in Abschnitt 4.5.

4.1 Objektverfolgung im Implicit Shape Model

Der in Kapitel 3 vorgestellte Objektdetektionsansatz arbeitet datengetrieben ausgehend von Bildmerkmalen[1]. In diesem Kapitel wird ein Trackingverfahren vorgestellt, das auf diesem Implicit Shape Model basierten Detektionverfahren aufbaut und das Implicit Shape Model zur Kombination von Detektion und Tracking erweitert.

[1]Das Modellwissen, welches durch das Codebuch integriert wird, ist auf die Objektklasse und nicht auf eine bestimmte Objektinstanz bezogen. Es ist demnach nicht gleichzusetzen mit modellbasiertem Tracking.

4.1.1 Datenassoziation durch Projektion von Objekthypothesen

Der hier vorgestellte Ansatz zur Objektverfolgung basiert darauf, dass im System zum Zeitpunkt T bekannte Personenhypothesen der Hypothesenmenge Γ für den nächsten Zeitpunkt (für das nächste Bild) projiziert werden. Der Projektionsschritt besteht im wesentlichen aus einem Prädiktions- und einem Vergleichsschritt. Die Datenassoziation findet also anders als bei tracking-by-detection Methoden nicht nach der bottom-up Objektdetektion, sondern davor statt. Hierzu wird für die Merkmalsmenge Π_γ der Hypothese γ eine *Zustandsprädiktion* für den aktuellen Zeitpunkt T durchgeführt. Der Zustand umfasst dabei die Bildposition[2] (x, y) und die Skalierung des Merkmals. Die Prädiktion findet auf Basis der Zustandshistorie der gesamten Merkmalsmenge statt. Details zu dem eingesetzten Dynamikmodell werden in Abschnitt 4.3 erläutert.

Unter Nutzung der prädizierten Merkmalszustände wird die Merkmalsmenge Π_γ mit den aus dem aktuellen Eingabebild extrahierten Merkmalen Π_{img} verglichen. Ziel dieses Vergleichs ist die Durchführung der Datenassoziation zwischen dem vorherigen und aktuellen Zeitpunkt. Für die Bildposition eines Merkmals π_γ wird dabei ein Aufenthaltsradius mit der prädizierten Position als Zentrum festgelegt. Dieser Aufenthaltsbereich legt die Menge von Bildmerkmalen fest, die mit dem Merkmal $\pi_\gamma \in \Pi_\gamma$ assoziiert werden dürfen. Eine Assoziation kann also nur mit Bildmerkmalen, deren Zentrum sich innerhalb des Aufenthaltsbereichs τ_S befindet, stattfinden:

$$\beta_S(\pi_\gamma, \pi_{img}) = \begin{cases} 1, & \text{wenn } \sqrt{((\pi_{\gamma,x} - \pi_{img,x})^2 + (\pi_{\gamma,y} - \pi_{img,y})^2)} < \tau_S \\ 0, & \text{sonst} \end{cases}. \tag{4.1}$$

Gleiches gilt für die Skalierungstufe des Merkmals π_γ. Hier wird der Schwellwert allerdings wesentlich niedriger gewählt, da sich die Skalierung der Person und damit die Skalierungsstufe der Merkmale nur sehr langsam ändern sollte. Für ein Merkmal π_γ ergibt sich somit eine Menge $\Pi_{img,\pi_\gamma} \subseteq \Pi_{img}$ von Bildmerkmalen, mit denen eine Assoziation möglich ist. Innerhalb der jeweiligen Zuweisungsbereiche findet keine Unterscheidung von Aufenthaltswahrscheinlichkeiten statt. D.h., liegt die Bildposition des Merkmals im zulässigen Bereich wird dieses mit 1 beim Merkmalsvergleich gewichtet. Die Bildposition wird über die Verwendung als Ausschlusskriterium hinaus im Merkmalsvergleich somit nicht als zusätzliches Merkmal eingebracht. Dies ist darin begründet, dass sich die Bewegungsrichtung gerade im Fall von artikulierten Objekten oftmals abrupt ändert und somit die Positionsprädiktion zwar den Aufenthaltsbereich einschränken kann, aber kein verlässliches Merkmal zur Datenassoziation ist. Der Vergleich der Merkmalsdeskriptoren $\vec{\mathbf{x}}$ und ζ der Merkmale π_γ und π_{img} wird somit auf Basis des Deskriptorabstandes (SSD) nach Formel 3.16 durchgeführt. Die Assoziation erfolgt auf Basis dieses Vergleichswertes, wobei der tatsächliche Abstand $\delta(\vec{\mathbf{x}}, \vec{\zeta})$ im Deskriptorraum die Ähnlichkeit und damit die Assoziationsstärke $\beta_D(\pi_\gamma, \pi_{img})$ angibt:

$$\beta_D(\pi_\gamma, \pi_{img}) = \max\left(\frac{\delta(\vec{\mathbf{x}}, \vec{\zeta}) - \tau_D^{MAX}}{-\tau_D^{MAX}}, 0\right). \tag{4.2}$$

Es werden somit nur Merkmalszuweisungen zugelassen, die innerhalb eines prädeterminierten Radius τ_D^{MAX} im Deskriptorraum liegen.

Diese Art der Datenassoziation ist schlüssig für den Fall eines einzelnen Merkmals. Tatsächlich besteht eine Hypothese aber aus einer Merkmalsmenge, deren Datenassoziation nicht unabhängig voneinander erfolgen kann, da es zwangsläufig Überschneidungen in den Mengen Π_{img,π_γ} der zulässigen Merkmale zwischen einzelnen Hypothesenmerkmalen π_γ gibt. Da ein Bildmerkmal, wie in Kapitel 3 diskutiert, nur exklusiv einem bestimmten Objekt zugeordnet werden kann, muss die Assoziation von Bildmerkmalen mit Hypothesenmerkmalen an dieser Stelle auch exklusiv durchgeführt werden. Insbesondere konkurrieren nicht nur die Merkmale Π_γ einer einzelnen Hypothese um aktuelle Bildmerkmale, sondern

[2] Diese ist durch die Position des Schlüsselpunkts gegeben.

natürlich auch die unterschiedlichen Hypothesen, was bedeutet, dass auch die Assoziation zwischen Hypothesen nicht unabhängig ablaufen kann. Ziel ist es, die beste Assoziation von Hypothesen- mit Bildmerkmalen zu finden die Exklusivität der Assoziation gewährleistet. Die beste Zuweisung von K Bildmerkmalen zu N Hypothesenmerkmalen wird dabei als die Zuweisung

$$k \rightarrow \nu(k), 1 \leq k \leq K, 1 \leq \nu(k) \leq N \tag{4.3}$$

$$k \neq n \Rightarrow \nu(k) \neq \nu(n) \tag{4.4}$$

definiert, die die Summe der Ähnlichkeiten

$$\beta_D^{ges} = \sum_{k=0}^{K} \beta_D(k, \nu(k)) \tag{4.5}$$

maximiert. Dieses Zuweisungsproblem kann durch den ungarischen Algorithmus [97, 130] effizient gelöst werden.

4.1.2 Fusion datengetriebener Objektdetektion mit Erwartungen

Der Datenassoziationsschritt liefert für alle $\pi_\gamma \in \Pi_\Gamma$ ein $\pi_{img} \in \Pi_{img}$ oder 0, wenn kein Merkmalspaar zustande kommt. Ebenso hat jedes $\pi_{img} \in \Pi_{img}$ ein $\pi_\gamma \in \Pi_\Gamma$ zugeordnet (die Abbildungen sind durch Hinzunahme der 0 in die Zielmenge injektiv). Aus dieser Form der Datenassoziation ergeben sich nun 3 unterschiedliche Merkmalstypen:

1. Bildmerkmale π_{img} mit einem assoziierten Hypothesenmerkmal π_γ:
   $\Pi_{ass} \subseteq \Pi_{img}, \nu(\pi_{ass}) \neq 0, \forall \pi_{ass} \in \Pi_{ass}$.
2. Bildmerkmale π_{img} ohne ein assoziiertes Hypothesenmerkmal π_γ:
   $\Pi_{iss} = \Pi_{img} \backslash \Pi_{ass}$.
3. Hypothesenmerkmale π_γ ohne ein assoziiertes Bildmerkmal π_{img}:
   $\Pi_{hna} \subseteq \Pi_\Gamma, \nu(\pi_{img}) \neq \pi_{hna}, \forall \pi_{hna} \in \Pi_{hna}, \pi_{img} \in \Pi_{img}$

Aus diesen 3 Merkmalstypen ergibt sich nun der gemeinsame Merkmalsraum $\Pi_{tot} = \Pi_{ass} \cup \Pi_{iss} \cup \Pi_{hna}$, wobei die Merkmale durch ihre Typisierung eindeutig Klassen zugeordnet sind.

Ausgehend vom Merkmalsraum Π_{tot} wird die Objektdetektionsprozedur ausgeführt. Wie in Abbildung 4.1 zu sehen ist wird die Merkmalsmenge Π_{tot} wie bei der in Kapitel 3 beschriebenen Standard-Objektdetektion dazu zunächst mit dem Codebuch verglichen und der Voting-Raum durch die aktivierten Codebucheinträge aufgebaut. An dieser Stelle erfolgt nun die eigentliche Erweiterung des ISM zum Tracking, die sich auf *zwei wesentliche Veränderungen* stützt:

Zunächst wird die *Gewichtung* einer einzelnen Vote im Voting-Raum nun anhand des *Merkmalstyps* adaptiert. Formel 3.19 (Seite 29) aus der Personendetektion wird dazu um den Faktor P_{typ}^{π}, welcher den Merkmalstyp in das Vote Gewicht einbringt, erweitert:

$$V_{\vec{m}}^{w} = p(C_i|\pi) \cdot p(V_{\vec{k}}|C_i) \cdot P_{typ}^{\pi}. \tag{4.6}$$

Der Faktor P_{typ}^{π} einer Vote $V_{\vec{m}}$ wird dabei durch den Typ des Merkmals π, das diese Vote generiert hat, bestimmt. Somit wird das Gewicht einer Vote in der Voting-Prozedur nicht mehr nur durch Ähnlichkeiten von Merkmalsdeskriptoren bestimmt, sondern ist auch von der zeitlichen Stabilität eines Merkmals abhängig. Mit Stabilität ist hier nicht nur die Stabilität des dem Merkmal zugrundeliegenden

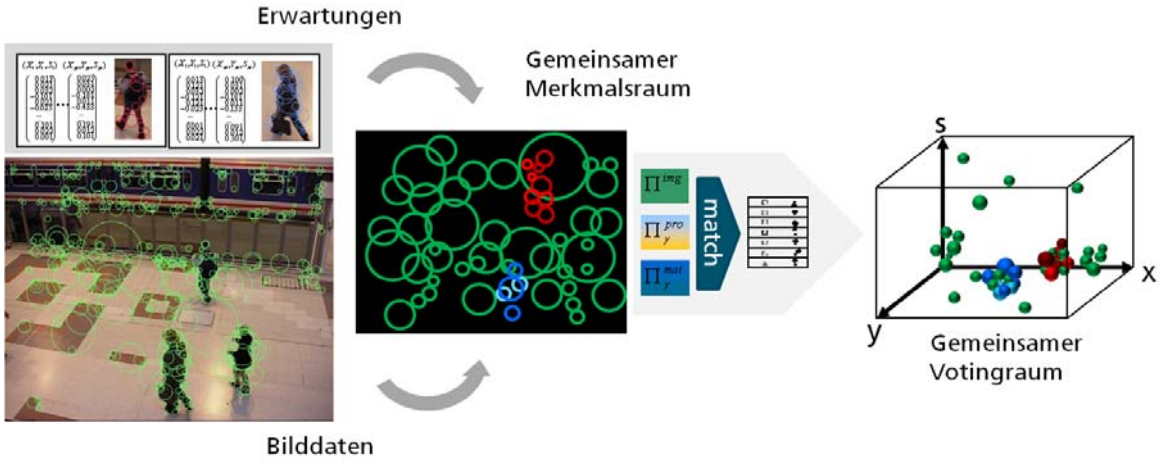


Abbildung 4.1: Schritte im Tracking bis zum Aufbau des Voting-Raums. Zunächst wird durch Vergleich der im System bekannten Hypothesen mit den aktuellen Bildmerkmalen ein Merkmalsraum mit drei unterschiedlichen Merkmalstypen generiert. Dieser gemeinsame Merkmalsraum bildet die Basis der Objektverfolgung, bei der zunächst der gemeinsame Voting-Raum durch Vergleich der Merkmalsmenge mit dem Codebuch erstellt wird. In diesem Voting-Raum findet nun die Objektverfolgung durch die erweiterte Objektdetektion statt.

Schlüsselpunkt gemeint, die allerdings Voraussetzung für die Stabilität auf höherer Ebene ist, sondern insbesondere auch die Stabilität und Aussagekraft bei der Objektklassifikation.

Dazu wird P^{π}_{typ} für Merkmalstyp 1 (Π_{ass}) auf einen Wert >1 gesetzt. Hierdurch werden Merkmale, bei denen die Erwartung aus dem letzten Zeitpunkt und die aktuelle Bildevidenz übereinstimmen, in ihrer Gewichtung und damit in ihrem Einfluss auf die Objektdetektion verstärkt. Somit wird zum einen die Objektdetektion stabilisiert, zum anderen ist es so auch möglich, Merkmalsmodelle eines Objektes aufzubauen (dies wird detailliert in Abschnitt 4.2 diskutiert). Für von Merkmalstyp 2 (Π_{iss}) generierte Votes wird P^{π}_{typ} auf 1 gesetzt. Diese von reinen Bildmerkmalen generierten Votes behalten also die gleiche Gewichtung bei. Von Merkmalstyp 3 (Π_{hna}) generierte Votes erhalten eine Gewichtung von <1. Somit gehen diese von Hypothesenmerkmalen ohne übereinstimmendes Bildmerkmal generierten Votes zwar in die Detektion ein, allerdings mit einer verminderten Gewichtung. Durch Einbringen dieser tatsächlich im aktuellen Bild nicht vorhandenen Merkmale in die Detektion wird eine starke Stabilisierung der Detektion erreicht. Im weiteren wird hierdurch eine automatische Zustandsschätzung bzgl. der Ausdehnung des Objekts und der Zuverlässigkeit der Objekthypothese erreicht. Dies wird in Abschnitt 4.2 näher diskutiert. Insbesondere wird durch Hinzunahme dieser Merkmale auch das Tracking über kurzzeitige Verdeckungen hinweg ermöglicht (siehe Abschnitt 4.5).

Als zweite wichtige Veränderung werden Votes, welche aus Merkmalen der Typen 1 und 3 enstanden sind mit *Hypothesenidentifikatoren (ID)* versehen. Diese ergeben sich bei Typ 1 aus der bei der Datenassoziation bestimmten Zuordnung von Bildmerkmalen zu Objekthypothesen, bei Typ 3 direkt aus der ID der generierenden Objekthypothese. Dies ist in Abbildung 4.1 im *gemeinsamen Voting-Raum* durch unterschiedliche Farben gekennzeichnet. Durch diese Erweiterung wird die direkte Integration der Objektverfolgung in die Objektdetektion ermöglicht, da diese Votes hypothesenspezifisch sind und ausschließlich für eine bestimmte Hypothese abstimmen können. Votes, die aus Merkmalen des Typs 2 (reine Bildmerkmale) entstanden sind, werden mit einer allgemeinen ID versehen und können weiterhin für jede Objekthypothese abstimmen.

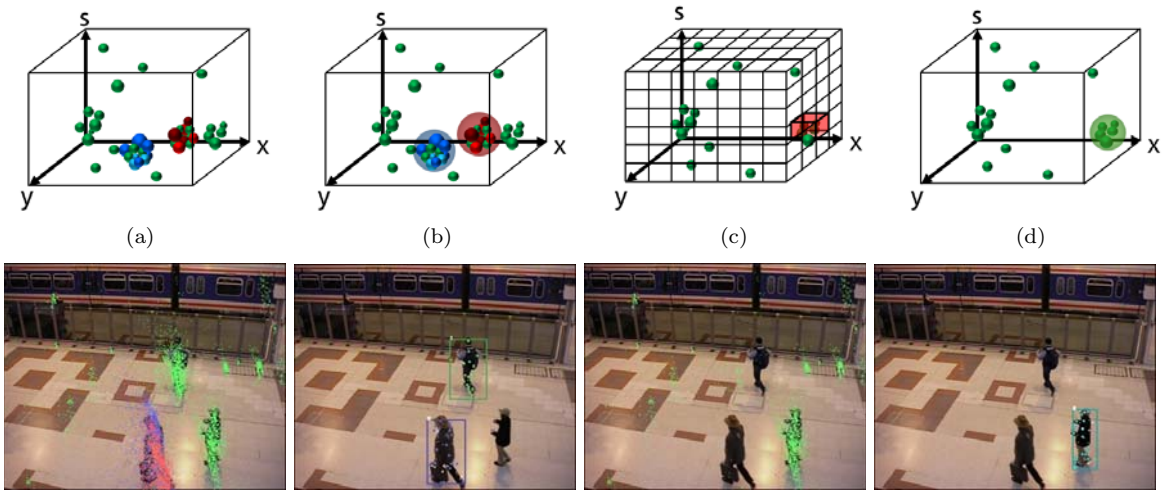


Abbildung 4.2: Objektverfolgung im Voting-Raum. (a) Gemeinsamer Voting-Raum. Hierbei sind die unterschiedlichen Vote-Typen farblich gekennzeichnet. Typ 1: Blau, Rot; Typ 2: Grün; Typ 3: Hellblau, Hellrot. Blau und rot gibt dabei die Zugehörigkeit zu unterschiedlichen Objekthypothesen an. (b) Mean-Shift-Suche zum Tracking bekannter Objekte. Hier ist keine initiale Maximasuche nötig, da die letzte Position der Objekte bekannt ist und als Startpunkt der Mean-Shift Maximasuche verwendet werden kann. (c) Maximasuche im diskretisierten Hough-Raum zur Detektion neuer Objekte. (d) Mean-Shift-Suche zur genauen Bestimmung der Maxima.

4.1.3 Tracking im Hough-Voting-Raum

Die Objektverfolgung wird nun durch Nutzung der Identitätsannotation der Votes direkt bei der Objektdetektion im Voting-Raum durchgeführt. Dazu wird die Objektdetektion um die Nutzung der ID-Annotation erweitert, in ihrer prinzipiellen Funktionsweise aber beibehalten. Der so erweiterte Voting-Raum ist in Abbildung 4.2 (a) dargestellt. Hier stellen die verschiedenen Farben die Zugehörigkeit zu den Objekthypothesen dar. Die Größe einer Vote gibt wie bereits bei der Objektdetektion ihre Stärke an, wobei diese hier durch den zusätzlichen Faktor P^{π}_{typ} bestimmt wird.

Wie in Abbildung 4.2 dargestellt kann das Tracking durch eine direkte, hypothesenspezifische Mean-Shift-Suche im Voting-Raum durchgeführt werden. Dazu wird, wie in Abbildung 4.2 (b) dargestellt, für alle bekannten Hypothesen unabhängig eine Mean-Shift-Suche ausgehend von der prädizierten Objektzentrumsposition gestartet. Da diese erwartete Position bekannt ist, ist keine Initialisierung durch Maxima im diskretisierten Hough Raum notwendig, um Anzahl und initiale Positionen der Mean-Shift-Suchen zu bestimmen. Jede hypothesenspezifische Mean-Shift-Suche bezieht ausschließlich Votes, die für die eigene Hypothese stimmen, sowie allgemeine Votes, die für beliebige Hypothesen stimmen, ein.

Hierdurch wird die Objektidentität automatisch gewahrt, da nach einer bestimmten Hypothese gesucht wird. Da bei der Suche allerdings auch allgemeine Votes des Typs 2 einbezogen werden, wird auch neue Information integriert und somit eine automatische Adaption des Identitätsmodells vollzogen. Ebenso werden Merkmale des Typs 3, die nicht mehr für das nun aktuelle Maximum abstimmen, automatisch aus der Hypothese entfernt. Somit erfolgt ein automatischer Ausschluss nicht mehr aktueller Informationen. Das alte Identitätsmodell einer Hypothese kann nun direkt durch die neue Information aus der aktuellen Maximasuche ersetzt werden, da hier bereits die Kombination der Information aus der Tracking-Historie mit der aktuellen Information stattgefunden hat. Es sind somit keine

nachgeschalteten Modellaktualisierungsheuristiken mehr notwendig. Ebenso hat auch die Zuverlässigkeitsbestimmung der Hypothese schon während des Votings stattgefunden. Die neue Zuverlässigkeit ergibt sich nach Formel 3.20 direkt aus der Stärke des Maximums.

Zur Detektion neuer, d.h. bisher unbekannter Objekte wird die Standard-Objektdetektion wie in Kapitel 3 beschrieben durchgeführt. Wie in Abbildung 4.2 (c) zu sehen ist, wird diese im reduzierten Voting-Raum durchgeführt. Der reduzierte Voting-Raum entsteht durch Entfernen der Votes von Merkmalen, die bereits für Objekthypothesen gestimmt haben. Diese sind schon einer Hypothese zugeordnet und können somit nicht für weitere Hypothesen stimmen. Ebenso werden alle Votes entfernt, die nur für bestimmte bekannte Objekthypothesen stimmen. Wie bei der Standard-Objektdetektion werden die Maximapositionen durch eine Mean-Shift-Suche genauer bestimmt (siehe Abbildung 4.2 (d)). Alle Maxima, die stark genug sind, werden als neue Objekthypothesen in das Trackingsystem übernommen. Die Detektion, genauer die an der Detektion beteiligten Bildmerkmale, bilden dabei das initiale Identitätsmodell der Hypothese. Ausgehend von diesem Modell wird im nächsten Schritt dann wiederum das Tracking für diese nun bekannte Hypothese durchgeführt.

4.2 Aufbau von Identitätsmodellen

Die im letzten Abschnitt beschriebene Strategie zur Objektverfolgung im ISM beruht, ähnlich wie ein Markov-Prozess, darauf, dass der aktuelle Zustand alle Informationen enthält, die bis zu diesem Zeitpunkt angefallen sind. Beim hier vorgestellten Tracking bedeutet dies, dass die Informationen in einer Detektion, welche die Fortsetzung einer bekannten Hypothese darstellt, die obsolete Information im Identitätsmodell der Objekthypothese ersetzen kann. Hinsichtlich der Stabilitätsmodellierung von Merkmalen durch den Typfaktor P^{π}_{typ} bezieht sich die Modellierung lediglich auf den letzten Zeitpunkt und nicht auf die gesamte Merkmalshistorie. Um die komplette Tracking-Historie eines Merkmals einzubeziehen wird der konstante Faktor P^{π}_{typ} in Gleichung 4.6 durch eine zeitabhängige, merkmalsspezifische Funktion $P^{\pi,t}_{typ}$ ersetzt. Für ein Merkmal $\pi_\gamma \in \Pi_\Gamma$ zum Zeitpunkt t ist der Typfaktor:

$$P^{\pi,t}_{typ} = P^{\pi,t-1}_{typ} \cdot \alpha_{typ}. \tag{4.7}$$

Hier ist $P^{\pi,t-1}_{typ}$ der Typfaktor des letzten Zeitpunkts und α_{typ} die typspezifische Adaptionsrate (vorher P^{π}_{typ}). Der Funktionswert wird in jedem Datenassoziationsschritt auf Basis des Merkmalstyps durch Multiplikation mit dem typspezifischen Adaptionsfaktor aktualisiert. Hierdurch wird die ganze Merkmalshistorie einbezogen und der aktuelle Wert abhängig vom Vorhandensein neuer Evidenz in den Bilddaten adaptiert. Bei der initialen Detektion einer Objekthypothese sind alle beteiligten Merkmale vom Typ 2 und starten somit mit einem Typfaktor von 1.0. Bei folgenden Detektionen (dem Tracking) der Objekthypothese hängt der Typfaktor davon ab, ob in den Bilddaten neue Evidenz für das entsprechende Merkmal vorhanden ist oder nicht. Sofern ein Merkmal zu einem Zeitpunkt mit einem Bildmerkmal assoziiert werden kann erhöht sich $P^{\pi,t}_{typ}$ durch Multiplikation mit $\alpha_{typ} > 1.0$. Merkmale, die wiederholt in den Bilddaten bestätigt werden steigen in ihrer Gewichtung, und damit im Einfluss auf die Detektion somit stetig an. Dies trägt sehr stark zur Stabilität der gesamten Hypothese bei, da durch wiederholt bestätigte Merkmale bereits eine a-priori Wahrscheinlichkeit für eine Hypothese, und insbesondere auch für eine Zentrumsposition vorhanden ist. Falls Merkmale nicht durch die Bilddaten bestätigt werden können, wird der Typfaktor durch Multiplikation mit $\alpha_{typ} < 1.0$ abgesenkt. Diese im Bild nicht vorhandenen Merkmale werden zwar weiterhin in die Detektion eingebracht, werden aber in ihrem Einfluss innerhalb der Detektion abgesenkt. Sollten Merkmale wiederholt nicht bestätigt werden, so sinkt $P^{\pi,t}_{typ}$ stetig ab und die Merkmale werden bei zu niedrigem Wert aus dem Modell entfernt. Dies ist auch in Abbildung 4.3 zu sehen.

Die Erweiterung durch den merkmalsspezifischen, zeitadaptiven Typfaktor hat drei Hauptaspekte hinsichtlich des Trackings:

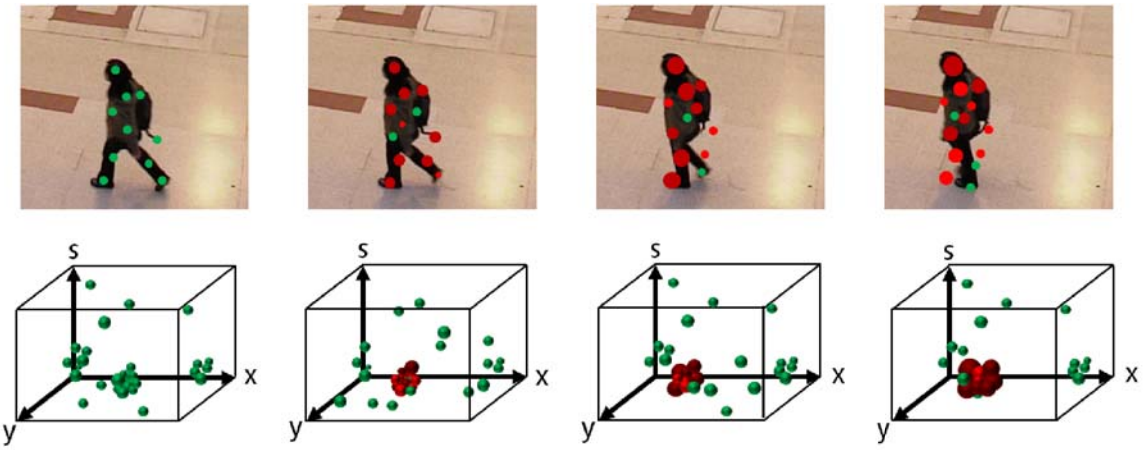


Abbildung 4.3: Aufbau von Identitätsmodellen während des Trackings. Ausgehend von einem Initialwert von 1 werden Merkmale, die durch die Bilddaten bestätigt werden, in ihrer Gewichtung angehoben, bzw. abgesenkt wenn sie nicht bestätigt werden. Der Merkmals-, bzw. Vote-Typ ist durch unterschiedliche Farben: Typ 1: Rot, Typ 2: Grün, Typ 3: Hellrot dargestellt. Das aktuelle Gewicht der Merkmale bzw. Votes ist durch die Größe der Kreise dargestellt.

(i) Wiederholt gesehene Merkmale haben einen größeren Einfluss auf das Hypothesenmaximum. Hierdurch wird die *Stabilität der Detektion*, sowohl was die Zeitstabilität als auch was die Ortsstabilität des Objektzentrums im Referenzsystem der Person angeht, erhöht.

(ii) Die *Zuverlässigkeit einer Hypothese* kann durch Summenbildung über alle beitragenden Votes direkt aus den für diese Hypothese abstimmenden Merkmalen geschlossen werden. Wurde eine Person über einen längeren Zeitraum korrekt verfolgt, so steigt die Zuverlässigkeit dieser Hypothese durch wiederholt bestätigte Merkmale an. Können die Merkmale einer Hypothese über einen längeren Zeitraum nicht im Bild bestätigt werden, so fällt die Zuverlässigkeit dieser Hypothese automatisch ab. Insbesondere ist die Dauer, für die eine Hypothese ohne Bestätigung durch Bildmerkmale aufrechterhalten werden kann damit auch abhängig von der Zeit, in der sie vorher korrekt verfolgt werden konnte. Dies ist schlüssig, da so evtl. auftretende Falschalarme schon nach kurzer Zeit wieder aus der Menge der Hypothesen entfernt, Objekte die stabil verfolgt werden konnten aber über längere Verdeckungen hinweg verfolgt werden können. Das Entfernen von Hypothesen aus der Hypothesenmenge erfolgt automatisch dadurch, dass die nicht bestätigten Bildmerkmale nach Absinken der Gewichtung nicht mehr ausreichen, um eine Detektion genügend großer Stärke zu generieren. Wichtig ist, dass für die gesamte Funktionalität keinerlei Heuristiken notwendig sind, sondern dies automatisch durch den Tracking-Ansatz gehandhabt wird.

(iii) Während des Trackings werden automatisch *Kurzzeit-Identitätsmodelle* der verfolgten Personen aufgebaut, wobei der Gewichtungsfaktor die aktuelle Relevanz eines bestimmten Merkmals im Modell bestimmt. Insbesondere erfolgt durch die Art der Objektverfolgung im Hough-Voting-Raum auch eine automatische Anpassung an Erscheinungsveränderungen der Person. Dies wird dadurch erreicht, dass die Maximasuche zur Hypothesenfortsetzung sowohl neue Bildmerkmale als auch bereits gesehene Hypothesenmerkmale integriert. Die automatische Adaption an Ansichtsveränderungen ist von großer Relevanz, da Erscheinungsveränderungen nicht nur durch Verdeckungen der Person, sondern auch durch Bewegung der Person in der Szene, sowie durch umgebungsbedingte Veränderungen wie z.B. Beleuchtungsveränderungen zustande kommen können. Die beim Tracking aufgebauten Kurzzeit-Identitätmodelle dienen außerdem als Basis für die bei der Personenwiedererkennung verwendeten Langzeit-Identitätsmodelle (siehe Kapitel 5).

4.3 Modellierung der Bewegungsdynamik

4.3.1 Kalman-Filter

Ein wesentliches Merkmal des hier vorgestellten Trackingverfahrens besteht darin, dass es nicht nur spezifisch für Personen, sondern für beliebige Objektklassen nutzbar ist. Aus diesem Grund soll auch bei der Auswahl des Dynamikmodells keine objektspezifische Modellierung stattfinden wie dies z.B. bei spezialisierten Ansätzen zur Personenverfolgung [5] häufig der Fall ist. Zur Dynamikmodellierung wird daher ein Kalman-Filter [93] mit Systemzustand $(x, y, v_x, v_y, a_x, a_y)$ eingesetzt. Dieses stellt eine einfache Form der Dynamikmodellierung dar und kann prinzipiell durch komplexere Modelle, wie nicht-lineare Kalman-Filter oder Partikelfilter, die ebenfalls unabhängig von Objektspezifika einsetzbar sind, ersetzt werden. Tatsächlich sind diese komplexeren Bewegungsmodelle bei dem hier vorgestellten Tracking-Ansatz nicht notwendig, da durch die in jedem Zeitschritt durchgeführte Korrespondenzbestimmung auf Merkmalsebene eine automatische Anpassung an das Bewegungsverhalten erfolgt.

Die im Datenassoziationsschritt bestimmten Merkmalskorrespondenzen bilden hierbei die Basis der Objektdynamikmodellierung, wobei nicht die Dynamik jedes einzelnen Merkmals modelliert wird, sondern ein Dynamikmodell, das als Observation eine Teilmenge der Hypothesenmerkmalsmenge verwendet, pro Objekthypothese genutzt wird. Hierzu werden zu einem Zeitpunkt T alle $\pi_{ass}^{\gamma,T} \in \Pi_\gamma$, also alle Hypothesenmerkmale die im aktuellen Schritt mit einem Bildmerkmal aktualisiert wurden, als Observation für das Kalman-Filter verwendet. Hierbei wird nicht die absolute Position der Bildmerkmale als Observation verwendet, sondern der Bildpositionsversatz des Merkmals zum Zeitpunkt $T-1$ (der nur bei Merkmalen des Typs 1 vorhanden ist). Die Messkovarianz wird hierbei durch den Anteil der Typ 1 Merkmale an der Gesamtmerkmalsmenge modelliert bzw. adaptiert. Hierbei werden die beim Merkmalsvergleich 4.1.1 bestimmten Ähnlichkeiten in Merkmalspaaren (die Ähnlichkeit zwischen $\pi_i^{\gamma,T-1}$ und assoziiertem $\pi^{img,T} = \nu(\pi_i^{\gamma,T-1})$) zur Gewichtung genutzt:

$$\rho_{(x/y/s)}^{\gamma} = \frac{\sum_{i=0}^{I} \beta_D(\pi_i^{\gamma,T-1}, \nu(\pi_i^{\gamma,T-1})) \cdot (\pi_{i,(x/y/s)}^{\gamma,T-1} - \nu(\pi_i^{\gamma,T-1})_{(x/y/s)})}{\sum_{i=0}^{I} \beta_D(\pi_i^{\gamma,T-1}, \nu(\pi_i^{\gamma,T-1})}. \tag{4.8}$$

Zur Prädiktion der Position der einzelnen Merkmale für den nächsten Zeitpunkt wird die gleiche Kalman-Filter Prädiktion für alle Merkmale genutzt. Die Prädiktion für das Objektzentrum, die für die Positionierung des Mean-Shift Kernels im Tracking benötigt wird, wird ebenfalls auf Basis der Dynamikmodellierung der Merkmalsmenge durchgeführt. Eine direkte Modellierung der Dynamik des Objektzentrums hat sich in der Praxis als zu instabil erwiesen. Gleiches gilt für eine Dynamikmodellierung auf Ebene der einzelnen Bildmerkmale. Hierzu müssten alle einzelnen Merkmale für sich von Beginn an in mehreren aufeinanderfolgenden Bildern wiedergefunden werden, um die Dynamik korrekt aufzeichnen zu können. Dies ist in der Praxis nicht gegeben, da einzelne Merkmale häufig nicht stabil genug im Bild wiedergefunden werden. Dies ist aufgrund der prinzipiellen Eigenschaften von schlüsselpunktbasierten Verfahren der Fall, aber auch weil bei einer bewegten Person fortlaufend Ansichtsveränderungen stattfinden. Da die Merkmalsdeskriptoren nur zu einem gewissen Maß ansichtsinvariant sind, ist es nicht zu vermeiden (und sogar erwünscht um beim Aufbau von Merkmalsmodellen nicht zu sehr zu generalisieren), dass die einzelnen Merkmalstracks Ausfälle haben, bzw. jeweils nur von kurzer Dauer sind.

4.3.2 Bewegungskompensation durch Merkmalsvergleich

Ein Nachteil des hier vorgestellten Tracking-Ansatzes im Vergleich zu reinen tracking-by-detection Ansätzen, die auf dedizierten Objektdetektoren aufbauen, besteht darin, dass das hier vorgestellte Verfahren im Datenassoziationsschritt die absolute Position, bzw. die Prädiktion für die absolute Position der Merkmale zur Datenassoziation nutzt. Für ein Merkmal ist dabei ein Bildbereich vorgegeben, der die Menge der zuweisbaren Merkmale bestimmt. Dieser Bildbereich um die prädizierte

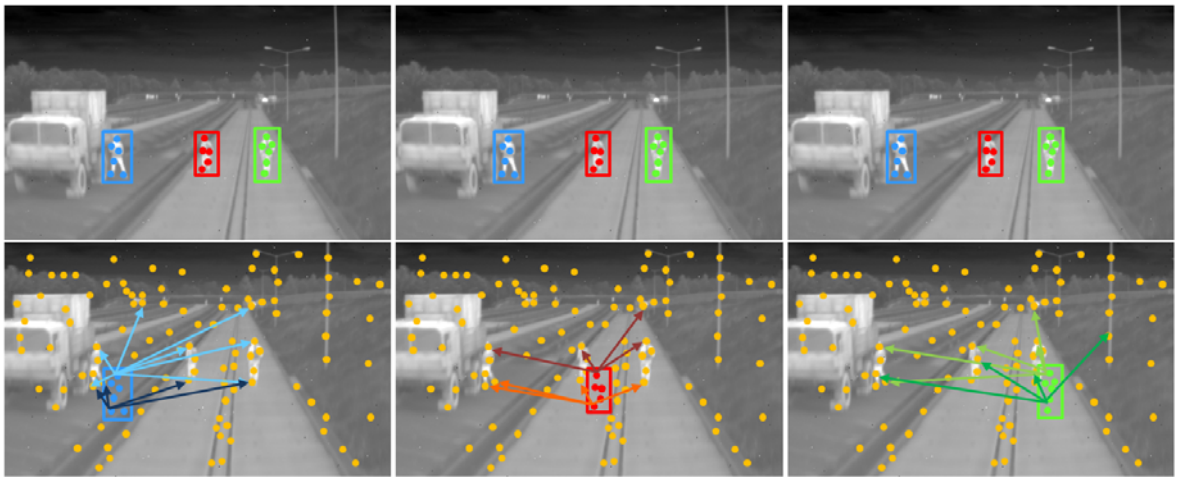
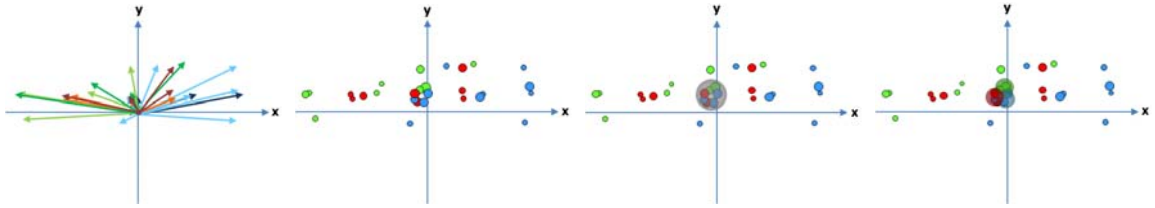


Abbildung 4.4: Berechnung der Versatzvektoren zwischen aufeinanderfolgenden Bildern einer Bildfolge. Die obere Reihe zeigt Trackingergebnisse des Zeitpunkts T. Die untere Reihe visualisiert die Berechnung der Versatzvektoren für das nächste Bild: Jedes Merkmal des Kurzzeit-Modells einer Person wird mit allen aktuellen Bildmerkmalen verglichen. Abhängig von der Ähnlichkeit werden Bildmerkmale aktiviert und der jeweilige Versatz zwischen Hypothesen- und Bildmerkmal gewichtet mit der Ähnlichkeit aufgezeichnet. Anhand dieser Versätze kann die Gesamtbewegung zwischen den Bildern kompensiert und somit die Bewegung der einzelnen Personen berechnet werden.

Position eines Merkmals gibt also die maximale Positionsunsicherheit vor, die zwischen zwei aufeinanderfolgenden Bildern einer Bildsequenzen vorhanden sein darf. Sind nicht nur die Objekte in der Szene, sondern auch der aufzeichnende Sensor bewegt, so kann die Bildbereichseinschränkung bei der Merkmalsassoziation dazu führen, dass korrekte Merkmalspaare aufgrund zu großer räumlicher Distanz nicht gebildet werden können. Da in der Praxis die Kamerabewegung aber inhärent mit in die Bewegungsschätzung des Kalman-Filters eingeht, wird gleichmäßige Kamerabewegung automatisch kompensiert. Unter der Annahme, dass die Positionsprädiktion für ein Merkmal bei Nichtbetrachtung der Sensorbewegung korrekt ist, gibt der bei der Datenassoziation gewählte Radius τ_S die maximale Sensorbewegung in Pixeln des Bildversatzes an, bei der eine korrekte Datenassoziation möglich ist. Probleme bei der Datenassoziation treten also lediglich bei stark unruhiger und nicht gleichmäßiger Kamerabewegung, bei der der Pixelversatz zwischen zwei Bildern durch starke Sensorbewegung den Radius τ_S überschreitet (siehe Auswertung in Abschnitt 4.5.2.2 für experimentelle Auswertung bei gleichmäßiger Kamerabewegung), auf.

Um eine Objektverfolgung auch in diesen Fällen von starker Sensorbewegung zu ermöglichen wird in diesem Abschnitt ein Modell zur merkmalsbasierten Bewegungskompensation vorgestellt, welches die Einschränkung durch τ_S bei der Datenassoziation entfernt. Dieses Verfahren macht keinerlei Annahmen über die absolute Position der Personen im Bild und erreicht dadurch, dass ein Tracking auch bei beliebigem Pixelversatz zwischen zwei Bildern und somit bei beliebig starker Kamerabewegung möglich ist. Die prinzipielle Idee des Ansatzes ist es, die Bewegung zwischen zwei Frames durch ein Merkmalsmatching zu kompensieren. Dabei werden die Vorteile des hier vorgestellten Trackingansatzes gegenüber reinen tracking-by-detection Methoden (wie erhöhte Stabilität und Tracking durch Verdeckungen) beibehalten.

Abbildung 4.4 zeigt die prinzipielle Idee der Bewegungskompensation. Die aktuellen Merkmale Π_γ im Kurzzeit-Identitätsmodell einer Hypothese werden mit allen Bildmerkmalen Π^T_{img} des aktuellen Zeitpunkts T verglichen. Im Gegensatz zur eindeutigen Zuordnung von Hypothesen- zu Bildmerkmalen



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

Abbildung 4.5: Bewegungskompensation zum Tracking von Personen. (a) Tansfer der Merkmalsversätze in einen kontinuierlichen 2D Hough-Voting-Raum. (b) Votes mit durch die Kreisgröße dargestellten Gewichten. (c) Determinierung des globalen Maximums durch Mean-Shift. Das globale Maximum in diesem Voting-Raum definiert die globale Kamerabewegung in Pixeln des Bildversatzes. (d) Hypothesenspezifische Mean-Shift Maximasuche zur Bestimmung der Personenbewegung.

im Datenassoziationsschritt, werden hier für ein Hypothesenmerkmal π_γ alle Bildmerkmale aktiviert, deren Ähnlichkeit β_D über einem Schwellwert τ_{DM} liegt. Aktiviert meint in diesem Kontext, dass der Pixelversatz zwischen dem Merkmal π_γ und dem aktivierten Merkmal π_{img} aufgezeichnet wird. Ziel dieses Schrittes ist es, vielversprechende Aufenthaltspunkte der Hypothesenmerkmale und damit der Hypothesen im nächsten Bild zu finden. Da ohne die Einschränkung der Merkmalsposition die Deskriptorähnlichkeit das wesentliche Merkmal der Zuordnung ist, werden die aufgezeichneten Merkmalsversätze wie in Formel 4.8 mit der jeweiligen Merkmalsähnlichkeit gewichtet. An dieser Stelle wird im Gegensatz zur Datenassoziation noch keine eindeutige Zuordnung zwischen Merkmalen getroffen, da diese ohne eine räumliche Einschränkung fehlerhaft sein könnte. Dies liegt an der Tatsache, dass beim Multi-Hypothesen Tracking von Objekten der gleichen Klasse häufig große visuelle Ähnlichkeiten zwischen den Objekten der Klasse auftreten. Z.B. ist beim Tracking von Personen im infraroten Spektralbereich kaum eine direkte visuelle Unterscheidung der Personen möglich. Dies macht das Tracking im infraroten im Vergleich mit dem Tracking im visuellen Spektralbereich schwieriger. Insbesondere kann nicht davon ausgegangen werden, dass eine eindeutige Diskriminierung auf Ebene eines einzelnen Bildmerkmals möglich ist. Durch den hier gewählten Ansatz, dass zunächst alle Merkmale mit ausreichender Ähnlichkeit für die Assoziation in Betracht gezogen werden, kann davon ausgegangen werden, dass das korrekte Merkmalspaar in der Menge enthalten ist.

Aus der Durchführung dieses Verfahrens für jede der Objekthypothesen ergibt sich pro Hypothese eine Menge $\kappa^{\gamma,T}$ von gewichteten Verschiebungsvektoren. Für den aktuellen Zeitpunkt T ergibt sich somit die Gesamtmenge von Verschiebungsvektoren $\kappa^{\Gamma,T} = \kappa^{1,T} \cup \kappa^{2,T} \cup \kappa^{N,T}$. Wie in Abbildung 4.5 dargestellt ist, werden diese Verschiebungsvektoren (4.5 (a)), in einen zweidimensionalen Hough-Voting-Raum (4.5 (b)), bestehend aus den zwei Bilddimensionen, transferiert. Hier stimmt jeder Verschiebungsvektor gewichtet für einen möglichen Versatz zwischen den zwei Bildern. Unter der Annahme, dass die relative räumliche Position der Objekte zwischen den beiden Bildern erhalten bleibt, und die Ähnlichkeit zwischen den korrekten Objekten auf Merkmalsebene zumindest vergleichbar groß wie die Ähnlichkeit zu den falschen Objekten oder Hintergrundstrukturen ist, kann angenommen werden, dass das Maximum in diesem Raum, das aus den Stimmen aller Hypothesen gebildet wird, an der Position des durch die Kamerabewegung verursachten Pixelversatzes liegt. Die Maximasuche (4.5 (c)) wird mit Mean-Shift durchgeführt. Dies hat den Vorteil, dass durch die Wahl eines geeignet großen Mean-Shift-Kernels gezielt Toleranzen bei der Maximalokalisierung zugelassen werden können. Dies ist notwendig, da die Eigenbewegung der Personen dazu führt, dass das Maximum nicht exakt an der durch die Kamerabewegung definierten Position liegt, sondern die Votes eher um diesen Punkt verteilt sind. Ausgehend vom globalen Maximum, welches der Kamerabewegung (in Pixelkoordinaten) entspricht, erfolgt die Bestimmung der hypothesenspezifischen Maxima (4.5 (d)). Dazu wird für jede Hypothese eine Mean-Shift-Suche gestartet, die allerdings nur Votes für diese bestimmte Hypothese,

also Versätze, die durch Merkmale dieser Hypothese generiert wurden, einbezieht. Die Suche startet dabei für alle Hypothesen ausgehend vom globalen Maximum. Ausgehend von dem Punkt der Kamerabewegung soll also hierbei für jede Hypothese nun noch die Eigenbewegung bestimmt werden. Die Wahl des globalen Maximums als Startpunkt ist notwendig, da es sonst zu Vertauschungen zwischen Objekten kommen könnte. Dies wäre bei einer direkten hypothesenspezifischen Maximasuche dann der Fall, wenn die Ähnlichkeit zwischen unterschiedlichen Objekten (der gleichen Klassen) zwischen den beiden Zeitpunkten höher wäre als die Ähnlichkeit des gleichen Objekts. Da dies insbesondere im infraroten Spektralbereich bei Objekten der gleichen Klasse nicht ungewöhnlich ist, muss dieser Fall einkalkuliert werden. Dadurch, dass bei der Suche nach dem globalen Maximum die Votes aller Hypothesen einbezogen werden, wird aber hier nicht nur die visuelle Ähnlichkeit einzelner Merkmale genutzt. Vielmehr wird hier Konsistenz auf Ebene der Merkmalsversätze aller Merkmale gefordert. Es werden zudem die räumlichen Abhängigkeiten zwischen den Bildmerkmalen inhärent eingebracht, ohne dass eine räumliche Einschränkung erfolgt. Die in diesem Verfahren inhärent ausgenutzte Forderung, dass sich die relativen Positionen der Objekte zueinander auch bei starker Kamerabewegung nur marginal durch die Eigenbewegung der Objekte ändern und somit als räumliche Konsistenzeigenschaft auch bei starker Kamerabewegung genutzt werden können, trägt somit über die reine visuelle Ähnlichkeit der Merkmalspaare zu einer konsistenten Merkmalszuordnung bei. Da die Merkmalsähnlichkeit aber direkt als Gewichtung in das Zuordnungsverfahren eingeht, wird hier also die räumliche Konsistenzforderung mit der visuellen Ähnlichkeitsforderung kombiniert um so eine korrekte Datenassoziation zu gewährleisten.

Bei Abschluss der Mean-Shift-Suche liefern die hypothesenspezifischen Maxima die Bewegung der Objekte in Bildkoordinaten. Zudem können die Merkmalskombinationen, welche die Verschiebungsvektoren im endgültigen Maximum generiert haben, auch direkt zur Datenassoziation verwendet werden, indem die Ähnlichkeitswerte als Eingabe des ungarischen Algorithmus (siehe Abschnitt 4.1.1) genutzt werden.

Der hier vorgestellte Ansatz hat für den Anwendungsfall der Objektverfolgung zwei Hauptvorteile gegenüber der expliziten Bewegungskompensation z.B. durch eine Homographieschätzung. Zunächst passt sich dieser Ansatz dadurch, dass die Bewegungskompensation die Merkmalszuweisungen für die Datenassoziation liefert, nahtlos in die merkmalsbasierte Tracking-Strategie ein. Außerdem ist es in der Praxis schwierig, eine Homographie zu bestimmen, wenn zahlreiche bewegte Objekte in der Szene vorhanden sind. In der hier vorgestellten Strategie wird dies durch die Ungenauigkeiten, die in der Suche nach dem globalen Maximum explizit zugelassen werden, kompensiert.

Es bleibt zu sagen, dass der Nachteil dieser Strategie darin besteht, dass sich der Aufwand für den Merkmalsvergleich durch die größere Menge an notwendigen Vergleichen – da jedes Bildmerkmal mit allen Hypothesenmerkmalen verglichen werden muss – erhöht. Durch Nutzung eines geeigneten Abstandsmaßes und einer effizienten Vergleichsstrategie, wie z.B. KD-Trees, kann dieser allerdings wiederum reduziert werden. So kann mit dem hier zum Merkmalsvergleich verwendeten Abstandsmaß (SSD, siehe Formel 3.16) zusammen mit dem Ähnlichkeitsschwellwert τ_{DM} eine Reduzierung der erforderlichen Vergleichsoperationen stattfindet, da angenommen werden kann, dass die Merkmale in dem Identitätsmodell eines bestimmten Objekts auch nur hohe Ähnlichkeiten zu Objekten der gleichen Objektklasse und zu einigen wenigen anderen im Bild vorhandenen Strukturen aufweisen. Da die Abstandsberechnung der Deskriptoren bei Überschreiten des Schwellwerts abgebrochen werden kann, ist es in den meisten Fällen nicht notwendig, die Deskriptoren in voller Länge miteinander zu vergleichen. Tests hierzu haben ergeben, dass die Distanzberechnung von 87% der Merkmale vorzeitig abgebrochen werden kann. Im Schnitt findet der Abbruch dabei bei 53% des verwendeten Merkmalsdeskriptor statt, so dass insgesamt nur 53% der Vergleichsoperationen ausgeführt werden müssen.

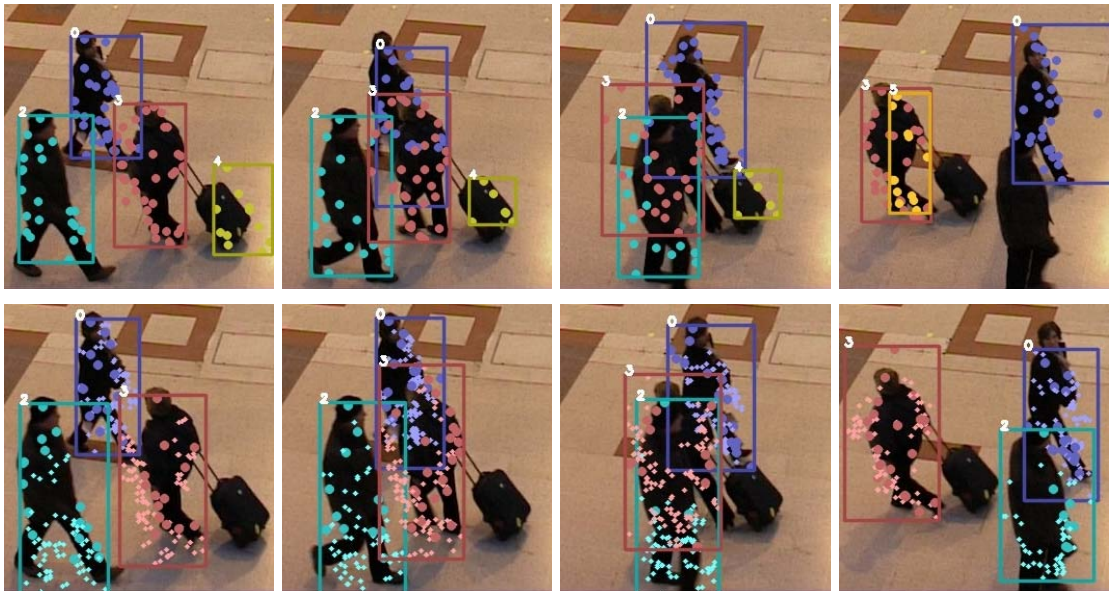


Abbildung 4.6: Vergleich tracking-by-detection (obere Reihe) mit dem in dieser Arbeit vorgestellten Tracking-Ansatz (untere Reihe)

4.4 Qualitative Bewertung

Zur Bewertung des hier vorgestellten Trackingverfahrens im Vergleich mit der Standard-Vorgehensweise bei einem tracking-by-detection Verfahren sei die Situation in Abbildung 4.6 betrachtet. Hier kreuzen sich die Laufwege von drei Personen, was zu einer Verdeckungssituation führt, bei der eine der Personen kurzzeitig total verdeckt wird. Im Einzelbild ist hier eine korrekte Detektion der verdeckten Person nicht möglich, so dass ein auf Einzelbilddetektionen aufbauendes Trackingverfahren die Person in dieser Situation nicht korrekt verfolgen kann, bzw. durch Heuristiken auf das Vorhandensein einer Person schließen muss.

Dies ist auch in der oberen Reihe von Abbildung 4.6 zu beobachten. Hier sind Ergebnisse eines Verfahrens, bei dem die Spurbildung auf Basis der Einzelbilddetektionen im Nachhinhein erfolgt, dargestellt. Zwar werden während der Verdeckung noch kurzzeitig zwei Hypothesen im Bild gefunden. Allerdings bauen beide Hypothesen hauptsächlich auf den Merkmalen der im Vordergrund sichtbaren Person auf, so dass in der Folge für eine der Hypothesen nicht mehr genügend Merkmale vorhanden sind, um weiterhin aufrecht erhalten zu werden. Wie in der unteren Reihe von Abbildung 4.6 zu sehen ist, ist es durch das hier vorgestellte Verfahren möglich, die Person auch über solche Verdeckungen hinweg korrekt zu verfolgen. Insbesondere kann der aktuelle Aufenthaltsort im Bild auch während einer Verdeckung korrekt eingeschätzt werden. Da dies durch die Projektion der Erwartungen und die Nutzung der drei Merkmalstypen direkt in das Trackingverfahren integriert ist, müssen keine Heuristiken, die solche speziellen Situationen behandeln, eingesetzt werden. Insbesondere kann eine Hypothese in Situationen, in denen eine Person über einen längeren Zeitraum Teilverdeckungen unterliegt, auch über einen längeren Zeitraum durch wenige Bildmerkmalskorrespondenzen aufrecht erhalten werden. Die verdeckten Teile einer Person können in diesen Situationen durch die Merkmale des Typs 3 (projizierte Hypothesenmerkmale ohne Bildmerkmalskorrespondenzen) als verdeckt identifiziert werden und dabei trotzdem zur Einschätzung des aktuellen Status der Person herangezogen werden. D.h. es wird automatisch geschätzt, wo die aktuell verdeckten Teile einer Person im Bild lokalisiert sind. Wie in der unteren Reihe von Abbildung 4.7 zu sehen ist, gilt gleiches auch für eine ähnliche Situation im

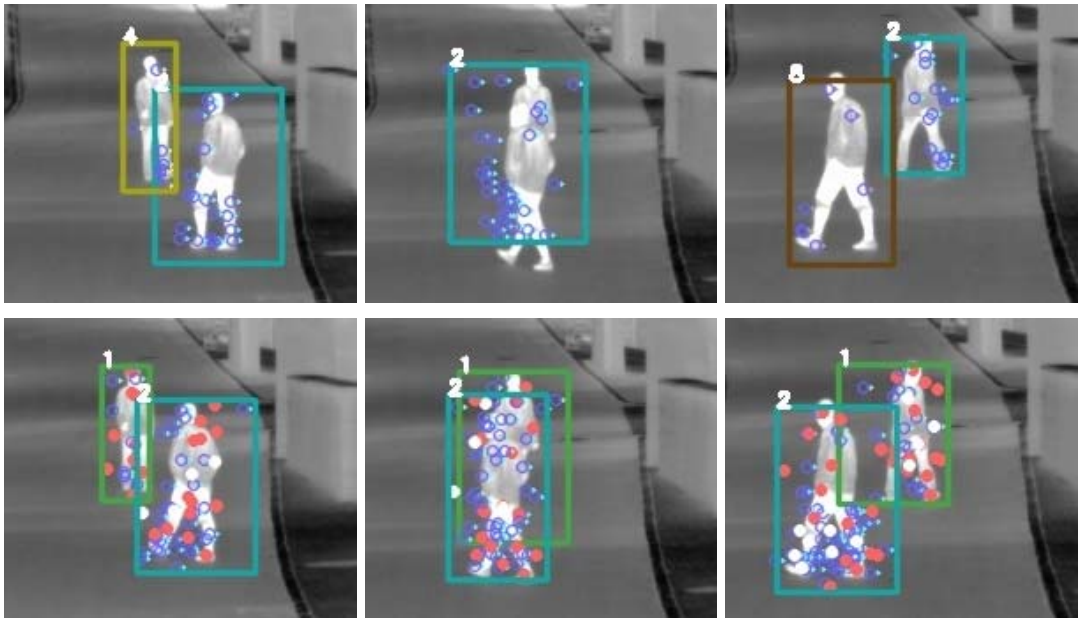


Abbildung 4.7: Vergleich tracking-by-detection (obere Reihe) mit dem in dieser Arbeit vorgestellten Tracking-Ansatz (untere Reihe)

infraroten Spektralbreich. Auch unter den hier schwierigeren Bedingungen für ein Trackingverfahren, nämlich dass eine Unterscheidung von Personen aufgrund der fehlenden Textur und der großen visuellen Ähnlichkeit von Personen eine große Herausforderung ist, ist ein korrektes Tracking bei der dargestellten Verdeckungssituation möglich. Das tracking-by-detection Verfahren mit nachträglicher Spurbildung scheitert hier, da für beide Personen während der Verdeckung nur eine Detektion generiert wird. Auch für mögliche Heuristiken, die eine solche Verdeckungssituation erkennen könnten und somit beide Hypothesen fortführen könnten, wäre diese Situation nicht unproblematisch, da die während der Verdeckungssituation generierte Detektion Merkmale beider Personen enthält und somit unabhängig davon, welcher Person sie zugeordnet würde, das zum Tracking genutzte Modell zumindest einer Person verfälschen würde.

4.5 Quantitative Auswertung

In diesem Abschnitt wird eine quantitative Auswertung des in diesem Kapitel vorgestellten Trackingverfahrens durchgeführt. Dazu werden Bildsequenzen aus verschiedenen Anwendungsbereichen, die unterschiedliche Herausforderungen für ein Trackingverfahren beinhalten, ausgewertet. Insbesondere erfolgt die Auswertung für Sequenzen aus dem infraroten sowie dem sichtbaren Spektralbereich, um die Unabhängigkeit des vorgestellten Verfahrens von der Sensormodalität an diesem Beispiel zu zeigen. Das Trackingverfahren bleibt bei beiden Modalitäten unverändert – lediglich die Trainingsdaten für den Personendetektor werden spezifisch gewählt.

4.5.1 Bewertungsmaße

Zur Bewertung des Trackings werden zusätzlich zu den bereits in Kapitel 3 eingeführten Performancemaßen zur Bewertung der Personendetektion weitere Bewertungskriterien eingeführt. Diese entsprechen den Standard CLEAR MOT[3] Bewertungskriterien wie sie in [24] vorgestellt wurden. Für weitere Details bzgl. der Performancemaße sei auf [24] verwiesen.

Die *Multiple Object Tracking Precision (MOTP)* gibt die Exaktheit von Objekthypothesen an:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_{i,t} gt_t^i}. \tag{4.9}$$

Hierbei ist d_t^i die Distanz (z.B. die Distanz der Objektzentren) zwischen einer korrekten Detektion und der Grundwahrheit. Da hier das Überschneidungskriterium zur Bewertung genutzt wird, wird nicht die Objektzentrumsdistanz, sondern der Überscheidungsgrad zwischen bounding box der Objekthypothese und der Grundwahrheit genutzt.

Die *Multiple Object Tracking Accuracy (MOTA)*:

$$MOTA = 1 - (\overline{m} + \overline{fp} + \overline{mm}) \tag{4.10}$$

bewertet die gesamte Trackingperformance unter Einbezug drei elementarer Fehlermaße. Der *miss-ratio (Falschnegativrate)* $\overline{m}$, der *false-positive-ratio (Falschpositivrate[4])* $\overline{fp}$ und der *mismatch-ratio (Falschzuordnungsrate)* $\overline{mm}$:

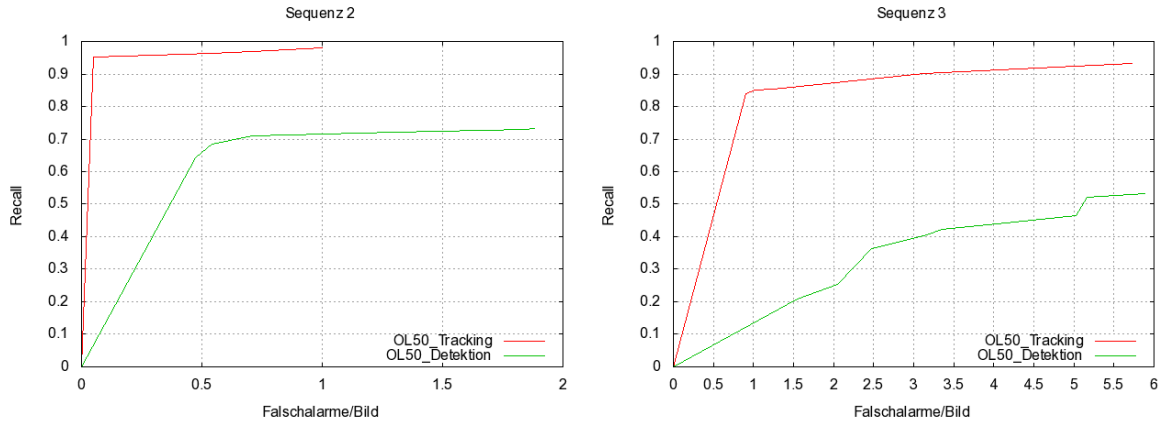
$$\overline{m} = \frac{\|m\|}{\|gt\|} \;\;,\;\; \overline{fp} = \frac{\|fp\|}{\|gt\|} \;\;,\;\; \overline{mm} = \frac{\|mm\|}{\|gt\|} \;, \tag{4.11}$$

die über die gesamte Bildsequenz akkumuliert werden. $\|gt\|$ ist hierbei die Anzahl der Objekte in der Grundwahrheit. Falschzuordnungen werden gezählt, wenn die Objekt-ID innerhalb des Tracking-Systems für das gleiche Objekt in der Grundwahrheit wechselt. Diese Auswertungsmetriken stellen den aktuellen Standard zur Trackingbewertung dar. Um auch einen Vergleich mit reinen Detektionsalgorithmen, die diese Metriken nicht nutzen, zu erlauben, werden in den Ergebnistabellen zusätzlich die bereits in Kapitel 3 genutzten Maße *recall* und *Falschalarme pro Bild* dargestellt.

Die Bewertung einer Objekthypothese als korrekt bzw. Falschalarm erfolgt wie in Abschnitt 3.6.4 durch das Überschneidungskriterium (Intersection over Union), das als Standard-Auswertekriterium in CLEAR MOT [24] genutzt wird. Bei allen folgenden Experimenten wird ein minimaler Überschneidungsgrad von 50% als Kriterium angelegt.

4.5.2 Infraroter Spektralbereich

In diesem Abschnitt erfolgt die Auswertung auf Daten aus dem infraroten Spektralbereich. Neben der Evaluierung zur Validierung der Funktionalität in dem für das Tracking schwierigeren infraroten Spektralbereich, liegt der Fokus hier besonders auf der Bewertung des Trackings bei bewegter Kamera. Hier soll aufgezeigt werden, in welchen Situationen das Kalman-Filter zur Dynamikmodellierung an seine Grenzen kommt und inwiefern eine Unterstützung durch das in Abschnitt 4.3.2 vorgestellte Modell zur Bewegungskompensation Abhilfe schaffen kann. Daneben soll auch gezeigt werden, inwiefern sich die Detektionsqualität, insbesondere hinsichtlich der Exaktheit der Detektion, im Vergleich zur Auswertung in Abschnitt 3.6.4 durch das Tracking verbessern lässt.



(a) (b)

Abbildung 4.8: ROC-Kurven (Recall in Abhängigkeit der Falschalarme pro Bild) für (a) Sequenz 2 und (b) Sequenz 3. Jedes Diagramm zeigt jeweils einen Graphen für die Performance des Trackings (rot) und die Performance der Einzelbilddetektion (grün) aus Kapitel 3. Die Auswertung erfolgte in beiden Fällen mit einem 50% Überschneidungskriterium.

4.5.2.1 Statische Kamera

Zur Auswertung bei statischer Kamera werden die bereits in Abschnitt 3.6.4 zur Auswertung der Detektionsqualität herangezogenen Daten des öffentlichen OTCBVS Datensatzes [46] genutzt. Diese Sequenz (2) beinhaltet nur wenige Schwierigkeiten für die Personenverfolgung, da es zu keinerlei Verdeckungen kommt und sich nur 2 Personen in der Szene aufhalten. Hier soll gezeigt werden, inwiefern sich die Detektionsqualität, insbesondere bzgl. der Genauigkeit der bounding box, durch das Tracking im Vergleich zur Einzelbilddetektion verbessert. Dazu sind in Abbildung 4.8 (a) Graphen für den re-

Tabelle 4.1: Tracking-Ergebnisse von Sequenz 1-7.

Sequenz	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frames	71	400	201	417	416	1664	216
Objekte (#ids)	301(7)	763(2)	1471 (8)	1119 (8)	1637 (13)	5426 (12)	417 (3)
MOTP	0.67	0.62	0.76	0.67	0.76	0.77	0.66
$\overline{m}$	0.43	0.05	0.16	0.16	0.35	0.20	0.1
$\overline{fp}$	0.05	0.02	0.12	0.08	0.04	0.12	0.1
$\overline{mm}$	0	0	0.001 (2)	0.002 (4)	0	0.001 (5)	0
MOTA	**0.52**	**0.93**	**0.72**	**0.76**	**0.61**	**0.68**	**0.79**
Recall	0.57	0.95	0.84	0.84	0.64	0.80	0.9
Falschalarme/Bild	0.22	0.04	0.94	0.28	0.11	0.37	0.25

[3]Serie von Workshops zum Thema „Classification of Events, Activities, and Relationships“ und „Multiple Object Tracking“.
[4]Falschpositiv wird synonym zu Falschalarm verwendet

call als Funktion der Falschalarme pro Bild, jeweils für das Tracking und die Einzelbilddetektion aus Kapitel 3 aufgetragen. Die Auswertung erfolgte jeweils mit dem 50% Überschneidungskriterium (siehe Abschnitt 3.6.3, Formel 3.22). Wie zu sehen ist, wird durch das Tracking eine immense Verbesserung der Detektionsperformance, genauer gesagt der Genauigkeit der Detektion hinsichtlich der bounding box, erreicht. So wird ein recall von 95% bei nur 0,04 Falschalarmen pro Bild erreicht. Ähnliche Werte wurden bei der Einzelbilddetektion in Abschnitt 3.6.4 für das bounding box Kriterium erreicht. Hier werden diese nahezu perfekten Werte allerdings bei einer 50% Überschneidungsanforderung erreicht. Es ist also deutlich zu sehen, dass das Tracking die Genauigkeit der Detektion immens verbessert. Die Trackingperformance für diese Sequenz (2) kann Tabelle 4.1 entnommen werden.

4.5.2.2 Bewegte Kamera

Die Auswertung der Personenverfolgung bei bewegter Kamera erfolgt auf Daten, die bei einer IOSB-Messkampagne aufgezeichnet wurden. Hierbei werden Sequenzen mit unterschiedlichem Schwierigkeitsgrad betrachtet. Der Schwierigkeitsgrad ist dabei durch Variationen in der Personenskalierung, durch Personenverdeckungen und insbesondere durch die Stärke der Kamerabewegung definiert.

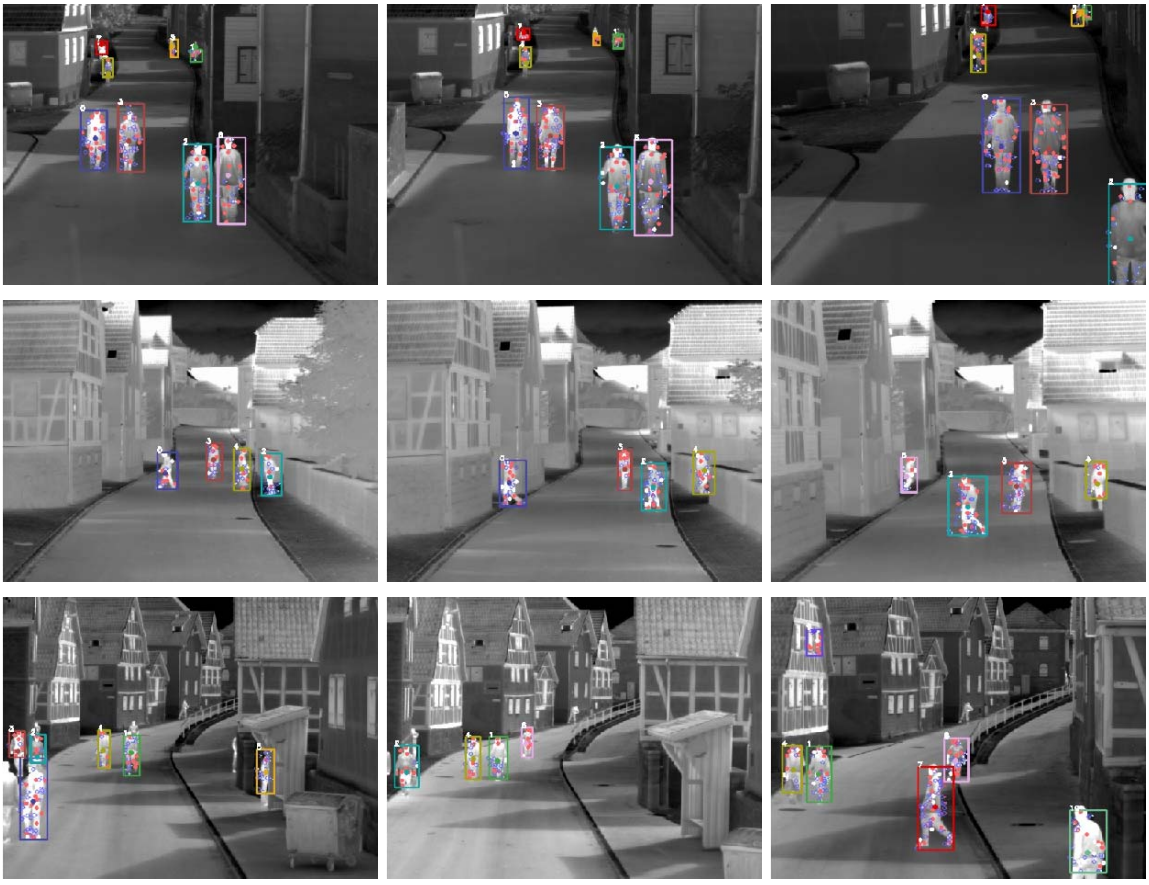


Abbildung 4.9: Beispielergebnisse des Trackings bei bewegter Kamera für Sequenz 3 (Reihe 1), Sequenz 4 (Reihe 2) und Sequenz 5 (Reihe 3).

Abbildung 4.10: Beispielergebnisse des Trackings bei bewegter Kamera für Sequenz 6.

Vor der Auswertung der Trackingperformance soll aber auch hier ein Vergleich der Detektionsperformance des Trackings und der Performance der Einzelbilddetektion aus Kapitel 3 erfolgen. Dazu wird wie bei Sequenz 2, die Detektionsperformance von Einzelbilddetektion und Tracking bei 50% Überschneidungsanforderung für Sequenz 3 (siehe Abschnitt 3.6.4) verglichen. Dieser Vergleich ist in Abbildung 4.8 (b) dargestellt. Wie anhand der Graphen zu sehen ist, wird hier bereits bei nur einem Falschalarm pro Bild ein recall von 85% erreicht. Dies ist eine Verbesserung von über 70% im Vergleich zur Einzelbilddetektion. Auch hier ist dies der höheren Genauigkeit der bounding boxes, aber auch der erhöhten zeitlichen Stabilität (keine Ausfälle der Detektion), die durch das Tracking erreicht wird, zu verdanken. Die Trackingperformance für diese Sequenz ist in Tabelle 4.1 dargestellt.

Zur Auswertung der Trackingperformance (diese beinhaltet auch eine Auswertung der Detektionsperformance) werden drei zusätzliche Sequenzen ausgewertet. Diese enthalten alle bei der Personenverfolgung potentiell auftretenden Herausforderungen, wie Personen auf unterschiedlichen Skalierungsstufen und Verdeckungen zwischen Personen. Alle Sequenzen wurden von einer bewegten Kamera mit einer Auflösung von 640x480 aufgezeichnet.

Ergebnisse dieser Sequenzen sind in Tabelle 4.1 dargestellt. Beispielergebnisse für die Sequenzen befinden sich in Abbildung 4.9 und Abbildung 4.10. Wie anhand der Auswertung zu sehen ist, wird in allen Sequenzen gute Performance erreicht.

Eine nähere Betrachtung soll von Sequenz 6 erfolgen, da hier die größten Herausforderungen für ein Tracking-System gegeben sind. Insgesamt enthält die Sequenz 5426 Auftreten von 12 unterschiedlichen Personen, wobei sich 9 Personen dicht gedrängt in einer engen Straße bewegen. Dabei kreuzen sich die Laufwege der Personen häufig, was zu totalen Verdeckungen der Personen führt. Dies ist insbesondere in diesen Infrarotsequenzen eine große Schwierigkeit, da die Personen selbst für den menschlichen Beobachter nicht voneinander zu unterscheiden sind und sich zudem die Kamera bewegt, was die Verlässlichkeit der Bildposition bei der Datenassoziation verringert. Personen treten hier auf stark unterschiedlichen Skalierungsstufen, in einem Größenbereich von 7x25 bis 52x150, auf. Ebenso kommt es hier nicht nur zu normalen, durch Kamera- und Personenbewegung sowie durch Verdeckungen

Abbildung 4.11: Vergleich des Tracking ohne Bewegungskompensation (obere Reihe) mit dem Tracking mit Bewegungskompensation (untere Reihe).

verursachten Erscheinungsveränderungen, sondern durch eine Explosion in der Szene kommt es zu extremen Beleuchtungsveränderungen wie sie sonst kaum zu beobachten sind. Beispielbilder dieser Gegebenheiten sind in Abbildung 4.10 zu sehen. Auch unter diesen schwierigen Bedingungen erreicht das hier vorgestellte Trackingverfahren gute Ergebnisse. So wird ein recall von 80% erreicht, wobei die fehlenden 20% hauptsächlich von 2 Personen im Hintergrund der Szene verursacht werden, die erst ab der Mitte der Sequenz detektiert werden. Die 12% Falschalarmrate wird hier hauptsächlich von kleinen Ungenauigkeiten in der bounding box verursacht, was bei den zahlreichen Verdeckungen nicht verwunderlich ist. Es gibt keinen „echten Falschalarm“ in der Sequenz. Lediglich wird die Hypothese einer Person, die den Sichtbereich der Kamera verlässt für eine zu lange Zeit aufrechterhalten, so dass hier kurzzeitig ein Falschalarm zu sehen ist. Mit nur 5 Identitätsvertauschungen bei zahlreichem Kreuzen der Laufwege ist die Performance des eigentlichen Trackings kaum steigerbar. Insgesamt wird hier eine MOTA von 68% und eine MOTP von 77% erreicht, was für eine Sequenz dieses Schwierigkeitsgrads eine sehr gute Performance darstellt.

4.5.2.3 Stark bewegte Kamera

In Abschnitt 4.3.2 wurde die Problematik des hier vorgestellten Trackingverfahrens bei sehr stark bewegter Kamera geschildert und ein Ansatz vorgestellt, der Abhilfe hierfür schafft.

Um diese Problematik zu verdeutlichen, sind in Abbildung 4.11 Beispielbilder einer Bildsequenz dargestellt, in der Probleme für das Tracking aufgrund abrupter Kamerabewegung entstehen. Die in der oberen Reihe gezeigten Ergebnisse sind mit dem hier vorgestellten Trackingverfahren ohne spezielle Bewegungskompensation entstanden – die in der unteren Reihe mit zusätzlicher Bewegungskompensation. Wie zu sehen ist, findet hier eine starke abrupte Bewegung der Kamera nach links statt. Dies hat, wie in der oberen Reihe zu sehen ist, ohne Bewegungskompensation zur Folge, dass die Personen nicht mehr korrekt verfolgt werden können und eine Identitätsvertauschung, in der Form, dass eine neue Hypothese für die gleiche Person generiert wird, stattfindet. In der unteren Reihe ist zu sehen,

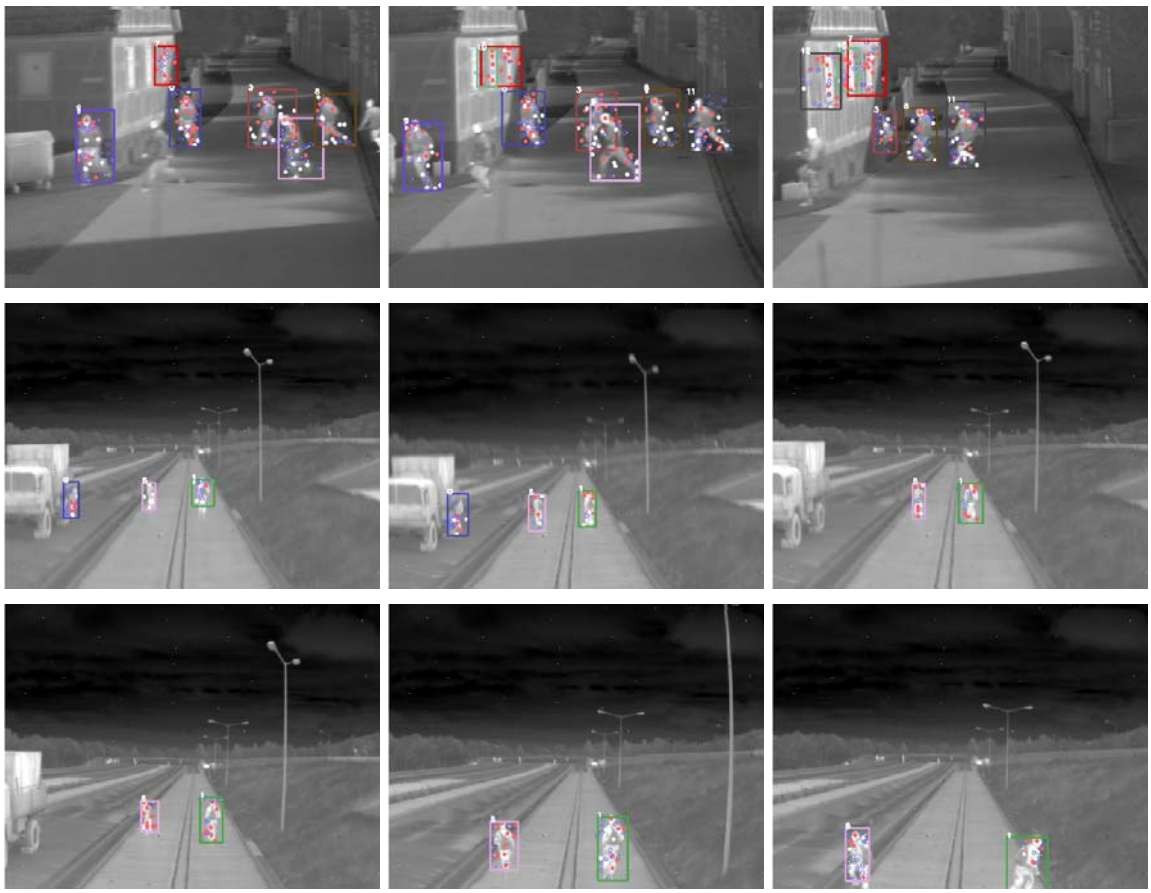


Abbildung 4.12: Beispielergebnisse des Trackings bei starker Kamerabewegung in Sequenz 1 (Reihe 1) und Sequenz 7 (Reihe 2-3). (Abbildungen aus [88])

dass in dieser Situation durch die Bewegungskompensation ein korrektes Tracking möglich wird.

Die quantitative Auswertung bei starker Kamerabewegung erfolgt in zwei Bildsequenzen. Die Ergebnisse der Auswertung von Sequenz 1 aus Abschnitt 3.6.4, bei der starke Kamerabewegung durch einen Kameraschwenk nach links verursacht wird, sind in Tabelle 4.1 zu sehen. Beispielbilder dieser Sequenz sind in Abbildung 4.12 abgebildet.

Die Auswertung in einer zweiten Sequenz (7), die zur Auswertung von Trackingverfahren unter extremen Bedingungen auf einer Fahrzeugteststrecke aufgezeichnet wurde, bestätigt die Funktionalität der integrierten Bewegungskompensation. Hier bewegt sich ein Fahrzeug auf einer mit künstlichen Bodenwellen versehen Strecke, was zu starker Kamerabewegung in vertikaler Richtung führt. Dies hat starke Versätze der Personen in Bildkoordinaten und auch Verzeichnungen im Bild selbst zur Folge. Wie in Tabelle 4.1 und in den Beispielergebnissen in Abbildung 4.12 zu sehen ist, wird auch hier eine gute Trackingperformance ohne Identitätsvertauschungen erreicht.

Tabelle 4.2: Ergebnisse des Trackings in verschiedenen Sequenzen aus dem sichtbaren Spektralbereich.

Sequenz	**INRIA**	**Pets06-1**	**Pets06-2**	**Hockey**	**Pets09**
Frames	1043	185	250	101	795
Objekte (#ids)	2142(3)	793 (8)	808 (6)	1026 (14)	4968 (20)
MOTP	0.55	0.76	0.69	0.65	0.70
$\overline{m}$	0.5	0.06	0.27	0.17	0.18
$\overline{fp}$	0.1	0.008	0.16	0.04	0.09
$\overline{mm}$	0.001 (3)	0	0	0	0.004 (21)
MOTA	**0.4**	**0.93**	**0.57**	**0.79**	**0.73**
Recall	0.5	0.94	0.73	0.83	0.82
Falschalarme/Bild	0.21	0.04	0.58	0.4	0.57

4.5.3 Sichtbarer Spektralbereich

In diesem Abschnitt erfolgt die Auswertung der Personenverfolgung in verschiedenen Sequenzen aus dem sichtbaren Spektralbereich. Nachdem im letzten Abschnitt insbesondere die Performance des Ansatzes bei bewegter Kamera betrachtet wurde, liegt der Fokus hier auf der Unabhängigkeit des Ansatzes von der Anwendungsumgebung. Dazu erfolgt die Auswertung in Bildsequenzen aus unterschiedlichen Szenarien.

Tabelle 4.2 zeigt die Ergebnisse der Auswertung. Wie zu sehen ist, wurde das Tracking in 5 Sequenzen getestet.

Sequenz *INRIA* stammt aus dem CAVIAR Testdatensatz [2]. Dieser wurde ursprünglich zur Evaluierung von Aktionserkennungssystemen aufgezeichnet, bietet aber auch für die reine Personenverfolgung einige Herausforderungen. So wurde die Sequenz mit einer niedrigen Auflösung von 384x288 Pixeln aufgezeichnet. Wie in den Beispielbildern dieser Sequenz in Abbildung 4.13 Reihe 1 zu sehen ist, erscheinen die Personen hier in sehr geringer Größe und insbesondere auch mit sehr geringem Kontrast zum Hintergrund. Eine weitere Schwierigkeit sind die starken Beleuchtungsveränderungen, die durch Lichteinfall durch eine Glasfront verursacht werden. Insbesondere variiert die Ansicht der Personen, gegeben durch die Kameraoptik, sehr stark von annähernd Seitenansicht im hinteren Bereich der Szene zu Draufsicht im vorderen Bereich. Hier wird eine MOTA von 0.4 erreicht, wobei es bei einem recall von 0.5 kaum Falschalarme gibt.

Die beiden Sequenzen *Pets06-1* und *Pets06-2* stammen aus dem PETS[5] 2006 Datensatz [3]. Dieser wurde an einem Bahnhof mit einer Auflösung von 720x576 aufgezeichnet. Personen sind hier in relativ hoher Auflösung mit gutem Kontrast zu sehen. Die Herausforderung in diesen Sequenzen liegt nicht in den schwierigen Umgebungsverhältnissen, sondern in der Anzahl der zu verfolgenden Personen und in den gegenseitigen Verdeckungen dieser. Verdeckungen werden hier durch Personengruppen, bzw. durch das Kreuzen der Laufwege von Personen verursacht.

Wie die Ergebnisse in Tabelle 4.2 und Abbildung 4.13 (Reihe 2) zeigen, wird dies ohne Falschzuordnung und einer MOTA von 0.93 bzw. 0.57 sehr gut erreicht.

[5]Serie von Workshops zur „Performance Evaluation of Tracking and Surveillance“.

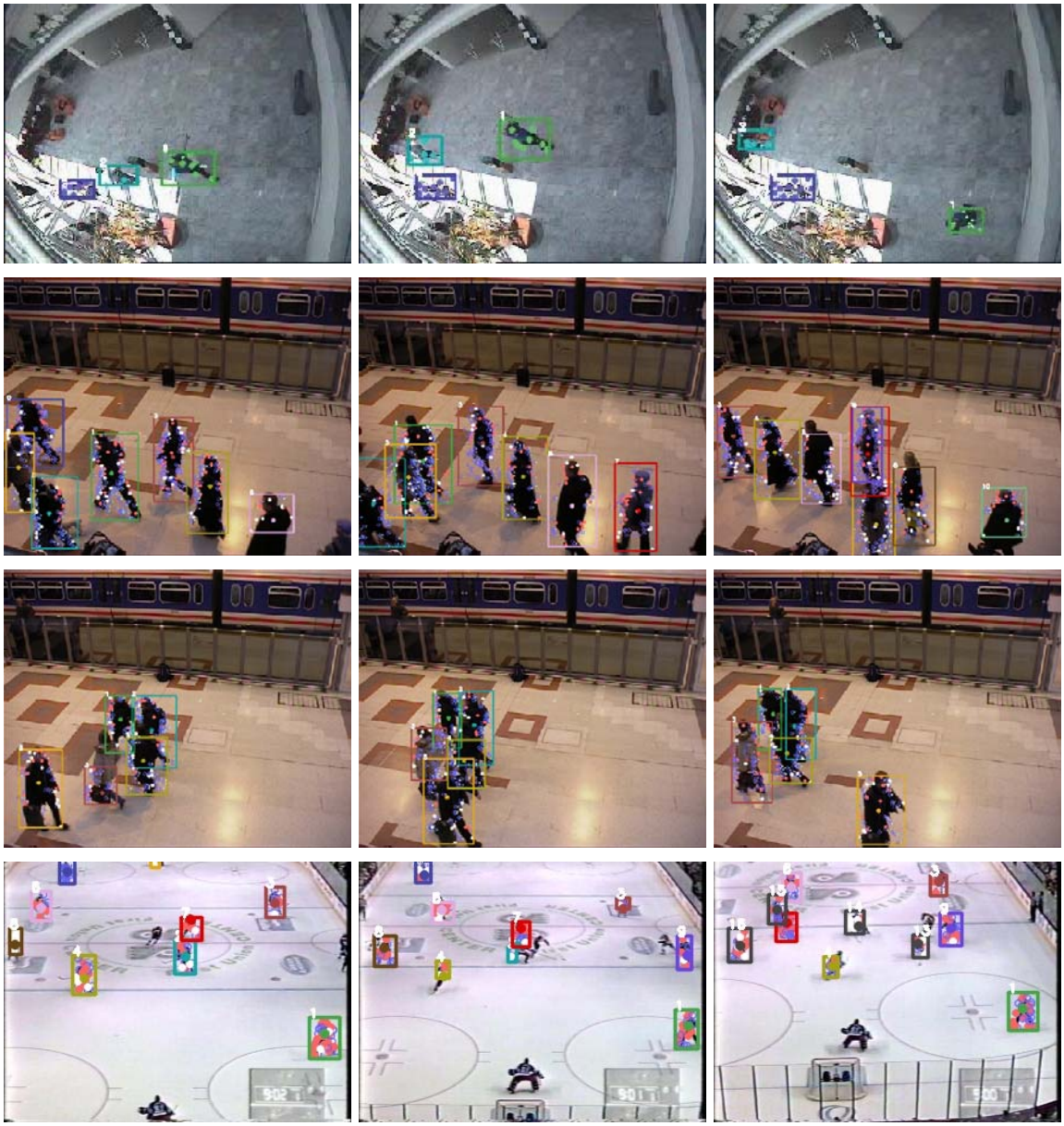


Abbildung 4.13: Beispielergebnisse des Trackings für verschiedene Sequenzen aus dem sichtbaren Spektralbereich: INRIA (Reihe 1), Pets06-1 (Reihe 2), Pets06-2 (Reihe 3) und Hockey (Reihe 4).

Nach diesen beiden Sequenzen aus typischen Überwachungsszenarios wird mit *Hockey* eine Testsequenz aus einem völlig anderen Anwendungsbereich, dem Sport, genauer einem Ausschnitt aus einem Eishockey-Spiel, betrachtet. Gerade weil diese Sequenz aus einem anderen Anwendungsbereich kommt wird sie als Testsequenz gewählt, um zu zeigen, dass das hier vorgestellte Verfahren allgemein anwendbar und nicht auf den Überwachungsbereich beschränkt ist. Insbesondere stehen für diese Sequenz auch Ergebnisse anderer state-of-the-art Verfahren zur Verfügung, so dass ein Performancevergleich erfolgen kann. Die Sequenz enthält einen Ausschnitt eines Eishockey-Spiels und wurde mit einer Auflösung von

320x240 aufgezeichnet. Die Herausforderung hier besteht in der Spieler- und Kamerabewegung sowie darin, dass die Spieler aufgrund gleicher Kleidung visuell nicht zu unterscheiden sind. Dies bringt für das Tracking die gleichen Schwierigkeiten wie im infraroten Spektralbereich mit sich, wobei dies, wie umfangreich gezeigt wurde, das hier vorgestellte Verfahren nicht vor große Probleme stellt. Eine weitere hier auftretende Schwierigkeit ist, dass die Personen nur in sehr geringer Größe von ca. 10x15 Pixeln zu sehen sind.

Tabelle 4.3: Trackingergebnisse der „Hockey“ Sequenz.

Verfahren	MOTP	**MOTA**	$\overline{m}$	$\overline{fp}$	$\overline{mm}$ (absolut)
Okuma et al,.	0.51	0.68	0.31	0.0	11
Breitenstein et al.	0.57	0.77	0.22	0.01	0
Jüngling	0.65	0.79	0.17	0.04	0

Die Ergebnisse sind in der Übersichtstabelle 4.2, sowie in der Vergleichstabelle 4.3 zu sehen. Es gibt keine einzige Falschzuordnung in der gesamten Sequenz und die MOTA Rate liegt bei 0.79. Damit zeigt das ISM-Tracking (siehe Tabelle 4.3) bessere Performance als das state-of-the-art Verfahren von Breitenstein et al. [31, 29], sowie wesentlich bessere Performance als die Methode von Okuma et al. [134]. Insbesondere übersteigt nicht nur die MOTA, welche die gesamte Performance von Detektion und Tracking angibt die der anderen Verfahren, sondern ebenso die MOTP, die die Genauigkeit der Detektion angibt. Dies zeigt, dass das ISM-Tracking, was zu jedem Zeitpunkt eine Modellaktualisierung mit den Bilddaten vornimmt, ein Verfahren wie das von Breitenstein et al. [29], das sich zu großem Teil auf eine Positionsprädiktion durch einen Partikelfilter stützt, in der Genauigkeit übertrifft.

Ein weiterer Vergleich des ISM-Trackings mit aktuellen state-of-the-art Verfahren erfolgt auf Sequenz *Pets09*. Diese Sequenz stammt aus dem aktuellen PETS 2009 Datensatz[6] [4] und stellt den aktuellen Standard-Benchmark für Trackingverfahren dar. Die Sequenzen des PETS 2009 Datensatzes wurden im Außenbereich aufgenommen und enthalten verschiedene gestellte Sequenzen, in denen gezielt herausfordernde Situationen für Tracking-Systeme herbeigeführt werden. So ändern in der hier genutzten Sequenz einige Personen abrupt ihre Laufrichtung und gehen rückwärts weiter, um die Schwierigkeit für eine Positionsprädiktion zu erhöhen. Desweiteren werden längere Teilverdeckungssituationen sowie zahlreiches Kreuzen der Laufwege von Personen herbeigeführt, was die Identitätswahrung im Tracking zusätzlich erschwert.

Die Ergebnisse dieser Sequenz sind in Tabelle 4.2, sowie Beispielbilder in Abbildung 4.15 zu sehen. Der Vergleich mit aktuellen state-of-the-art Verfahren [16, 23, 22, 30, 148, 170, 103, 7] aus den beiden 2009er PETS Workshops ist in Abbildung 4.14 zu sehen. Die Ergebnisse der anderen Verfahren stammen aus [50][7].

Wie hier zu sehen ist, liegt die Performance des ISM-Trackings in der Spitzengruppe. Bei MOTA auf Platz 4, wobei nur ein geringer Abstand von 2% bzw. 3% zu Platz 3 und 2 besteht. Bei MOTP liegt das ISM-Tracking sogar auf Platz 1. Insgesamt kann also festgestellt werden, dass die Performance des hier vorgestellten ISM-Trackings mit der aktueller state-of-the-art Verfahren konkurrieren kann. Hierzu ist insbesondere auch anzumerken, dass einige der Vergleichsverfahren im Gegensatz zum ISM-Tracking mehrere Kameraansichten des Datensatzes und Information über den Szenenhintergrund nutzen. Dies betrifft insbesondere auch das Verfahren „Berclaz_LP“, welches zwar die beste Gesamtperformance aufweist, hierzu aber auch die Daten von 7 Kameras nutzt. Bei Nutzung der Daten lediglich einer

[6] Sequenz: S2.L1,12.34

[7] Die Ergebnisse einiger Verfahren waren in den jeweiligen Artikeln nicht exakt ausgewiesen. Diese wurden aus dem Graphen in [50] abgelesen und können deshalb um 1-2% nach oben oder unten abweichen.

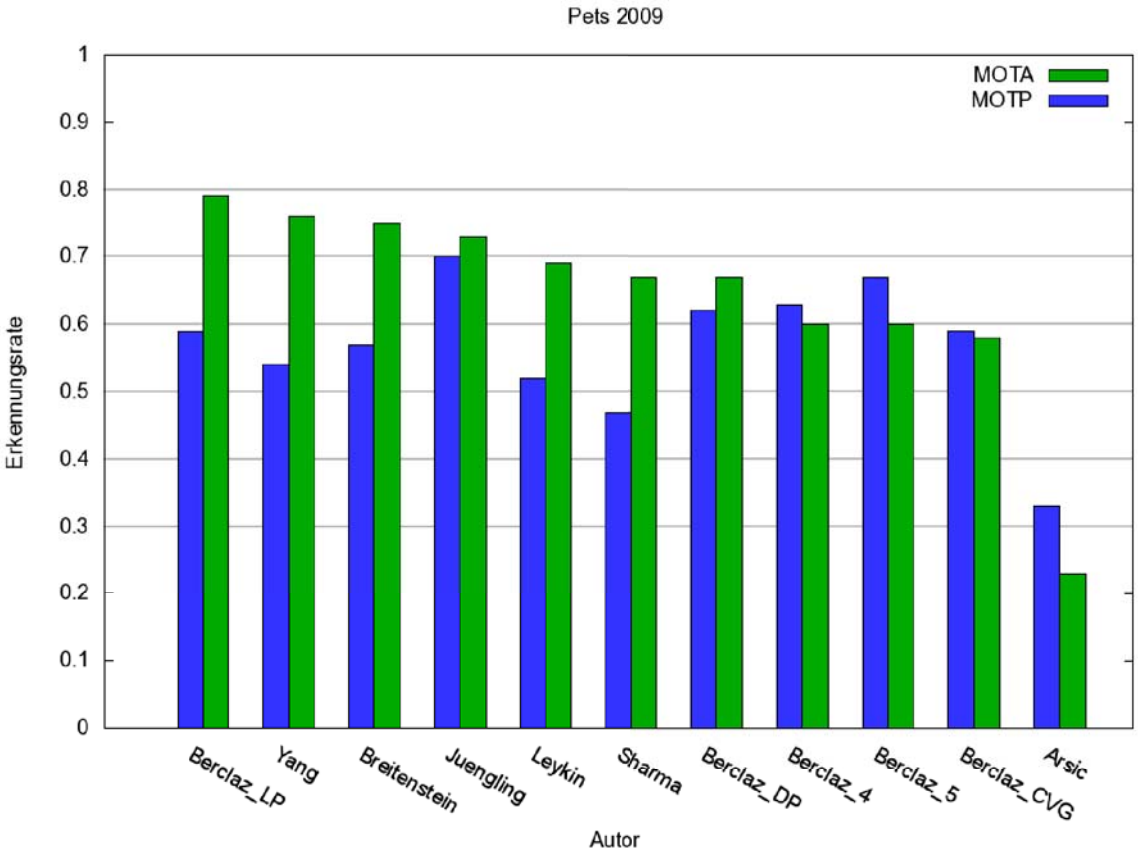


Abbildung 4.14: Tracking-Ergebnisse für Sequenz S2.L1,12.34 des PETS 2009 Datensatzes.

Kamera (siehe [22]) sinkt die MOTA des Verfahrens um mehr als 20% auf unter 0.6 ab. Dieses Verfahren, genauso wie das zweitbeste Verfahren von Yang et al. [170], baut insbesondere auch auf einer Personendetektion durch Vordergrundsegmentierung auf. Es wird also eine bekannte Szene und stationäre Kamera angenommen. Das insgesamt drittbeste Verfahren von Breitenstein et al. [30] macht diese Annahmen[8] nicht und ist deshalb auch am ehesten mit dem ISM-Tracking vergleichbar.

[8] Allerdings ist dieses Verfahren nicht mit dem ISM-Tracking bzgl. anderer Aspekte der Systemgenerizität vergleichbar. So werden dort z.B. Farbinformationen genutzt und das Verfahren ist aufgrund des genutzten Bewegungsmodells nicht ohne weiteres bei bewegter Kamera nutzbar.

Abbildung 4.15: Beispielergebnisse des Pets 2009 Trackings.

4.6 Ergebnisse anderer Objektklassen und hierarchische Objektverfolgung

Um die Objektunabhängigkeit des vorgestellten Trackingansatzes exemplarisch zu demonstrieren, sind in Abbildung 4.16 Beispielergebnisse des Trackings anderer Objektklassen gezeigt. Ebenso sind hier Beispielergebnisse der hierarchischen Objektverfolgung, die auf der hierarchischen Objektdetektion aus Abschnitt 3.5 aufbaut, dargestellt.

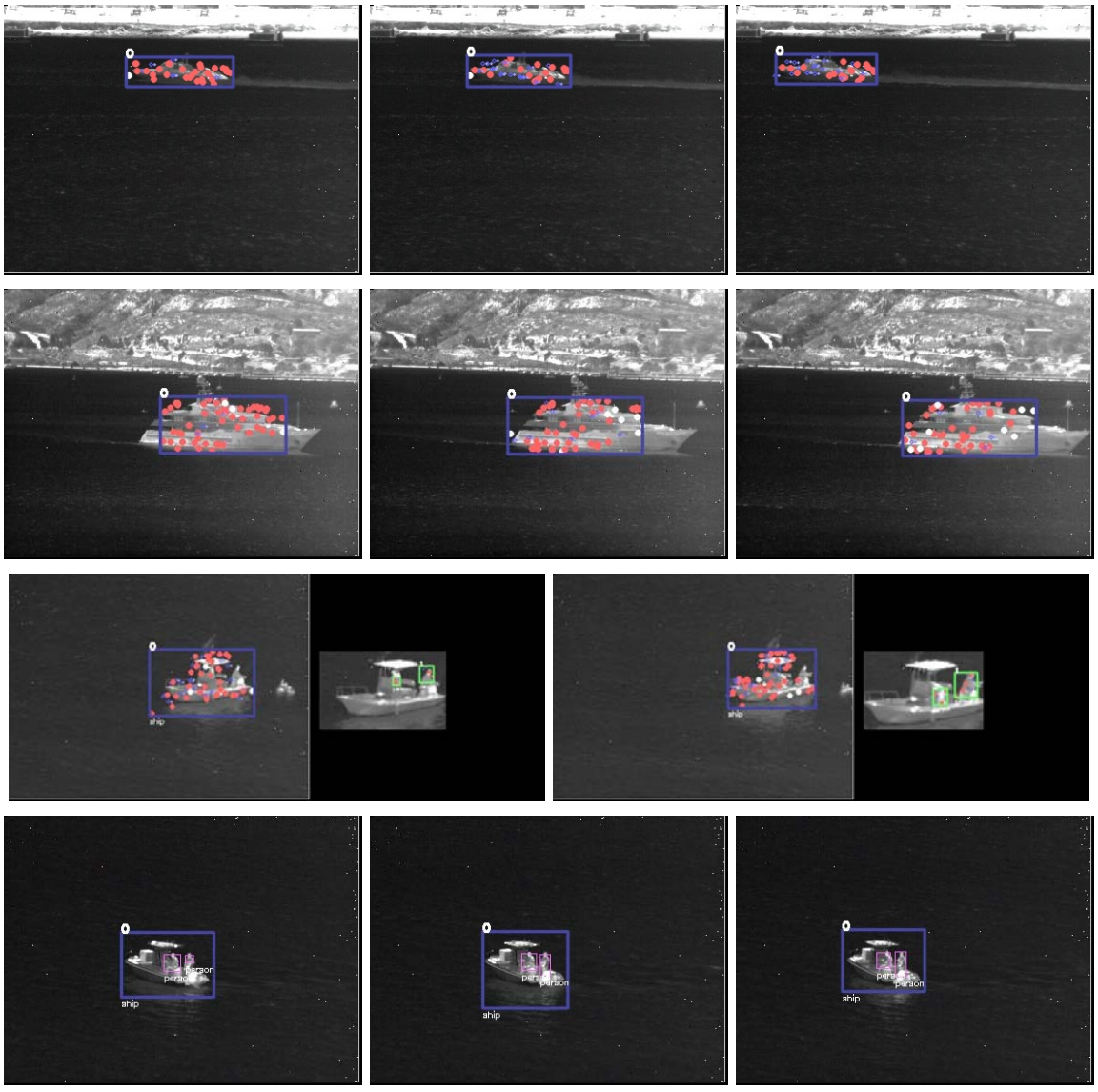


Abbildung 4.16: Ergebnisse des Trackings von Schiffen und der hierarchischen Objektverfolgung von Personen auf Schiffen.

Kapitel 5

Personenwiedererkennung

In diesem Kapitel wird ein Modell zur Personen-, bzw. allgemein Objektwiedererkennung vorgestellt, das auf den beiden vorherigen Kapiteln, der Personendetektion (Kapitel 3) und Personenverfolgung (Kapitel 4) insofern aufbaut, als dass die Personenverfolgung dazu genutzt wird, Personen in Bildsequenzen zu Detektieren und zu Verfolgen. Beim Tracking in Kapitel 4 werden Personen verfolgt, während sie sich in einer zeitlich zusammenhängenden Periode im Sichtbereich einer Kamera aufhalten. Sobald sie den Sichtbereich der Kamera verlassen, wird der „Track“ und damit die eindeutige Identitätszuordnung über der Zeit im System beendet. Sofern die gleiche Person den Sichtbereich der gleichen oder einer anderen Kamera wieder betritt, wird beim Tracking, wie es in Kapitel 4 vorgestellt wurde, eine neue ID vergeben. Es kann also kein Zusammenhang zwischen den beiden Tracks, die sich auf die gleiche Person beziehen, gezogen werden. An dieser Stelle greift die Personenwiedererkennung.

Unter Personenwiedererkennung versteht man, dass eine Person, die sich bereits im Sichtbereich der Kamera (oder einer Kamera in einem Multi-Kamera-System) aufgehalten hat, nach Verlassen und erneutem Eintreten in den Sichtbereich der Kamera als die zuvor bereits gesehene Person identifiziert wird. Es geht also nicht darum, eine Person wirklich mit Namen als z.B. Michael Müller zu identifizieren, sondern darum, eine Verbindung zwischen zeitlich nicht direkt zusammenhängendem Auftreten einer Person im Sichtbereich des Systems zu ziehen. Dabei kann sich, je nach Einsatzbereich, der Sichtbereich des Systems auch auf mehrere Kameras in einem Kameranetzwerk beziehen. D.h., die Aufgabenstellung für die Wiedererkennung kann die Wiedererkennung einer Person in der gleichen Kamera, aber auch die Wiedererkennung einer Person in unterschiedlichen Kameras, was natürlich aufgrund von unterschiedlichen Kameraperspektiven und Umgebungsbedingungen ein wesentlich schwierigeres Problem darstellt, bedeuten.

Der hier vorgestellte Ansatz zielt auf die Nutzung in unterschiedlichen Szenarien ab. So ist eine Nutzung sowohl in Szenarien sinnvoll, in denen die Person in hoher Auflösung frontal zu sehen ist und sogar das Gesicht der Person sichtbar ist – hier würde automatisch auch die Gesichtsinformation in das Modell eingebunden und zur Wiedererkennung genutzt, und zwar ohne dafür spezielle Heuristiken bzw. Verfahren zu benutzen, als auch in Szenarien, in denen die Personen nur in geringer Auflösung zu sehen sind. In diesen typischen Überwachungsszenarien, in denen biometrische Merkmale, wie Gesicht oder Iris nicht zur Wiedererkennung oder gar Identifikation nutzbar sind, ermöglicht das hier vorgestellte Verfahren trotzdem eine Wiedererkennung.

Vorteile des hier vorgestellten Ansatzes gegenüber Wiedererkennungsverfahren, die ebenfalls auf solche schwierigen Bedingungen abzielen sind:

(i) Die *Integration* der Wiedererkennung in ein Detektions- und Trackingverfahren. Die Kurzzeit-Identitätsmodelle aus Abschnitt 4.2 werden hierzu zu Langzeit-Identitätsmodellen, welche die gesamte Erscheinung einer Person modellieren, ausgebaut. Diese Modelle werden online während des Trackings

aufgebaut, wodurch im Gegensatz zu Verfahren, die auf einem offline durchzuführenden Trainingsschritt aufbauen oder solchen, die manuelle Annotation der Personen voraussetzen, eine Nutzung der Wiedererkennung in einem realen System möglich ist.

(ii) Die auch auf dieser Stufe bewahrte *Systemgenerizität*. D.h., dass auch zur Wiedererkennung ausschließlich lokale Bildmerkmale und keine sensorspezifischen Merkmale wie Farbe genutzt werden. Dies ermöglicht als erstes erscheinungsbasiertes Verfahren überhaupt insbesondere auch die Wiedererkennung im infraroten Spektralbereich. Dadurch, dass die Personenwiedererkennung direkt auf dem Tracking aufbaut und keine anderen Methoden benutzt, wird auch hier die Unabhängigkeit vom Anwendungsszenario und die Objektgenerizität beibehalten. Das gesamte System bis zur Wiedererkennung ist also auch im mobilen einsetzbar.

Im Detail wird die Wiedererkennung in einem *mehrstufigen* Verfahren mit ansteigender Komplexität und Distinktivität durchgeführt:

Auf *Stufe 1* werden die Codebuch-Aktivierungssignaturen der Personen zur Wiedererkennung genutzt. Diese erlauben einen effizienten Vergleich von Identitätsmodellen, da hier lediglich Vektoren der Codebuchdimension verglichen werden. Auf *Stufe 2* werden die Aktivierungssignaturen durch die ISM-Ortsverteilung erweitert. Hierdurch wird zusätzlich zur reinen Erscheinung auf Stufe 1 also auch Ortsinformation, die insbesondere die spezifische Personenform sowie den Auftretensort bestimmter Texturen modelliert, einbezogen. Hierdurch erhöht sich die Distinktivität im Vergleich zu Stufe 1, wobei die Komplexität nur unwesentlich erhöht wird. Auf *Stufe 3* werden die SIFT-Merkmalsdeskriptoren verglichen. Hier wird die Codebuchsignatur zur Indizierung und die ISM-Ortsverteilung zur Modellierung der Struktur genutzt. Durch dieses mehrstufige Verfahren wird maximale Distinktivität mit Effizienz verbunden. Insbesondere wird hierbei sowohl die Form einer Person als auch die auf dem Personenabbild gefundene Textur modelliert.

Desweiteren wird eine Methode zur Erhöhung der Ansichtsinvarianz der Personenwiedererkennung durch Generierung ansichtsspezifischer Identitätsmodelle vorgestellt. Hier wird eine ISM-Ansichtstransformation vorgestellt, die es erlaubt, Modelle, die in unterschiedlichen Ansichten generiert wurden, ineinander zu überführen um somit einen korrekten Modellvergleich zu gewährleisten.

Das Kapitel ist wie folgt aufgeteilt: Zunächst wird in Abschnitt 5.1 das zur Wiedererkennung genutzte Modell sowie dessen Aufbau vorgestellt. In Abschnitt 5.2 wird der 3-stufige Ansatz zum Vergleich von Identitätsmodellen vorgestellt. Abschnitt 5.3 widmet sich der Ansichtsinvarianz der Personenwiedererkennung, wobei in Abschnitt 5.3.1 der Aufbau von ansichtsspezifischen Identitätsmodellen und in Abschnitt 5.3.2 die Auswahl des Modells zur Wiedererkennung vorgestellt wird. In Abschnitt 5.3.3 wird auf die Thematik der Ansichtsinvarianz beim Modellvergleich eingegangen und in Abschnitt 5.3.4 wird hierzu eine Modelltransformation anhand des Beispiels der Spiegelung vorgestellt. In Abschnitt 5.4 wird die experimentelle Validierung der ISM-Wiedererkennung durchgeführt.

5.1 Individualisierung des Allgemeinen: Ein Modell zur Objektwiedererkennung

Die hier vorgestellte Personenwiedererkennung baut auf den beiden in den Kapiteln 3 und 4 vorgestellten Verfahren zur Personendetektion und -verfolgung auf. Dabei bedeutet aufbauen hier nicht nur, dass eine im Tracking bestimmte Position (bounding box) einer Person zur Berechnung von Merkmalen für die Wiedererkennung genutzt wird, sondern, dass die für die Personendetektion und -verfolgung genutzten lokalen Bildmerkmale (SIFT) auch unmittelbar zur Wiedererkennung verwendet werden. Die auf den ersten beiden Stufen vorhandene allgemeine Anwendbarkeit des Modells wird also auch auf der dritten Stufe – bei der Personenwiedererkennung – bewahrt.

Dazu werden die Kurzzeit-Identitätsmodelle aus Abschnitt 4.2, die während des Trackings aufgebaut werden, zu Langzeit-Identitätsmodellen zur Wiedererkennung ausgebaut. Insbesondere wird auch zur

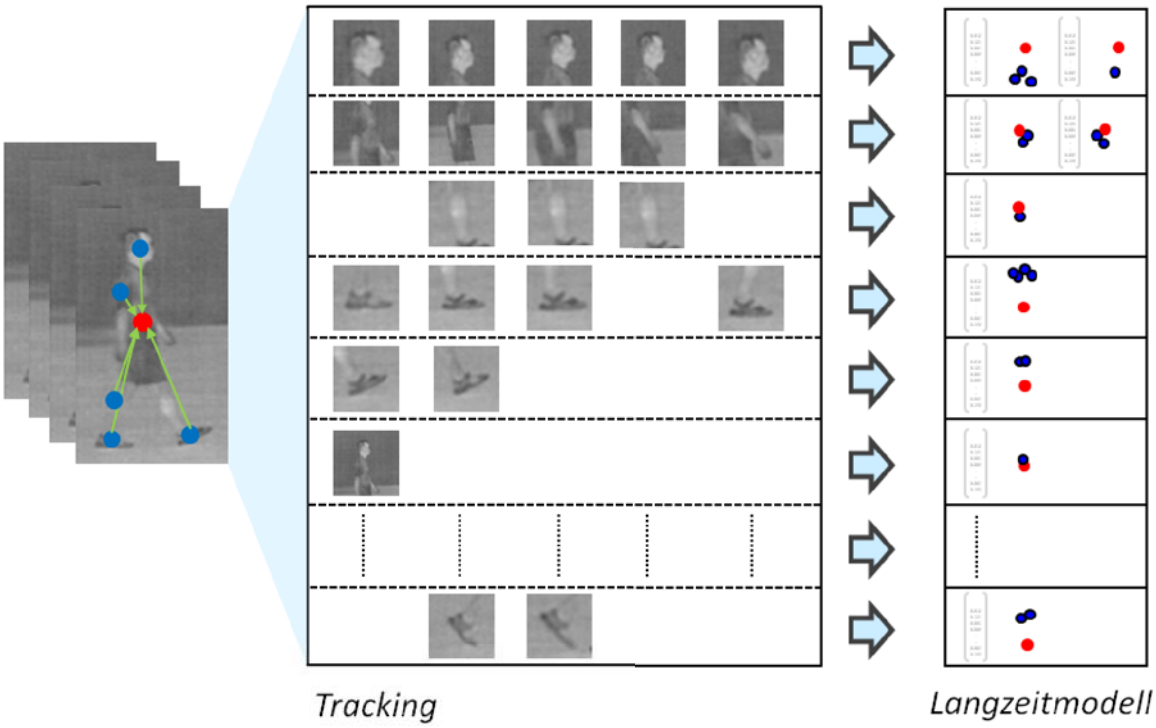


Abbildung 5.1: Aufbau von Merkmalsmodellen während des Trackings.

Wiedererkennung das Implicit Shape Model genutzt. Hier wird das allgemeine Modell, welches zur Detektion von Instanzen von Objektklassen entwickelt und in Kapitel 4 für das Tracking erweitert wurde, auf einer weiteren Ebene genutzt. Dabei wird aus dem allgemeinen Modell ein individuelles Modell, so dass es zur Wiedererkennung von Individuen nutzbar ist.

5.1.1 Aufbau von Merkmalsmodellen

Zur Wiedererkennung werden die in Abschnitt 4.2 vorgestellten Kurzzeit-Identitätsmodelle, die während des Trackings aufgebaut und zum Tracking verwendet werden, zu Langzeit-Identitätsmodellen ausgebaut. Dies ist notwendig, da die Kurzzeit-Modelle hochdynamisch sind und lediglich die aktuelle Erscheinung der Person modellieren sollen. Merkmale, welche die Person zwar in der Vergangenheit gut modelliert haben, mit der aktuellen Erscheinung der Person aber nichts zu tun haben, und somit nicht für das aktuelle Tracking der Person von Nutzen sind, werden aus dem Kurzzeit-Identitätsmodell automatisch entfernt. Zur Wiedererkennung ist es aber nötig, möglichst viele Informationen über mögliche Erscheinungen einer Person im Bild zu erhalten, da die Personenwiedererkennung invariant gegenüber Escheinungsveränderungen sein soll. Wird in das Modell also Information über die Erscheinungsvarianzen der Person integriert, kann bei der Wiedererkennung durch Rückgriff auf diese Information eine möglichst große Invarianz hergestellt werden. Somit soll die gesamte während des Trackings akquirierte Information in das Langzeit-Identitätsmodell einfließen.

Dazu wird das Kurzeit-Modell, wie in Abbildung 5.1 skizziert, zu einem Langzeit Modell ausgebaut. Durch das Bilden von Merkmalskorrespondenzen während des Trackings (siehe Abschnitt 4.1.1) ergibt sich ein durch $P_{typ}^{\pi,t}$ (siehe Formel 4.7) ausgedrückter Stabilitätswert. Dieser kann als notwendiges Zugangskriterium für das Langzeitmodell verwendet werden, da er ein Maß für die Relevanz bestimmter

Merkmale innerhalb des Kurzzeitmodells darstellt. Somit wird verhindert, dass Merkmale, die instabil sind oder durch Detektionsunzulänglichkeiten (z.B. Hintergrundmerkmale) in das Kurzzeitmodell integriert wurden, in das Langzeitmodell übernommen werden und dieses verfälschen. Merkmale, die dieses Stabilitätskriterium erfüllen, werden in das Langzeitmodell übernommen. Hier werden die durch die zeitliche Korrespondenzbildung im Tracking eindeutigen Merkmalsidentitäten zur Ablage in Merkmalsclustern genutzt. Anders als in den Kurzzeit-Trackingmodellen, bei denen nur jeweils der aktuelle Deskriptor (das aktuelle Bildmerkmal, welches durch Korrespondenz mit einem bereits in der Hypothese vorhandenen Merkmal eine Fortsetzung einer Merkmalstrajektorie bildet) für eine Merkmalstrajektorie gespeichert wird, werden in den Langzeitmodellen zur umfassenden Beschreibung alle Merkmalsdeskriptoren ab dem Zeitpunkt, an dem ein Merkmal die Stabilitätsanforderung erfüllt, abgelegt. Diese umfassende Erfassung aller Merkmalsausprägungen ist hier zielführend, da eine möglichst umfassende Beschreibung der Person erwünscht ist. Da hierdurch allerdings auch eine sehr große Datenmenge entsteht, ist es wünschenswert diese zu reduzieren. Dies kann zum Einen durch Selektion nur bestimmter Merkmale geschehen (eine Beschreibung dieser Merkmalsselektion erfolgt in Abschnitt 5.1.2), ebenfalls aber durch ein Clustering der Daten.

Die im Langzeitmodell abgelegten Merkmale werden zur Absenkung des Speicherbedarfs, insbesondere aber auch zur Reduzierung des späteren Vergleichsaufwands geclustert. Da der Modellaufbau und damit auch das Clustering online während des Trackings stattfinden muss, kann zum Clustering kein optimaler, die Datenmenge am besten reduzierender Ansatz genutzt werden, da hierfür die gesamte Datengrundlage zum Zeitpunkt des Clusterings zur Verfügung stehen muss. Das Clustering wird also hier online durch ein einfaches Verfahren durchgeführt, das für jeden dem Langzeitmodell neu hinzugefügten Deskriptor den Abstand zu den vorhandenen Clustern prüft. Befindet sich der Abstand unter einem Schwellwert τ_L, so wird der neue Deskriptor dem Cluster hinzugefügt und das Clusterzentrum, welches als Repräsentant des Clusters genutzt wird, neu berechnet. Ansonsten wird ein neuer Cluster mit dem Deskriptor als Repräsentanten erstellt. Wichtig ist in diesem Zusammenhang, dass das Clustering nur innerhalb der Deskriptoren einer Merkmalstrajektorie erfolgt. Es kann also hier bereits von einer durch den Deskriptorabstand zur Korrespondenzbildung τ_D^{MAX} gegebenen Mindestähnlichkeit ausgegangen werden, wobei für das Clustering hier eine stärkere Ähnlichkeit gefordert wird, um eine zu große Generalisierung zu vermeiden.

Zur Erstellung des Langzeitmodells werden die Clusterzentren als visuelle Beschreibung abgelegt. Zusätzlich zu dem Deskriptor wird pro Cluster die Ortsverteilung der Zentrumsversätze, sowie ein Vektor mit den Codebuchaktivierungsstärken abgelegt. Die Ortsverteilung sowie die Codebuchaktivierungsstärken werden unter Betrachtung aller dem Cluster zugeordneten Merkmale erstellt. D.h. jedes Merkmal, welches zur Erstellung des Clusters beigetragen hat, hat einen Eintrag in der Ortsverteilung und trägt zur Berechnung des Aktivierungsvektors bei. Zur Erstellung des Codebuchaktivierungsvektors wird pro Codebucheintrag jeweils die stärkste Aktivierung dieses Eintrags aus der Merkmalsmenge übernommen. Wichtig ist außerdem, dass die Ortsverteilung nicht die ursprünglichen Zentrumsversätze, unter denen die Merkmale im Bild gesehen wurden, festhält, sondern dass hier die skalierungsnormalisierten Versätze abgelegt werden. Hierdurch wird die ISM-Wiedererkennung skalierungsinvariant. Die Skalierungsnormalisierung findet mit der Skalierungsstufe des jeweiligen Merkmals statt.

Das Langzeitmodell einer Person enthält somit pro Tracking-Merkmalstrajektorie einen Modellcluster. Dieser enthält eine Menge von Merkmalsclustern, die ihrerseits eine Ortsverteilung und Codebuchaktivierungsstruktur enthalten. An dieser Stelle wird also die spezifische ISM-Ausprägung einer Person festgehalten und damit die Individualisierung des ISM zur Personenwiedererkennung vollzogen. Aktivierungsstruktur und Ortsverteilung dienen beim Modellvergleich in Abschnitt 5.2 zusätzlich zu den eigentlichen Deskriptoren als Merkmal und werden zur Reduzierung des Vergleichsaufwands und zur Erhöhung der Distinktivität der Modelle genutzt.

5.1.2 Merkmalsselektion

Während des Trackings werden Bildmerkmale gesammelt und im Identitätsmodell der Person gespeichert. Sobald Merkmale die Kriterien hierfür erfüllen, werden sie vom Kurzzeit-Trackingmodell in das Langzeitidentitätsmodell überführt. Dieses nimmt alle Merkmale (welche die Kriterien erfüllen) einer Hypothese über die gesamte Lebensdauer einer Hypothese auf, um alle vorkommenden Ansichtsveränderungen in das Modell zu integrieren. Zur Absenkung des Speicherbedarfs werden die Merkmale im Langzeitmodell geclustert. Trotzdem ist die Merkmalsmenge immens groß, was selbst im Fall eines effizienten Ansatzes zum Vergleich von Merkmalsmodellen zu hohem Vergleichsaufwand für die Merkmalsmengen führt. Da viele der auf einer Person gefundenen Merkmale sehr allgemeine Merkmale (wie z.B. Kanten an der Seite der Person) sind, die nicht unbedingt zur Diskriminierung der Person beitragen, kann die Merkmalsmenge durch Entfernen dieser reduziert werden. Insbesondere wird hiermit auch die Diskriminierungsfähigkeit des Modells erhöht, da das Ziel ist, nur die besonders diskriminierenden Merkmale auszuwählen.

Da in dem hier vorgestellten Ansatz im Gegensatz zu anderen Ansätzen [180, 72, 70] nicht angenommen wird, dass alle Personen vorher bekannt sind und die Aufgabe lediglich darin besteht, Klassifikatoren zur Unterscheidung der bekannten Personenmenge zu trainieren, ist es nicht möglich, speziell solche Merkmale auszuwählen, welche die vorhandenen Personen am besten diskriminieren. Allerdings ist es zur Erhöhung der Distinktivität der Identitätsmodelle möglich, solche Merkmale auszuwählen die besonders distinktiv und besonders signifikant für einzelne Personen sind. Um diese online während des Trackings auszuwählen, werden die Merkmale vor der Überführung in das Langzeitmodell mit einem allgemeinen Personenmodell verglichen. Merkmale, die eine hohe Ähnlichkeit zu diesem allgemeinen Modell haben, werden als nicht distinktiv genug für die Wiedererkennung eingestuft und nicht in das Identitätsmodell übernommen. Hierdurch werden auch ansichts- und artikulationsspezifische Merkmale gefiltert die nicht für die Wiedererkennung geeignet sind (da diese auch bei der gleichen Person variabel sind).

Das allgemeine Personenmodell wird in einem offline-Schritt aus einer Menge von durch Tracking generierten Trainingsmodellen erstellt. Hierbei werden Merkmale, die wiederholt in den Modellen unterschiedlicher Personen auftreten als allgemeine Merkmale definiert und in das allgemeine Personenmodell übernommen. In der Praxis wird die Merkmalsselektion für das allgemeine Modell durch ein Clustering der in den Trainingsmodellen enthaltenen Merkmale durchgeführt. Cluster, die eine hohe Anzahl an beitragenden Merkmalen und eine geringe Varianz aufweisen, werden in das allgemeine Modell übernommen. Dieser Schritt ist ähnlich zur Erstellung des zur Personendetektion genutzten Codebuchs, wobei dabei mehr Wert auf die Erstellung eines generellen Modells gelegt wird. Insbesondere kann der hier vorgestellte Schritt zur Erstellung eines allgemeinen Personenmodells, wie in Abschnitt 5.3.1.1 beschrieben, auch ansichtsspezifisch erfolgen um den maximalen Nutzen aus der Merkmalsselektion zu ziehen.

Integration zusätzlicher Merkmale

Die Merkmale, die während des Trackings gesammelt werden sind alle über das allgemeine Personencodebuch aktiviert worden. Zur Wiedererkennung sind aber gerade auch Merkmale relevant, die nur auf bestimmten Personen gefunden werden und somit nicht unbedingt über das generelle Codebuch integriert werden können. Um solche Merkmale zu integrieren, werden zusätzliche SIFT-Merkmale im Bereich der bounding box einer Person extrahiert. Aus der Menge der extrahierten Merkmale wird durch zeitliche Betrachtung eine Submenge ausgewählt, welche die gleiche Bewegungsdynamik wie die Tracking-Merkmale aufweisen. Hierdurch werden Hintergrundmerkmale, die im Falle von bewegten Personen eine andere Bewegungsdynamik als die Personenmerkmale aufweisen, ausgefiltert. Die verbleibende Merkmalsmenge wird dem Langzeitmodell der Person hinzugefügt, wobei hier die zusätzlichen Kriterien aus Abschnitt 5.1.1 gelten. Die Integration der Merkmale findet dabei auf gleichem Weg statt wie bei den während des Trackings akquirierten Merkmalen.

5.2 Ein effizienter Ansatz zum Vergleich von Merkmalsmodellen

Wie in Abschnitt 5.1.1 beschrieben wurde, werden während des Trackings Langzeit-Identitätsmodelle von Personen aufgebaut. Jedes dieser Modelle besteht aus einer Menge von Merkmalsclustern, welche die Erscheinung der modellierten Person beschreiben. Um diese zur Wiedererkennung nutzen zu können, muss ein Ansatz zum Vergleich dieser Modelle entwickelt werden. Dabei kommt es auf zwei wesentliche Aspekte an:

(i) Die vorhandene Information soll umfassend ausgenutzt werden um größtmögliche Distinktivität der Modelle zu gewährleisten. Dies bedeutet, dass nicht nur die Merkmalsdeskriptoren, sondern die gesamte ISM-Ausprägung inklusive der Ortsinformation genutzt wird, um einen möglichst diskriminativen Vergleich zu ermöglichen. Die Schwierigkeit hierbei besteht darin, die Invarianz des Modells gegenüber Erscheinungsveränderungen trotz der hohen Diskriminativitätsanforderungen beizubehalten.

(ii) Der Vergleich der Identitätsmodelle soll möglichst effizient gestaltet werden. Effizient bezieht sich in diesem Zusammenhang auf den Rechenaufwand, der zum Vergleich eines Anfragemodells mit Modellen in einer Datenbasis erforderlich ist.

Beide Punkte, Diskriminativität des Modells und geringer Vergleichsaufwand sind besonders relevant, da in der Praxis oft große Datenbasen mit vielen Personen vorliegen.

In diesem Abschnitt wird ein dreistufiger Ansatz zum Vergleich von Merkmalsmodellen vorgestellt, der diese Anforderungen erfüllt. Die drei Vergleichsstufen sind dabei mit ansteigender Komplexität und Distinktivität modelliert, so dass sie in einem integrierten System aufeinander aufbauen. Das Klassifikationssystem ist so ausgelegt, dass Stufe 1 und Stufe 2 einen sehr effizienten Vergleich von Merkmalsmodellen erlauben, die Distinktivität im Vergleich zu Stufe 3 aber geringer ist. Diese Stufen sollen zur Filterung von Modellen genutzt werden, um die Menge der zu vergleichenden Modelle, und damit die Komplexität auf Stufe 3 zu reduzieren. In diesen Stufen sollen möglichst viele Modelle in der Datenbasis ausgeschlossen werden, wobei sicherzustellen ist, dass das korrekte Modell in der Menge der verbleibenden Modelle enthalten ist.

Auf *Stufe 1* werden die *Codebuchsignaturen* der Merkmalsmodelle verglichen. Diese Stufe hat sehr geringe Komplexität, da hier nur Vektoren der Codebuchdimension N miteinander verglichen werden müssen. Aufgrund des hohen Maßes an Abstraktion ist allerdings auch die Distinktivität dieser Stufe im Vergleich mit den folgenden Stufen geringer.

Auf *Stufe 2* werden die *ISM-Aktivierungen* miteinander verglichen. Zusätzlich zu den auf Stufe 1 verwendeten Codebuchsignaturen, wird hier noch die ISM-Ortsverteilung einbezogen. Im Vergleich zu Stufe 1 erhöht sich durch die zusätzlich genutzte Information die Distinktivität, allerdings auch der Vergleichsaufwand.

Auf *Stufe 3* werden die eigentlichen *Deskriptoren* der Modellcluster verglichen. Diese Stufe beinhaltet die größte Distinktivität, damit allerdings auch den größten Vergleichsaufwand. Da auf den Stufen 1 und 2 bereits ein Großteil der Modelle ausgeschlossen werden können, muss der Vergleich auf diesen Stufen allerdings nur noch für eine kleine Menge von Modellen durchgeführt werden. Die eigentliche Klassifikation, also die Personenwiedererkennung findet auf dieser Stufe statt.

5.2.1 Stufe 1: Codebuchsignaturen

5.2.1.1 Aufbau von Signaturen

Der Aufbau der Codebuchsignaturen geschieht auf Basis des in Abschnitt 5.1.1 beschriebenen Langzeit-Identitätsmodells. Zum Aufbau der Gesamtsignatur werden die Signaturen der einzelnen Modellcluster des Identitätsmodells zusammengeführt. Die Signaturen der Cluster basieren auf den Merkmalssigna-

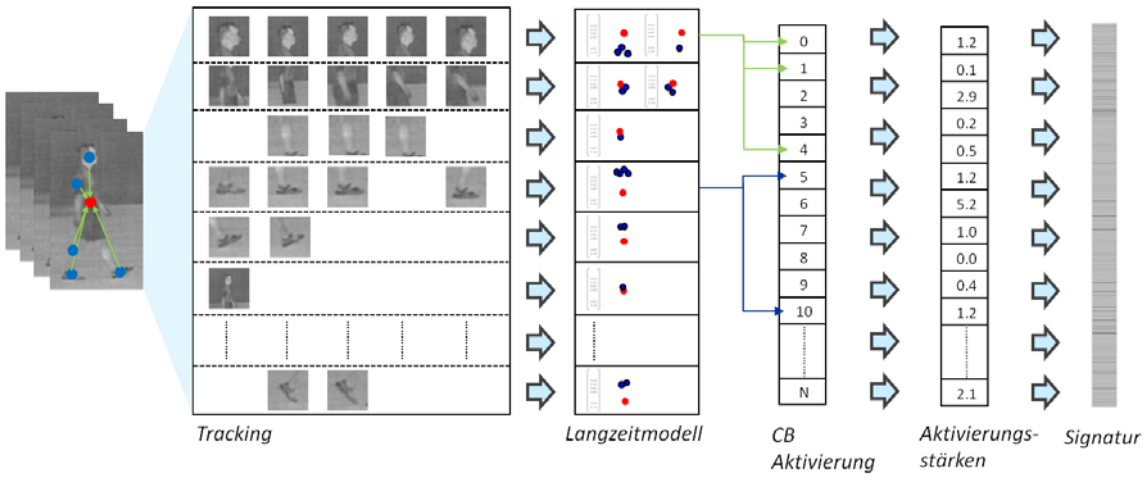


Abbildung 5.2: Aufbau von Codebuchsignaturen: Die Codebuchaktivierung jedes im Langzeitmodell vorhandenen Merkmalsclusters wird der Gesamtsignatur des Modells hinzugefügt. Diese bildet sich durch Akkumulation der Aktivierungssignaturen der einzelnen Cluster.

turen, die bereits bei der Objektdetektion durch Vergleich des Bildmerkmals mit dem Codebuch generiert wurden und beinhalten Informationen darüber, welche Codebuchprototypen der Deskriptor wie stark aktiviert hat. Die Aktivierungsstärke liegt durch Normalisierung mit dem Schwellwert τ_ρ^{train} (Formel 3.17) für jeden Prototyp im Intervall $[0, 1]$, wobei typischerweise weniger als 5 Einträge ungleich 0 sind. Wie in Abbildung 5.2 zu sehen ist, wird die Gesamtsignatur eines Modells durch Zusammenführen der Signaturen der Modellcluster generiert. Dazu wird pro Codebucheintrag n die Signatur aller I Cluster θ_i eines Modells der Gesamtsignatur Θ hinzugefügt:

$$\Theta_n = \sum_{i=0}^{I} \theta_{i,n}. \tag{5.1}$$

n ist hierbei die Aktivierung des n-ten Codebuch-Prototypen. Diese Gesamtsignatur modelliert dabei die Gesamtaktivierung durch einen einzelnen Vektor der Codebuchdimension N (die Dimension liegt typischerweise im Bereich von 200-1000).

Die Generierung der Codebuchsignaturen kann auch als Quantisierung in ein diskretes Histogramm gesehen werden. Wobei das Histogramm nicht den gesamten Wertebereich des Deskriptorraums abdeckt, sondern durch das Codebuch gegeben ist. In der Literatur werden diese Ansätze auch als *Bag-of-Words (BoW)* oder *Bag-of-Features (BoF)* bezeichnet und häufig im Bereich des Information Retrievals genutzt. Dort wird der Ansatz neben der ursprünglichen Anwendung des Text Retrievals auch im Bereich des Image Retrievals zur Beschreibung eines Bildes durch einen visuellen Wortschatz (visual words) genutzt [150]. Der Unterschied des hier vorgestellten Ansatzes zu diesen Ansätzen ist mehrschichtig. Zum Aufbau der Signaturen wird hier eine zeitliche Betrachtung durch das Tracking vorgenommen. Die Information wird also zeitlich akkumuliert und insbesondere auch einer zeitlichen Konsistenzprüfung (nicht alle Merkmale werden in das Langzeitmodell übernommen) unterzogen. Die Aktivierungen werden hierbei unscharf vorgenommen. Ein großer Unterschied ist auch die Anwendung. Hier wird eine Unterscheidung von Objekten der gleichen Klassen angestrebt, wobei die Ansätze in anderen Arbeiten vornehmlich zur Beschreibung gesamter Bildinhalte genutzt wurden. Die Anforderungen bei der Nutzung des Ansatzes sind also unterschiedlich, was auch durch die duale Nutzung des Codebuchs zur Personendetektion und -beschreibung verdeutlicht wird.

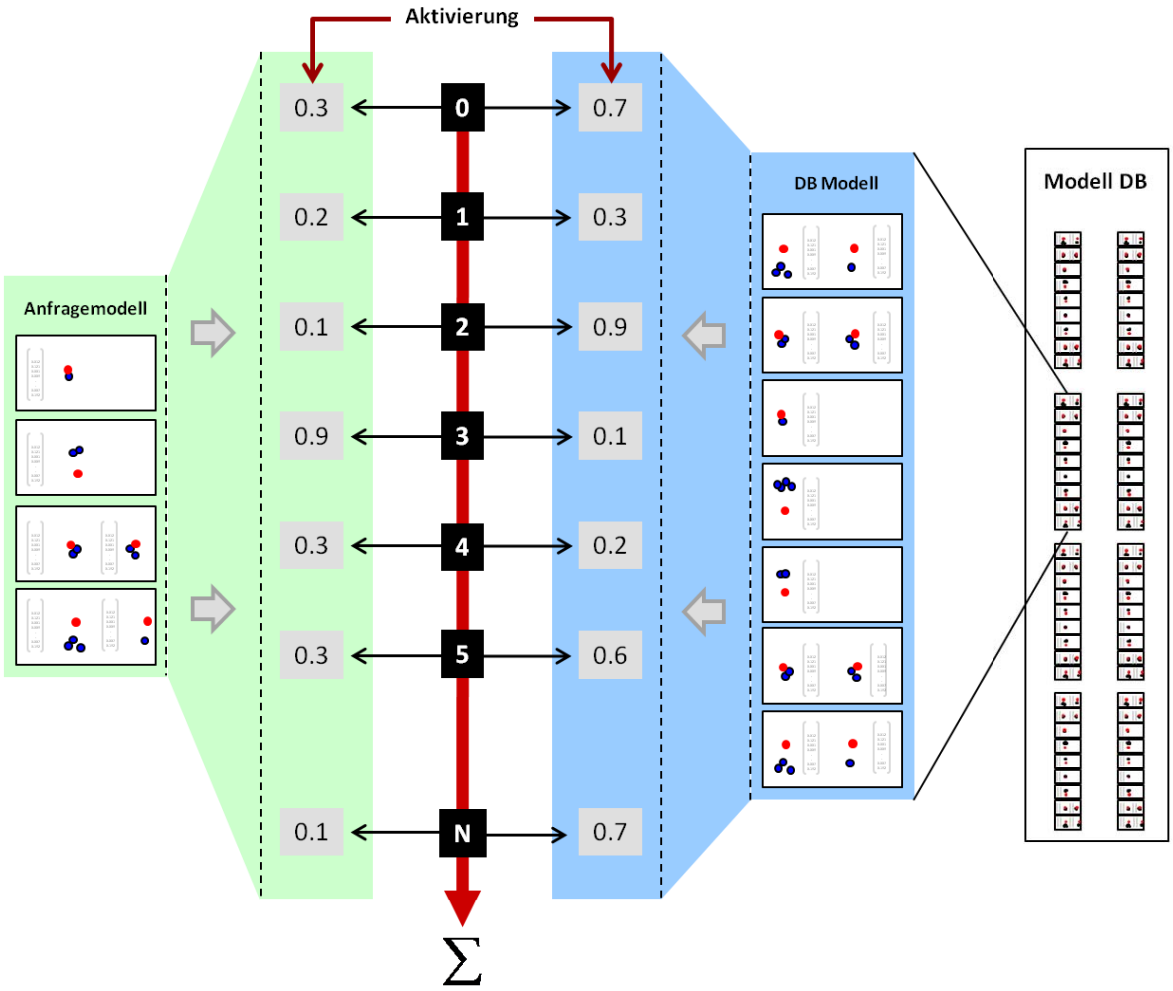


Abbildung 5.3: Vergleich von Codebuchsignaturen: Die Codebuchsignatur wird auf Basis aller Cluster des Identitätsmodells aufgebaut. Zur Wiedererkennung wird die Signatur mit den Codebuchsignaturen aller Datenbasismodelle verglichen, wobei der Vergleich durch Summenbildung über den absoluten Differenzen der einzelnen Signatureinträge stattfindet.

5.2.1.2 Vergleich von Signaturen

Die erste Vergleichsstufe zwischen Merkmalsmodellen ist der Vergleich der Codebuchaktivierungssignaturen. Diese beziehen noch nicht die Ausprägungen der im Bild gefundenen Deskriptoren ein, sondern stellen die Identitätsmodelle in Form von Aktivierungsstärken von Codebucheinträgen dar. Der Vorteil hierbei liegt auf der Hand. Im Gegensatz zu einer Menge von hochdimensionalen Merkmalsdeskriptoren muss hier nur ein Vektor der Codebuchdimension N verglichen werden. Der Nachteil ist, dass die Diskriminierungsfähigkeit eingeschränkt wird.

Diese erste Stufe kann wie alle folgenden Stufen direkt zur Klassifikation des Anfragemodells genutzt werden. Allerdings kann angenommen werden, dass diese Stufe aufgrund der im Vergleich zu den folgenden Stufen geringeren Diskriminierungsfähigkeit nicht zur Diskriminierung größerer Datenbestände geeignet ist. D.h., es kann davon ausgegangen werden, dass mehrere Modelle von ähnlich

erscheinenden Personen auch sehr ähnliche Aktivierungsprofile aufweisen und deshalb anhand der Aktivierungsprofile keine eindeutige Klassifikation vorgenommen werden kann. Da aber angenommen werden kann, dass das richtige Modell in der Datenbasis ebenfalls eine genügend hohe Ähnlichkeit der Aktivierungsstruktur aufweist, kann diese Stufe zur Auswahl einer Menge von Modellen aus der Datenbasis genutzt werden um damit den Vergleichsaufwand für die folgenden Stufen zu reduzieren. Diese Aussagen werden in der experimentellen Validierung in Abschnitt 5.4.1 belegt.

In der Praxis gibt es zwei unterschiedliche Möglichkeiten diese erste Stufe zur Reduzierung des Vergleichsaufwands in den folgenden Stufen durch Reduktion der zu vergleichenden Datenmenge zu nutzen. Die erste Möglichkeit ist die Einführung eines festen Schwellwerts für die Aktivierungsähnlichkeit. Modelle, deren Aktivierungsähnlichkeit diesen unterschreitet, können schon in dieser Stufe ausgeschlossen werden. Für diese Modelle muss bei der Identitätsbestimmung in folgenden Stufen also kein Vergleich mehr durchgeführt werden. Bei dieser Art der Verwendung ist sicherzustellen, dass der Schwellwert nicht zu niedrig gewählt wird, so dass das korrekte Modell nicht auch ausgefiltert wird. Eine andere Verwendungsmöglichkeit der Aktivierungsähnlichkeit ist die zur Auswahl der K besten Modelle aus der Datenbank. Hierbei muss K so gewählt werden, dass das korrekte Modell in den K besten Modellen enthalten ist. Der Vorteil dieser Methodik ist, dass unabhängig von der Datenbankgröße und Ähnlichkeit zwischen den Personen immer eine feste Anzahl an Modellen in der nächsten Stufe ausgewertet werden muss und somit eine feste Laufzeitanforderung eingehalten werden kann.

Die Ähnlichkeit zweier Identitätsmodelle ζ und η basiert dabei auf der Summe der Abstände über alle Codebucheinträge:

$$\chi_{AC}(\zeta,\eta) = \frac{1}{N}\sum_{n=0}^{N}\left(X - |\frac{|\zeta_n|}{\zeta_T} - \frac{|\eta_n|}{\eta_T}|\right). \tag{5.2}$$

Wobei hier eine Normalisierung mit der Tracking-Zeit ζ_T bzw. η_T der jeweiligen Modelle stattfindet. X ist hierbei eine Konstante die zur Umwandlung des Abstandes in Ähnlichkeit genutzt wird, wobei die Wahl dieser Konstante unkritisch ist. Der Vergleich der Codebuchsignaturen ist in Abbildung 5.3 visualisiert.

5.2.2 Stufe 2: ISM-Aktivierungen

Auf dieser Stufe werden zusätzlich zu den Codebuchaktivierungen die räumlichen Merkmalsverteilungen in den Vergleich einbezogen. Pro Codebucheintrag ist somit zusätzlich eine 2D-Ortsverteilung zu vergleichen (ca. $N * 2 * 10^2$ zusätzliche Vergleiche bei durchschnittlich 10 Positionseinträgen pro Codebucheintrag). Insbesondere wird hierdurch aber die Distinktivität erhöht, da nun auch miteinbezogen wird, wo auf einer Person ein bestimmter Codebucheintrag aktiviert wurde. Auf dieser Stufe wird, wie in Abbildung 5.4 zu sehen ist, also die eigentliche ISM-Ausprägung für den Vergleich von Identitätsmodellen genutzt:

$$\chi_{Off}(\zeta,\eta) = \frac{1}{N}\sum_{n=0}^{N}\left[\left(X - |\frac{|\zeta_n|}{\zeta_T} - \frac{|\eta_n|}{\eta_T}|\right)\cdot\beta_S(\zeta_n,\eta_n)\right]. \tag{5.3}$$

Hierbei ist X die gleiche Konstante wie in Abschnitt 5.2.1.2. $\beta_S(\zeta_n,\eta_n)$ definiert die Bedingung für die Übereinstimmung der Ortsverteilungen:

$$\beta_S(\zeta_n,\eta_n) = \begin{cases} 1, & \text{wenn } \min_{i,k}(dist_{eukl}(\zeta_i,\eta_k)) < \delta_S^{MAX} \\ 0, & \text{sonst} \end{cases}. \tag{5.4}$$

δ_S^{MAX} gibt dabei die obere Grenze des erlaubten Abstands zwischen den Ortsverteilungen an. Es muss innerhalb der Verteilungen also mindestens ein Element vorhanden sein, welches diesen (euklidischen) Abstand unterschreitet. Ist dies nicht der Fall, so wird $\beta_S(\zeta_n,\eta_n)$ in Formel 5.3 und damit auch die Aktivierungsähnlichkeit dieses Eintrags zu 0.

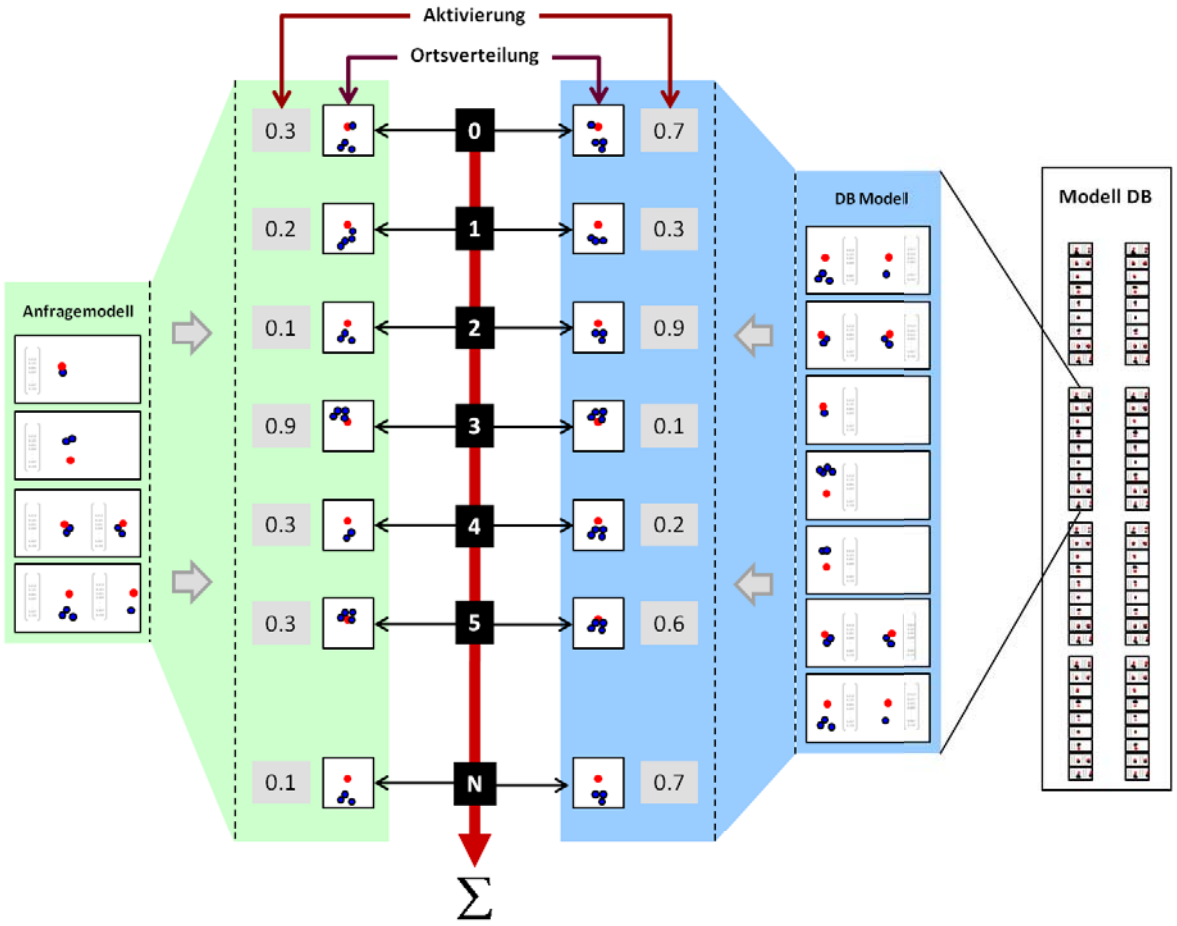


Abbildung 5.4: Vergleich von ISM-Aktivierungen: Zusätzlich zu Stufe 1 werden hier die Ortsverteilungen der Codebuchaktivierungen in den Vergleich einbezogen. Nur Codebucheinträge, die eine Übereinstimmung in den Ortsverteilungen aufweisen, werden in die Ähnlichkeitsbestimmung einbezogen.

5.2.3 Stufe 3: Merkmalsdeskriptoren

Auf dieser Stufe werden die in den Identitätsmodellen vorhandenen Modellcluster direkt verglichen. Anders als auf den vorherigen Stufen geht der Vergleich direkt von den Modellclustern des Anfragemodells aus. Für jeden dieser Cluster wird für ein Datenbasismodell der Modellcluster mit der höchsten Ähnlichkeit gesucht.

Um auch hierbei den Vergleichsaufwand zu reduzieren, wird wie bereits in Stufe 1 und 2 die Codebuchaktivierung von Merkmalen, hier zur Indizierung der zu vergleichenden Deskriptoren, genutzt.

Wie in Abbildung 5.5 zu sehen ist, werden vor dem Vergleich der Merkmalsdeskriptoren eines Clusters die Aktivierungsprofile verglichen. Sofern hier eine Übereinstimmung vorhanden ist, wird der Vergleich auf der nächsten Stufe, den Ortsverteilungen, durchgeführt. Diese dienen wiederum als Ausschlusskriterium für den eigentlichen Vergleich der Clusterdeskriptoren. Tatsächlich muss der aufwändigere Vergleich der Clusterdeskriptoren also nur für eine Teilmenge der Cluster erfolgen. Die Hinzunahme der Ortsverteilungen hat neben der erhöhten Effizienz natürlich auch die Eigenschaft, dass die Diskriminativität des Modells erhöht wird. Dies ist der Fall, da die Ortsverteilung kodiert, an welcher

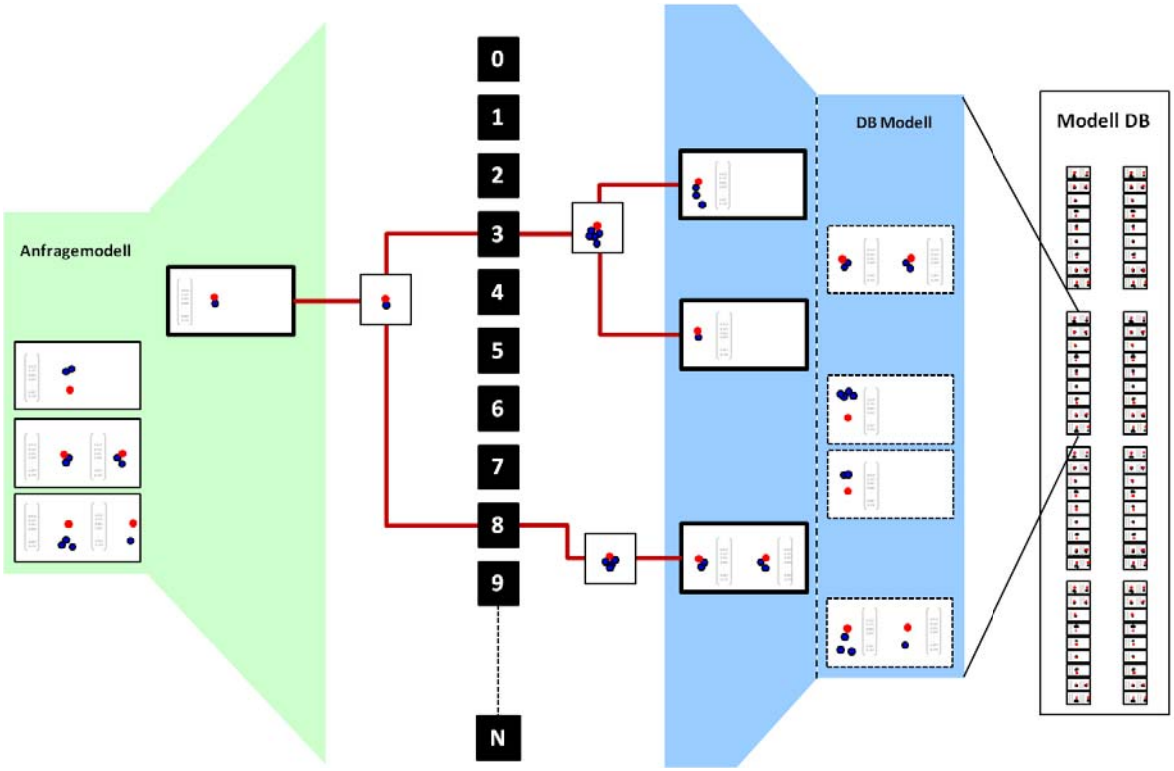


Abbildung 5.5: Vergleich von Merkmalsdeskriptoren: Für jeden Modellcluster des Anfragemodells wird der ähnlichste Cluster im Datenbasismodell gesucht. Der Vergleich der Clusterdeskriptoren findet dabei nur statt, wenn die Codebuchaktivierungen und Ortsverteilung der Cluster eine Übereinstimmung ergeben. Die Ähnlichkeit zweier Identitätsmodelle ergibt sich aus der Summe der Ähnlichkeiten aller Cluster des Anfragemodells.

Stelle welche Merkmale auftreten und somit die tatsächliche Form der Person beschreibt. Der Vergleich der Aktivierung bzw. der Ortsverteilung, wird hier nach dem gleichen Schema wie in Abschnitt 5.2.2 durchgeführt.

Die Ähnlichkeit $\chi_{DESC}(\zeta, \eta)$ zwischen einem Anfragemodell ζ mit K Modellclustern (jeder dieser Einträge enthält wiederum mehrere Cluster, siehe Abschnitt 5.1.1) und einem Datenbasismodell η wird durch die Summe über alle Modellcluster des Anfragemodells berechnet:

$$\chi_{DESC}(\zeta, \eta) = \frac{\sum_{k=0}^{K}(\beta_{DS}(\zeta_k, \eta))}{\phi(\zeta, \eta)}. \tag{5.5}$$

$\phi(\zeta, \eta)$ ist der Normalisierungsfaktor, der die jeweiligen Tracking-Zeiten ζ_T und η_T der Modelle einbezieht:

$$\phi(\zeta, \eta) = \zeta_T + \eta_T. \tag{5.6}$$

Anstatt dieses Normalisierungsfaktors können auch andere Normalisierungsmethoden, wie z.B. Einbezug der Anzahl der Modellcluster, genutzt werden.

Die Ähnlichkeit $\beta_{DS}(\zeta_k, \eta)$ eines Modellclusters ζ_k mit dem Datenbasismodell η ergibt sich aus dem minimalen Deskriptorabstand des Clusters mit allen Modellclustern des Datenbasismodells:

$$\beta_{DS}(\zeta_k, \eta) = \delta_{DS}^{MAX} - (min(min_i(\delta_{DS}(\zeta_k, \eta_i)), \delta_{DS}^{MAX})). \tag{5.7}$$

δ_{DS}^{MAX} gibt hierbei den Maximalwert für den Deskriptorabstand vor. Der Vergleich zweier Deskriptorcluster innerhalb eines Modellclusters:

$$\delta_{DS}(\zeta_k, \eta_i) = \min_{a,b}(\nu_{AC}(\zeta_{k,a}, \eta_{i,b}) \cdot \nu_S(\zeta_{k,a}, \eta_{i,b}) \cdot \delta_D(\zeta_{k,a}, \eta_{i,b})), \tag{5.8}$$

bezieht vor Berechnung des Deskriptorabstandes δ_D:

$$\delta_D(\upsilon, \psi) = \sum_{i=0}^{I}(\upsilon_i^D - \psi_i^D)^2 \tag{5.9}$$

die Zugangsfunktionen ν_{AC} und ν_S ein, die eine Übereinstimmung der Codebuchsignaturen im Fall von ν_{AC}, bzw. der Ortsverteilungen im Fall von ν_S prüfen:

$$\nu_S(\upsilon, \psi) = \begin{cases} 1, & \text{wenn } \min_{i,k}(dist_{eukl}(\upsilon_i^S, \psi_k^S)) < \delta_S^{MAX} \\ \infty, & \text{sonst} \end{cases}, \tag{5.10}$$

$$\nu_{AC}(\upsilon, \psi) = \begin{cases} 1, & \text{wenn } \exists\, \upsilon_i^{AC} \in \upsilon_I^{AC}, \upsilon_i^{AC} = \psi_j^{Ac} \\ \infty, & \text{sonst} \end{cases}. \tag{5.11}$$

In der Praxis ist es häufig sinnvoll, nicht die absolute Ähnlichkeit zweier Merkmalsmodelle, sondern das Verhältnis von bester und zweitbester Übereinstimmung in der Datenbasis zur Klassifikation zu nutzen. Der Vorteil hierbei ist, dass dieser Wert unabhängig von der absoluten Ähnlichkeit ist. Somit muss kein in der Praxis schwierig zu bestimmender Klassifikationsschwellwert für einen Absolutwert gesetzt werden.

Um zu gewährleisten, dass Clusterdeskriptoren, die innerhalb des erlaubten Maximalabstandes von δ_{DS}^{MAX} liegen, in jedem Fall eine Überschneidung ihrer Codebuchaktivierungsstruktur aufweisen und somit keine Rückweisung anhand der Codebuchaktivierung erfolgt, muss folgendes Kriterium erfüllt sein:

$$\tau_{AC}^{train} \geq (\delta_{DS}^{MAX} + \tau_{\rho}^{train}). \tag{5.12}$$

Hierbei ist τ_{AC}^{train} der Maximalabstand zwischen Deskriptor und Codebuchprototyp, bei dem der Aktivierungsstruktur des Deskriptors eine Aktivierung für einen bestimmten Codebucheintrag hinzugefügt wird. Dieser Abstand muss folglich größer sein als die Summe des Maximalabstands τ_{ρ}^{train}, der bei der Detektion (siehe Formel 3.17) für die Aktivierung von Codebucheinträgen verwendet wird, und des Maximalabstands für Deskriptoren bei der Wiedererkennung.

Somit wird verhindert, dass, wie in Abbildung 5.6 für den vereinfachten zweidimensionalen Fall dargstellt, eine Zuweisung von Deskriptoren, die innerhalb des Maximalabstands δ_{DS}^{MAX} liegen, durch fehlende Überschneidung in der Codebuchaktivierungsstruktur verhindert wird.

5.3 Ansichtsinvarianz der Wiedererkennung

An diesem Punkt soll auf die Thematik der Ansichtsinvarianz bei dem bisher vorgestellten Ansatz zur Personenwiedererkennung eingegangen werden. Dazu wird eine Vorgehensweise zur Generierung ansichtsspezifischer Identitätsmodelle vorgestellt. Im Weiteren wird die Ausnutzung der ansichtsspezifischen Modelle zur Wiedererkennung bei Vorliegen von unterschiedlichen Ansichten des Anfrage- und Datenbasismodells vorgestellt. Hierzu wird eine Technik zur Ansichtstransformation eines Identitätsmodells am Beispiel der Spiegelung vorgestellt.

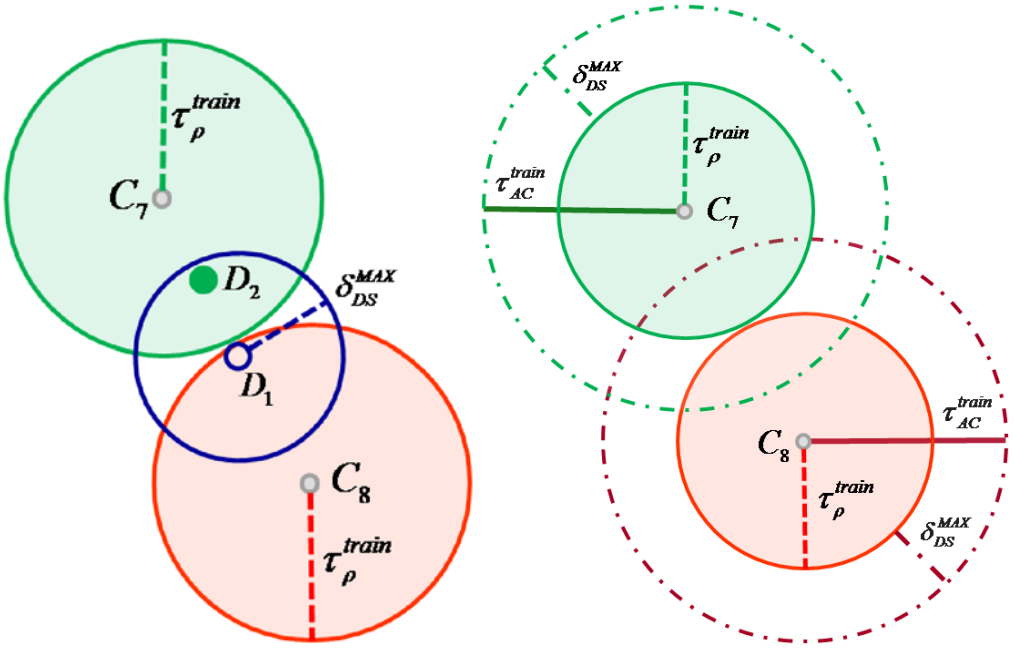


Abbildung 5.6: Darstellung der Abhängigkeiten zwischen den maximalen Distanzen für die Codebuchaktivierung bei Detektion und Wiedererkennung.

5.3.1 Generierung von ansichtsspezifischen Identitätsmodellen

Da während des Trackings einer Person unterschiedliche Ansichten einer Person vorkommen können – durch Bewegung der Person in der Szene oder bei Aktualisierung eines Modells aus einer anderen Kamera in Kameranetzwerken – sind die Identitätsmodelle einer Person nicht ansichtsspezifisch. Dies ist grundsätzlich gewünscht, da die Wiedererkennung nicht auf eine bestimmte Ansicht einer Person beschränkt, sondern über unterschiedliche Ansichten hinweg möglich sein soll.

Die Vermischung von Ansichten innerhalb eines Modells hat aber Nachteile. So wird die Distinktivität des Modells durch die Vermischung von Merkmalen aus unterschiedlichen Ansichten verringert. Insbesondere ist nicht bekannt, welche Merkmale aus welcher Ansicht der Person stammen. Somit ist es beim Vergleich mit einem anderen Modell nicht möglich, nur die Merkmale der passenden Ansicht zum Vergleich auszuwählen, sondern es erfolgt ein Vergleich mit allen Merkmalen. Dies ist nicht wünschenswert, da hierbei viele zufällige Übereinstimmungen aus unterschiedlichen Ansichten zu ungewollten Ähnlichkeiten zwischen Modellen unterschiedlicher Personen führen können.

Es ist also wünschenswert, unterschiedliche Modelle für verschiedene Ansichten aufzubauen um (i) die Distinktivität des Modells zu erhöhen und (ii) die Ansichtsinvarianz der Wiedererkennung durch Ausnutzung der bekannten Ansichten der wiederzuerkennenden Person und der Personen in der Datenbank zu erhöhen.

5.3.1.1 ISM-basierte Ansichtbestimmung

Zum Aufbau von ansichtsspezifischen Modellen ist es notwendig, die jeweils vorliegende Ansicht einer Person in der Szene zu kennen. Die Ansichtsbestimmung kann automatisch während des Trackings der Person durchgeführt werden.

Zur Bestimmung der aktuellen Ansicht einer getrackten Person wird hierzu das aktuelle Modell mit Trainingsmodellen in der Datenbasis verglichen. Die Trainingsmodelle werden offline erstellt, in dem nach dem in Abschnitt 5.1.1 beschriebenen Schema, Modelle von Personen aus unterschiedlichen Ansichten generiert werden. Die bekannten Ansichten der Trainingsmodelle werden dazu genutzt, ein allgemeines Modell für jede Ansicht zu erstellen. Hierzu werden die Trainingsmodelle der gleichen Ansicht, wie in Abschnitt 5.1.1 vorgestellt, zuammengeführt.

Der Modellvergleich zur Ansichtsbestimmung wird nach dem gleichem Schema wie die Wiedererkennung durchgeführt (siehe Abschnitt 5.2). Hierbei wird zur Ansichtenbestimmung allerdings das Kurzzeitmodell der aktuell getrackten Person genutzt, da dieses nicht die gesamte Historie, sondern nur die Historie in einem kurzen Zeitfenster enthält. Die Ansichtsklassifikation auf diesem Weg ist möglich, da die nach dem hier beschriebenen Schema aufgebauten Modelle nicht nur die Identitätsinformation, sondern in großem Umfang auch die Ansichtsinformation enthalten.

Tests dieser Methode zur Ansichtenerkennung auf dem CASIA A Datensatz[1] [1] zeigen sehr gute Performance. Hier reichen die von vier unterschiedlichen Personen generierten Trainingsmodelle aus, um 100% korrekte Ansichtsklassifikation zu erreichen. Die Ansichtsklassifikation bezieht sich hierbei auf die Auswahl einer Ansicht aus einer diskreten Menge von 6 Trainingsansichten.

Die hier vorgestellte Methodik wird auch genutzt, um die in Abschnitt 5.1.2 beschriebene Merkmalsselektion zu verbessern. Dazu wird zur Merkmalsauswahl nicht mehr ein einzelnes Modell für alle Ansichten verwendet, sondern es werden ansichtsspezifische Modelle generiert, die dann zur Merkmalsselektion genutzt werden. So ist es möglich, allgemeine, ansichtsspezifische Merkmale auszufiltern, die bei der Wiedererkennung zwischen Ansichten kontraproduktiv wären. In der Praxis kann das gleiche allgemeine, aber ansichtsspezifische Modell zur Ansichtsklassifikation und zur Merkmalsselektion genutzt werden.

5.3.1.2 Ablage der ansichtsspezifischen Modelle

Beim Aufbau des Identitätsmodells einer Person während des Trackings wird wie in Abbildung 5.7 zu sehen ist, die Ansicht der getrackten Person nach dem in Abschnitt 5.3.1.1 vorgestellten Schema bestimmt. Hierbei ist abhängig von den vorhandenen Trainingsdaten eine Klassifikation der aktuellen Ansicht in eine diskrete Menge von Ansichten möglich. In dem in Abbildung 5.7 gezeigten Beispiel sind dies 8 verschiedene Ansichten die eine Klassifikation in Abstufungen von 45° erlauben. Für jede der 8 Ansichten kann nun ein unabhängiges Identitätsmodell generiert werden. Dabei wird die Ansichtsklassifikation zur Auswahl des Modells, in dem die Bildmerkmale abgelegt werden, genutzt.

Zur Ablage der ansichtsspezifischen Modelle muss für jeden Zeitpunkt, d.h. jedes Bild einer Bildsequenz, während des Trackings einer Person eine Ansichtsbestimmung erfolgen. Sofern eindeutig eine Ansicht bestimmt werden kann, werden die Merkmale, welche die Kriterien nach Abschnitt 5.1.1 erfüllen, in dem Modell der determinierten Ansicht abgelegt. Kann zu einem Zeitpunkt keine eindeutige Ansichtsklassifikation erfolgen, werden die Merkmale in keinem Modell abgelegt. Alternativ zu diesem Vorgehen könnte auch ein weiteres Modell für eine Person eingeführt werden, das die Merkmale nicht klassifizierter Ansichten enthält. Eine andere Möglichkeit wäre auch die Ablage im Modell der letzten bekannten Ansicht oder eine Interpolation der Ansicht bei nicht klassifizierten Ansichten.

Durch diesen Schritt liegen nun potentiell mehrere ansichtsspezifische Modelle einer Person vor. Da beim Tracking einer Person typischerweise nicht alle Ansichten abgedeckt sind, ist es möglich, dass bei der Wiedererkennung die Ansicht des Anfragemodells nicht bei allen Modellen in der Datenbasis enthalten ist. In diesem Fall muss bei der Wiedererkennung also die beste Vergleichsansicht ausgewählt werden.

[1] Dieser Datensatz wird auch zur Auswertung der Ansichtsinvarianz der Personenwiedererkennung in Abschnitt 5.4.2 verwendet.

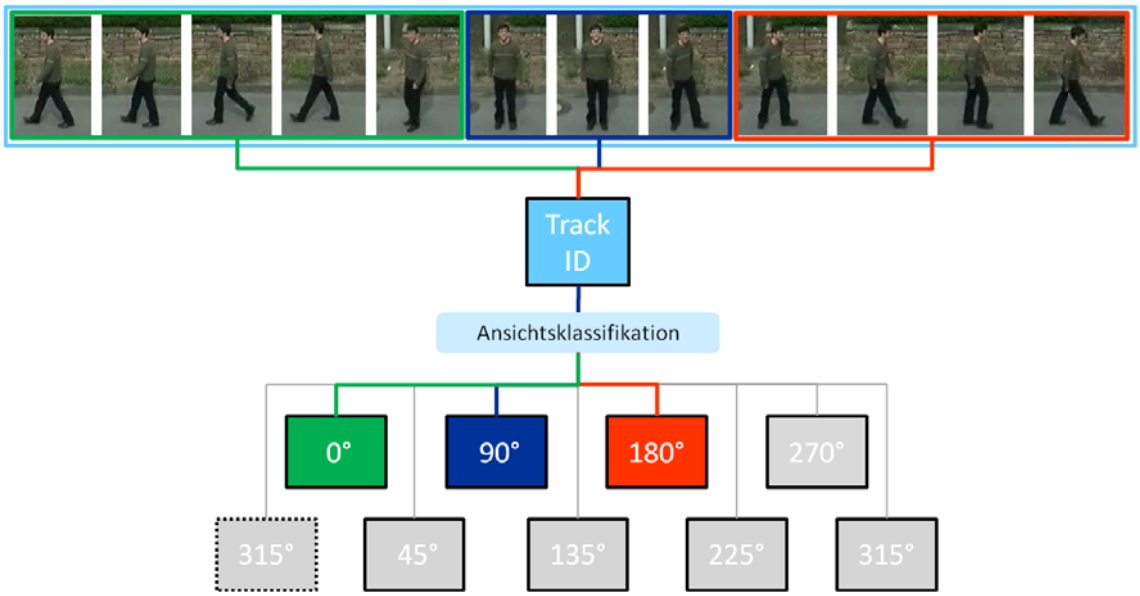


Abbildung 5.7: Ablage von ansichtsspezifischen Modellen anhand des Beispiels „Gehrichtungsveränderung während des Trackings“.

5.3.2 Ansichtenauswahl zum Modellvergleich

Durch die Ansichtsbestimmung wird es möglich zu jedem Zeitpunkt während des Trackings einer Person, die der Kamera zugewandte Seite der Person, und damit die aktuelle Ansicht der Person in der Kamera zu bestimmen. Diese Information kann bei der Personenwiedererkennung genutzt werden, indem für jede Ansicht separat ein Identitätsmodell aufgebaut wird. Durch die vollautomatische Ansichtsbestimmung kann dies online erfolgen.

Bei Identitätsbestimmung der aktuell getrackten Person durch Vergleich mit den Personenmodellen in der Datenbasis kann nun für den Vergleich die passende Ansicht der zu vergleichenden Person ausgewählt werden. Hierdurch verringert sich nicht nur die Menge der zu vergleichenden Merkmale, insbesondere erhöht sich auch die Distinktivität des Modells.

Die Auswahl der passenden Ansicht der Person in der Datenbank kann allerdings nicht immer direkt erfolgen, da in der Datenbasis nicht immer alle Ansichten der Person vorliegen. In realen Überwachungssituationen (z.B. an Bahnhöfen) ist es eher so, dass sich eine Person gezielt durch den Sichtbereich der Kamera bewegt und somit immer von einer bestimmten Ansicht sichtbar ist, die zwischen den unterschiedlichen Kameras in einem Kameranetz aber natürlich variieren kann. Somit ist es nicht immer möglich, die korrekte Vergleichsansicht direkt auszuwählen. Um den Vergleich auch in diesen Fällen zu ermöglichen, wird jeweils die Ansicht ausgewählt, die der aktuellen Ansicht der Anfrageperson am nächsten ist und somit den akkuratesten Vergleich zulässt.

Wie in Abbildung 5.8 zu sehen ist, wird hierzu jeweils das vorhandene Modell ausgewählt, das die geringste Winkeldifferenz zum Anfragemodell aufweist.

Zusätzlich zur Auswahl der besten Vergleichsansicht kann die vorliegende Information über die Ansicht der Person auch zur Parametrisierung der Personenwiedererkennung genutzt werden. So ist insbesondere der Parameter δ_S^{MAX}, der den Maximalabstand zwischen den Ortsverteilungen angibt, an dieser Stelle relevant. Dieser kann, sofern Anfrage- und Datenbasismodell in der gleichen Ansicht vorliegen, niedriger gewählt werden als bei Vorliegen der Modelle in unterschiedlichen Ansichten.

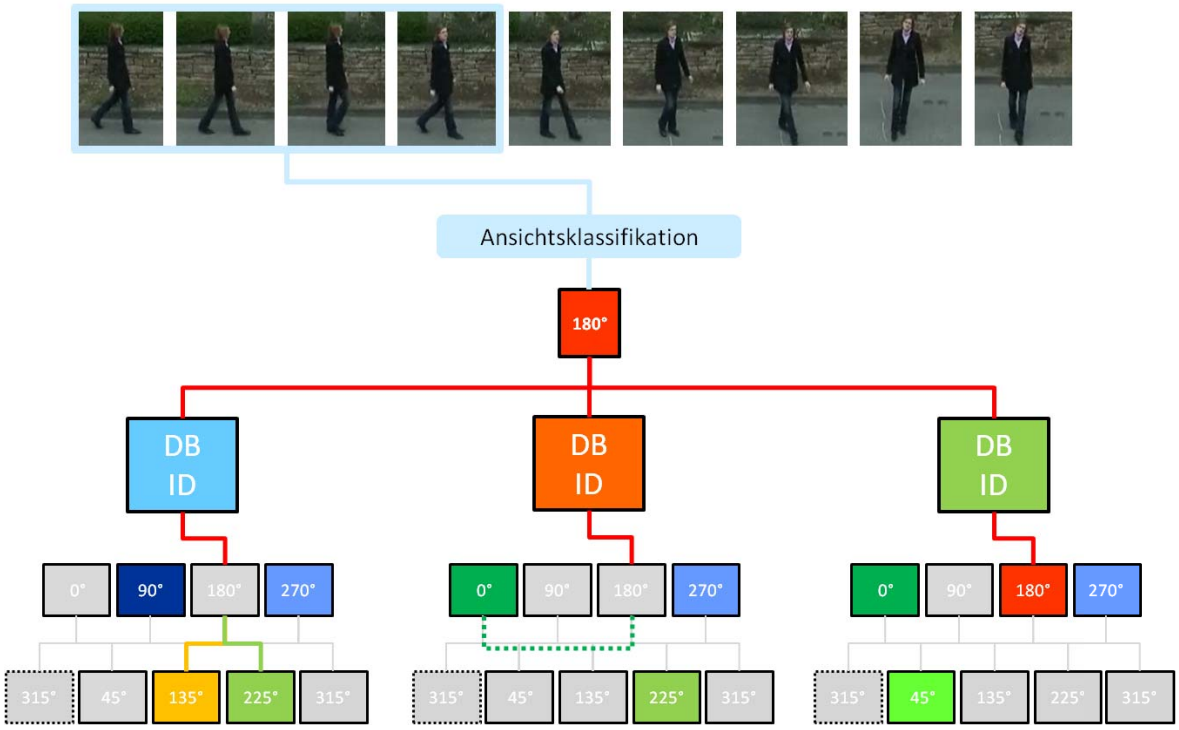


Abbildung 5.8: Ansichtsauswahl für den Modellvergleich.

5.3.3 Ansichtsinvarianz beim Modellvergleich

Die in Abschnitt 5.3.2 beschriebene Methodik verbessert zwar die Distinktivität der Modelle bei der Wiedererkennung, ist für sich genommen aber noch nicht ausreichend, um eine weitgehende Ansichtsinvarianz der Wiedererkennung zu gewährleisten. Dies ist der Fall, da in der Praxis nur selten alle Ansichten einer Person für ein Modell in der Datenbasis vorliegen. Dies kann dazu führen, dass ein Anfragemodell mit dem Modell einer stark abweichenden Ansicht verglichen werden muss, die aufgrund der starken Abweichung keinen korrekten Vergleich mit der Anfrageansicht zulässt. Dies ist in der Praxis allerdings nicht immer zu vermeiden, da in vielen Fällen nur stark abweichende Ansichten, zwischen denen keine Verbindung hergestellt werden kann, vorliegen. Ein solcher Fall liegt z.B. vor, wenn eine Person in der Anfrage frontal von vorne sichtbar ist, in der Datenbasis aber nur aus einer Seitenansicht vorliegt. Zwischen diesen Ansichten kann kaum eine Verbindung hergestellt werden, da es bei Menschen keine visuelle Korrelation zwischen der Front- und Seitenansicht geben muss.

In anderen Fällen ist allerdings denkbar, dass eine visuelle Verbindungen zwischen den Ansichten besteht, diese Ähnlichkeit aber durch die Invarianzen der zur Personenbeschreibung gewählten Merkmale nicht abgedeckt ist. Im hier konkret vorliegenden Fall der Merkmalsbeschreibung durch SIFT sind durch Rotation ineinander überführbare Ansichten ein Beispiel dafür. Da die Rotationsinvarianz des SIFT-Deskriptors im Kontext dieser Arbeit, aus den in Abschnitt 3.1.1.3 beschriebenen Gründen nicht genutzt wird, könnte die Lösung hierfür sein, entweder bei Erstellung der Merkmalsmodelle die Rotationsinvarianz zu nutzen, was aber die Distinktivität einschränken würde, oder aus den bekannten Ansichten der beiden Modelle gezielt die vorliegende Rotation zwischen den Ansichten auf die SIFT-Deskriptoren einer Ansicht anzuwenden, um die Modelle ineinander zu überführen. Dieses Beispiel einer Transformation zwischen Merkmalsmodellen wäre auf Basis der SIFT-Deskriptoren leicht durchzuführen, hat aber keine praktische Relevanz.

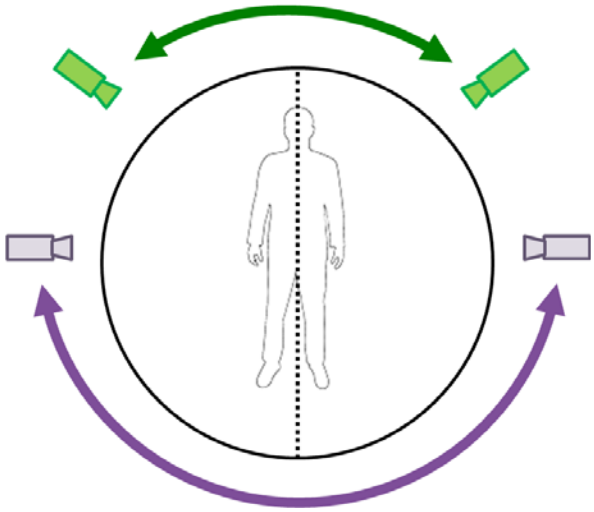


Abbildung 5.9: Durch die Symmetrie von Personen ist die eine Hälfte der Sichtsphäre direkt in die andere überführbar.

Ein Beispiel für eine in der Praxis in Überwachungsszenarien häufig vorkommende Transformation ist die Spiegelung. Diese kommt immer dann vor, wenn sich eine Person einmal von links nach rechts durch das Sichtfeld der Kamera bewegt, ein anderes Mal von rechts nach links[2]. Gegenüber Spiegelungen ist der SIFT-Deskriptor nicht invariant, hierzu müsste also eine explizite Transformation eingeführt werden.

Zunächst soll aber die Rechtfertigung einer solchen Transformation betrachtet werden. Diese ergibt natürlich nur Sinn, wenn sie mit den realen Gegebenheiten bei Personen übereinstimmt, sprich, wenn eine so geartete Symmetrie vorliegt.

Dies ist bei Personen typischerweise der Fall, da eine typische Person, wie in Abbildung 5.9 dargestellt, achsensymmetrisch für eine Spiegelung an der z-Achse ist. Dies wird in der Realität auch durch Kleidung nur selten aufgehoben, so dass diese Symmetrie ohne Einschränkung der Allgemeingültigkeit für Personen angenommen werden kann. Für die Ansichtsinvarianz der Personenwiedererkennung bedeutet dies, dass wie in Abbildung 5.9 dargestellt, bereits die Hälfte der möglichen Ansichten durch Spiegelung ineinander überführt werden können. Die Hälfte der Sichtsphäre ist also bereits durch eine einzige Transformation abgedeckt. Da diese Transformation in der Praxis die größte Relevanz hat, soll sie hier als Beispieltransformation durchgeführt, und auch in den Experimenten zur Personenwiedererkennung genutzt werden.

5.3.4 Spiegelung des Identitätsmodells

Die Spiegelung von SIFT-Merkmalen kann direkt auf Ebene des Deskriptors vorgenommen werden. Es ist also nicht notwendig, den der Deskriptorberechnung zugrunde liegenden Bildausschnitt zu transformieren und den Deskriptor unter Nutzung dieses transformierten Bildausschnitts neu zu berechnen. Dies ist auch nicht wünschenswert, da die Neuberechnung des SIFT-Deskriptors rechenintensiv ist und insbesondere zusätzlicher Speicherbedarf für die Ablage der jeweiligen Bildausschnitte im Modell erforderlich wäre.

Tatsächlich ist es zum Vergleich zweier durch Spiegelung ineinander überführbarer Modelle nicht

[2]Dieser Fall liegt auch häufig durch entgegengesetzte Positionierung der Kameras in Kameranetzen vor.

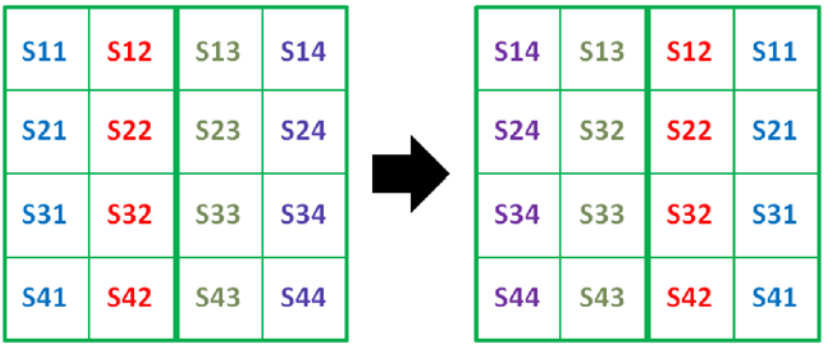


Abbildung 5.10: Spiegelung des SIFT-Deskriptors: Ortshistogramm.

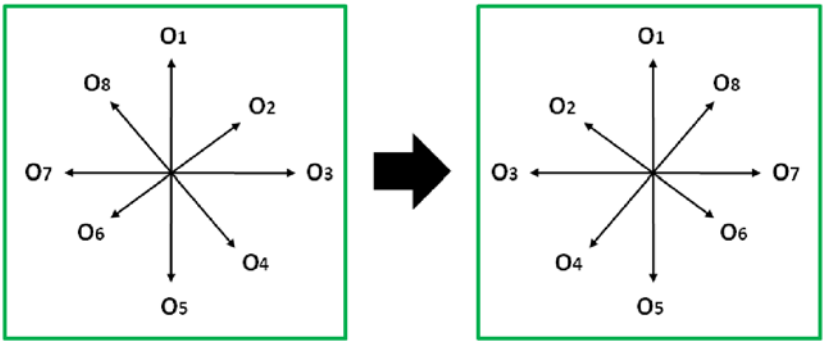


Abbildung 5.11: Spiegelung des SIFT-Deskriptors: Orientierungshistogramm.

notwendig, überhaupt explizit eine Transformation durchzuführen. Diese kann direkt beim Deskriptorvergleich, und damit ohne Geschwindigkeitsverlust durchgeführt werden [118]. Wie in Abbildung 5.10 und 5.11 zu sehen ist, ist es dazu lediglich notwendig, das Ortshistogramm (Abbildung 5.10) des Deskriptors an der y-Achse, sowie das Orientierungshistogramm (Abbildung 5.11) an der durch die Orientierungen 0° und 180° definierten Achse zu spiegeln. Für das Ortshistogramm wird die Spiegelung dabei wie folgt durchgeführt:

$$S'(i,j) = S(i,5-j). \tag{5.13}$$

Zusammen mit der Spiegelung der Orientierungshistogramme innerhalb des Ortshistogramms ergibt sich somit:

$$\begin{aligned} H'(i,j) = [&O((i,5-j),1), O((i,5-j),8), O((i,5-j),7), O((i,5-j),6), \\ &O((i,5-j),5), O((i,5-j),4), O((i,5-j),3), O((i,5-j),2)]. \end{aligned} \tag{5.14}$$

Die Spiegelung kann durch Vertauschung der Einträge also direkt auf dem 128-dimensionalen Deskriptorvektor während des Deskriptorvergleichs durchgeführt werden. Die Vertauschung kann pro Eintrag durch einen einzelnen „Hash-Table-Lookup“ erfolgen und erhöht den Rechenaufwand daher kaum.

Zur Spiegelung des Identitätsmodells auf Ebene des ISM muss lediglich die x-Komponente des Zentrumsversatzes $\epsilon_{x,y}$ invertiert werden:

$$\hat{\epsilon}_{x,y} = \epsilon_{-x,y}. \tag{5.15}$$

Da für den Modellvergleich auf allen Stufen die Codebuchaktivierungen der Merkmale genutzt werden, ist es notwendig, auch diese zu transformieren. Dazu müssen die Deskriptoren unter Nutzung

der Spiegelungstransformation mit dem Codebuch verglichen werden. Dies ist zusätzlicher Vergleichsaufwand, der zukünftig durch eine mögliche Anwendung der Spiegelung auf das gesamte Codebuch vermieden werden kann. Hierbei könnte für jeden Prototypen einmalig eine Menge an Prototypen, deren Spiegelung in einem bestimmten Ähnlichkeitsradius zu dem Prototypen liegt, bestimmt werden. Diese einmalig bestimmte Menge an Prototypen könnte bei Aktivierung dieses Prototypen durch ein Bildmerkmal dann direkt ohne zusätzliche Vergleichsoperationen als Spiegelungsaktivierung für dieses Merkmal genutzt werden. Eine weitere Möglichkeit wäre auch die in [155] vorgestellte Nutzung von unterschiedlichen Codebüchern und die Integration dieser in ein „Multi-View“ Modell.

Die hier für die Objektklasse Person gemachte Symmetrieannahme gilt nicht nur für Personen, sondern für viele Objektklassen – insbesondere für sogenannte „man-made-objects“[3]. Beispiele für in diesem Kontext relevante Objektklassen auf die dies zutrifft sind Autos oder Schiffe. Tatsächlich trifft dies aber auf die meisten im Alltag gebräuchlichen Objekte zu. Die allgemeine Anwendbarkeit des hier vorgestellten Ansatzes wird somit auch durch diese Annahme nicht wesentlich eingeschränkt.

Durch die vorgestellte Spiegelung der Modelle kann eine Ansichtsinvarianz der Modelle für die jeweils gegenüberliegende Halbkugel in der Sichtsphäre erreicht werden. Die verbleibenden Ansichten liegen in der Sichtsphäre näher zusammen, d.h. die aufgezeichneten Bilder unterliegen nicht so starken Transformationen. Für den konkreten Fall des SIFT-Deskriptors und des zur Wiedererkennung genutzten ISM bedeutet dies, dass nicht notwendigerweise eine explizite Berechnung der Transformationen notwendig wird, da in der Personenbeschreibung bereits Ansichtsinvarianz bis zu einem gewissen Grad enthalten ist.

Beim ISM ist die Ansichtsspezifität hauptsächlich durch die Zentrumsversätze der Merkmale gegeben. An dieser Stelle kann die Invarianz durch Aufweichung des Abstandskriteriums für die Ortsverteilungen erhöht werden. Hierbei ist zu beachten, dass bei Zulassen einer größeren Abweichung δ_S^{MAX} der Ortsverteilung natürlich auch Distinktivität des Modells verloren geht.

Beim SIFT-Deskriptor ist durch den Deskriptor selbst auch eine gewisse Invarianz gegenüber Transformationen, selbst affinen Transformationen gegeben. Lowe [117] gibt hierzu an, dass bei 50° Ansichtsveränderungen noch immer für mehr als 50% der Merkmale akkurate Korrespondenzen gefunden werden. Dies würde bedeuten, dass eine Personenwiedererkennung auch bei Veränderung des Winkels um bis zu 50° ohne explizite Transformation möglich sein sollte. Hierzu sei auch auf die experimentelle Validierung auf dem Datensatz CASIA A in Abschnitt 5.4.2, bei der die Ansichtsinvarianz der Personenwiedererkennung ausgewertet wird und die diese Annahmen unterstützt, verwiesen.

Die Erhöhung, insbesondere der Invarianz des Deskriptors gegenüber affinen Ansichtstransformationen kann aber auch durch explizite Einführung von Invarianz gegenüber solchen Transformationen erfolgen. Dazu könnten die in [126] beschriebenen Techniken zur Bestimmung und Normalisierung der lokalen affinen Transformation genutzt werden.

5.4 Auswertung

Die experimentelle Validierung der Personenwiedererkennung erfolgt in zwei Szenarien, die beide mit typischen Überwachungsszenarien vergleichbar sind, in denen keine biometrischen Merkmale wie Gesicht oder Iris nutzbar sind.

Wie Eingangs dieses Kapitels ausgeführt, hat eine Wiedererkennung von Personen unterschiedliche Anwendungsfälle. Die unterschiedlichen Anwendungsfälle sind dabei von unterschiedlichem Grad hinsichtlich der Herausforderung für die Personenwiedererkennung.

Ein möglicher Anwendungsfall ist z.B. die Unterstützung eines Operators bei der Suche nach einer bestimmten Person. So könnte anhand einer Beispielsequenz einer Person nach allen Auftreten dieser Person in der Datenbasis – die abgelegte Personenmodelle aus der Vergangenheit oder anderer Kameras

[3]Keine in der Natur vorkommenden Objekte, sondern von Menschen produzierte Gegenstände.

enthält – gesucht werden. Da in diesem typischen Anwendungsfall die Wiedererkennung lediglich zur Unterstützung eines menschlichen Operators dient, geht es hier lediglich um Arbeitserleichterung. Es reicht also aus, wenn die gesuchte Person unter einer Menge von N vorgeschlagenen Personen aus der Datenbasis ist.

Ein weitaus schwierigerer Fall liegt vor, wenn die Wiedererkennung Teil eines vollautomatischen Systems ist, bei dem kein menschlicher Eingriff erwünscht ist. Dies kann z.B. in Überwachungssystemen der Fall sein, bei dem eindeutig bestimmt werden muss, ob und wenn ja wann eine gerade in der Szene sichtbare Person bereits vorher in der Szene zu sehen war. Dies kann notwendig sein, wenn z.B. auffälliges Verhalten automatisch detektiert, die Zuordnung eines Gegenstands, z.B. einer abgestellten Tasche, zu einer Person, die den Sichtbereich der Kamera nach abstellen des Gegenstands kurzzeitig verlassen hat, determiniert, oder ein Verhaltensbild von Personen erstellt werden soll. In diesem Fall reicht es nicht aus, wenn die Anfrage an die Datenbasis nach der gerade in der Szene sichtbaren Person 10 Modelle ergibt, in denen die Zielperson enthalten ist, sondern es ist eine eindeutige Zuordnung notwendig.

In realen Anwendungen ergeben sich über diese Anforderung hinaus noch weitere Herausforderungen. So ist es bei Beobachtung einer Person in der überwachten Szene nicht notwendigerweise der Fall, dass diese Person bereits vorher gesehen wurde und somit in der Datenbasis vorhanden ist. Das Wiedererkennungsverfahren muss also auch entscheiden, ob eine Person überhaupt in der Datenbasis vorhanden ist, oder ob diese Person unbekannt ist. Dies ist eine typische *open-set* Klassifikationsaufgabe. Für das Wiedererkennungsverfahren ist diese Problemstellung weitaus schwieriger zu bewältigen als die reine Auswahl des am Besten passenden Modells aus der Datenbasis, unter der Annahme, dass die Person in der Datenbasis enthalten ist. Es reicht also nicht aus, einen Klassifikator zu finden, der imstande ist die richtigen Personen aus der Datenbasis auszuwählen. Zudem muss auch noch sichergestellt sein, dass der Klassifikator geeignet ist, beliebige unbekannte Personen als solche zu identifizieren.

Dieser, für die Personenwiedererkennung herausfordernsten Aufgabe widmet sich Abschnitt 5.4.1. Hier wird eine Personenwiedererkennung unter diesen Anforderungen auf dem im infraroten Spektralbereich aufgezeichneten CASIA C Datensatz [1] vorgenommen. Neben der Validierung des Ansatzes im infraroten Spektralbereich wird hier insbesondere eine Auswertung der verschiedenen Klassifikationsstufen vorgenommen. Da in der Literatur keine Verfahren zur Personenwiedererkennung im infraroten Spektralbereich existieren, erfolgt ein Vergleich der Performance mit dem Gangerkennungsverfahren aus [153].

Eine weitere Auswertung der Wiedererkennung, hier auf im sichtbaren Spektralbereich aufgezeichneten Daten, erfolgt in Abschnitt 5.4.2. Hier soll neben der Validierung der Wiedererkennung im sichtbaren Spektralbereich der Fokus insbesondere auf der Bewertung der Ansichtsunabhängigkeit des Modells liegen, wobei die in Abschnitt 5.3 vorgestellten Methoden zur Ansichtsbestimmung und -auswahl sowie die Spiegelung des ISM genutzt wird. Hier erfolgt ein Vergleich mit dem Verfahren aus [162], bei dem der gleiche Datensatz genutzt wurde.

Beide Auswertungen bauen auf dem in Kapitel 4 vorgestellten System zur Personenverfolgung auf. Es erfolgt also keinerlei manuelle Annotation oder Eingriff in das System, wie es in vielen anderen Wiedererkennungsansätzen (siehe Kapitel 6), die auf manuell annotierten Daten aufbauen, der Fall ist. Die Auswertung erfolgt mit den während der Personenverfolgung online aufgebauten Identitätsmodellen. Insbesondere wird für beide Auswertungen der gleiche Wiedererkennungsansatz verwendet, was die Szenario- und Sensorunabhängigkeit des Gesamtansatzes demonstriert.

Die Auswertung erfolgt dabei nicht mit festen Zeitintervallen, wie sie sonst bei „künstlichen“ Wiedererkennungsauswertungen genutzt werden, sondern unter realen Bedingungen. D.h. alle Unwägbarkeiten und Ungenauigkeiten die aus dem Tracking in der Realität hervorgehen, gehen auch in die Wiedererkennung ein.

Die beiden Auswertungen zeigen, dass Vergleichsverfahren aus dem Gangerkennungsbereich [153, 162], die auf den gleichen Datensätzen getestet wurden, in der Performance weit übertroffen werden.

5.4.1 Infraroter Spektralbereich

Der CASIA C Datensatz [1] enthält Bildsequenzen von einzelnen Personen, die den Sichtbereich der Kamera von links nach rechts, orthogonal zur Kamera durchqueren. Die Bildsequenzen wurden im infraroten Spektralbereich mit einer Auflösung von 320x240 mit 25 fps[4] aufgezeichnet.

Zur Auswertung wird eine Datenbasis mit 50 Personen aufgebaut[5]. Dies erfolgt durch Tracking der Personen in jeweils einer Bildsequenz. Pro Person wird jeweils ein Modell in der Datenbasis abgelegt. Zur Wiedererkennung werden die gleichen 50 Personen in einer zweiten Bildsequenz verfolgt und online mit den Modellen in der Datenbasis verglichen. Somit ergibt sich für jeden Zeitpunkt während des Trackings eine Ähnlichkeit der getrackten Person mit den Modellen in der Datenbasis und damit für jeden Zeitpunkt eine Klassifikation als eine der Personen in der Datenbasis oder eine Rückweisung des Modells. Eine solche Rückweisung kann erfolgen, da die Auswertung nach dem *open-set* Schema erfolgt. D.h. zusätzlich zu den 50 Testpersonen in der Datenbasis werden 10 unbekannte Testpersonen nach dem gleichen Schema verfolgt und mit der Datenbasis verglichen. Das System kann also nicht sicher sein, dass alle Testpersonen in der Datenbasis enthalten sind. Somit kann nicht einfach das Modell mit der höchsten Ähnlichkeit aus der Datenbasis ausgewählt werden, sondern es muss auch entschieden werden, ob diese Ähnlichkeit zur eindeutigen Klassifikation ausreicht, oder ob eine Rückweisung der Person als unbekannt erfolgt.

Beispielbilder der Personen dieses Szenarios sind in Abbildung 5.12 dargestellt.

5.4.1.1 Auswertungsmaße

Zur Auswertung werden folgende Maße genutzt, die sich aus der Aufgabendefinition ableiten lassen und ihren Ursprung in biometrischen Auswertung haben:

Die *False Rejection Rate (FRR)* ist die Rate von Personen die zwar in der Datenbasis vorhanden sind, die aber fälschlicherweise vom System als unbekannt zurückgewiesen wurden:

$$FRR = \frac{\#\text{Falschrückweisungen}}{\#\text{Personen in DB}}. \tag{5.16}$$

Die *False Acceptance Rate (FAR)* ist die Rate von Personen, die als bestimmte Person in der Datenbasis klassifiziert wurden, die aber tatsächlich nicht in der Datenbasis vorhanden sind, also hätten als unbekannt klassifiziert werden müssen:

$$FAR = \frac{\#\text{Falschangenommene}}{\#\text{Personen in DB}}. \tag{5.17}$$

Die *Misclassification Rate (MCR)* ist die Rate von Personen, die als falsche Person klassifiziert wurden:

$$MCR = \frac{\#\text{Falschklassifikationen}}{\#\text{Personen in DB}}. \tag{5.18}$$

Als gemeinsames Maß für die Korrektheit der Wiedererkennung wird die *Correct Classification Rate (CCR)*, welche die MCR und FRR einbezieht, genutzt:

$$CCR = 1.0 - MCR - FRR \tag{5.19}$$

$$= \frac{\#\text{Korrekte Klassifikationen}}{\#\text{Personen in DB}}. \tag{5.20}$$

[4]Frames per second (Bilder pro Sekunde)
[5]Dies ist eine Erweiterung um 35 Personen im Vergleich zu Jüngling und Arens [89]

Abbildung 5.12: Beispielbilder des CASIA C Testdatensatzes

Zur Klassifikation wird eine Kombination von einem Schwellwert θ für das Verhältnis der Ähnlichkeitswerte des ähnlichsten und des zweitähnlichsten Modells in der Datenbasis und des absoluten Ähnlichkeitswerts gewählt. Hierbei wird die zeitliche Entwicklung von θ und der Maximalwert der absoluten Ähnlichkeit betrachtet. Voraussetzung für die Klassifikation als eine bestimmte Person ist, dass die Eingabeperson mindestens 50% der getrackten Zeit als eine bestimmte Person mit $\theta \geq 1.4$ klassifiziert werden muss. Dieser Zeitraum muss zusammenhängend sein. Das zusätzliche Kriterium für einen Mindest-Ähnlichkeitswert wird eingeführt, um zu verhindern, dass unbekannte Personen, die nur sehr geringe Ähnlichkeit zu einer bestimmten Person aufweisen (z.B. durch eine einzelne Merkmalskorrespondenz) und keine zu anderen (also Ähnlichkeit 0) falsch als bekannte Person klassifiziert werden.

5.4.1.2 Auswertung

In der Praxis dienen Stufe 1 und 2 lediglich dem Ausschluss von unähnlichen Modellen in der Datenbasis zur Reduktion des Vergleichsaufwands auf Stufe 3, auf der die eigentliche Klassifikation vorgenommen wird. Um aber auch Stufe 1 und 2 bzgl. ihrer Klassifikationsfähigkeiten zu bewerten, wird in den hier vorgestellten Experimenten auch eine Klassifikation auf dieser Stufe vorgenommen.

Stufe 1: Codebuchsignatur Wie in Tabelle 5.1 zu sehen ist, werden bei Klassifikation auf Stufe 1, alleine durch die niedrigdimensionalen Codebuchsignaturen bereits sehr gute Wiedererkennungsraten von über 73% bei Einzelbildauswertung und 86% bei Auswertung der gesamten Sequenz erreicht. Allerdings wurde hier keine open-set Klassifikation durchgeführt. Wichtig für die Auswertung auf dieser Stufe ist außerdem, dass die Auswertung ergeben hat, dass sich das korrekte Modell jederzeit unter den Besten fünf Datenbasismodellen befindet. Somit können auf diesem Weg also für N=5 in der ersten Stufe, sofern diese nicht direkt als Klassifikationsstufe dienen soll, 90% aller Modelle ausgeschlossen werden. Für eine Datenbasis von 50 Personen muss der aufwändigere Deskriptorvergleich also nur für jeweils 5 Modelle durchgeführt werden. Dies ist ein immenser Anteil, insbesondere wenn man bedenkt, dass in der ersten Stufe lediglich N-dimensionale (mit Codebuchdimension N = 216 in diesem Experiment) Vektoren miteinander verglichen werden müssen.

Tabelle 5.1: Personenwiedererkennung CASIA C. Stufe 1: Codebuchsignaturen.

	FAR	**FRR**	**MCR**	**CCR**
Zeitpunkt	-	-	26.9	73.1
Sequenz	-	-	14.0	86.0

Stufe 2: ISM-Aktivierung Tabelle 5.2 zeigt die Ergebnisse der Auswertung auf Stufe 2. Wie zu sehen ist, verbessern sich die CCR-Raten um 3% auf 76% bzw. um 6% auf 92%. Dies ist der Hinzunahme der Zentrumsversätze auf dieser Stufe geschuldet, die die Distinktivität der Modelle erhöht. Gleichzeitig ist auch hier das korrekte Modell immer unter den besten 5 Modellen.

Tabelle 5.2: Personenwiedererkennung CASIA C. Stufe 2: ISM-Aktivierung.

	FAR	**FRR**	**MCR**	**CCR**
Zeitpunkt	-	-	24.0	76.0
Sequenz	-	-	8.0	92.0

Stufe 3: Merkmalsdeskriptoren Klassifikationsergebnisse für Stufe 3 sind in Abbildung 5.13 dargestellt. Hier wurde die in der Einführung beschriebene *open-set* Auswertung vorgenommen. Dargestellt sind zwei unterschiedliche Klassifikationsmethoden für Stufe 3. EDA ((E)inzelzeitpunkt (D)eskriptor A) und SDA ((S)equenz (D)eskriptor A) führen die Wiedererkennung jeweils ausschließlich auf Basis der bei der Detektion über das Codebuch aktivierten Merkmale durch. EDB und SDB nutzen zur Klassifikation zusätzlich die in Abschnitt 5.1.2 beschriebenen „zusätzlichen Merkmale“. Die Klassifikation findet auf Basis eines einzelnen Zeitpunkts (E) (ausgewertet wurden alle Einzelzeitpunkte der Sequenz) und auf Basis der gesamten Sequenz (S) statt. Für einzelne Zeitpunkte wurde die Auswertung jeweils für $\theta = 1.2$ und $\theta = 1.4$ durchgeführt.

Wie die Graphen zeigen, wird bei Einzelklassifikation die bessere Performance bei einem Klassifikationsschwellwert von 1.4 erreicht. Die CCR ist hierbei ähnlich wie bei 1.2, allerdings nimmt die

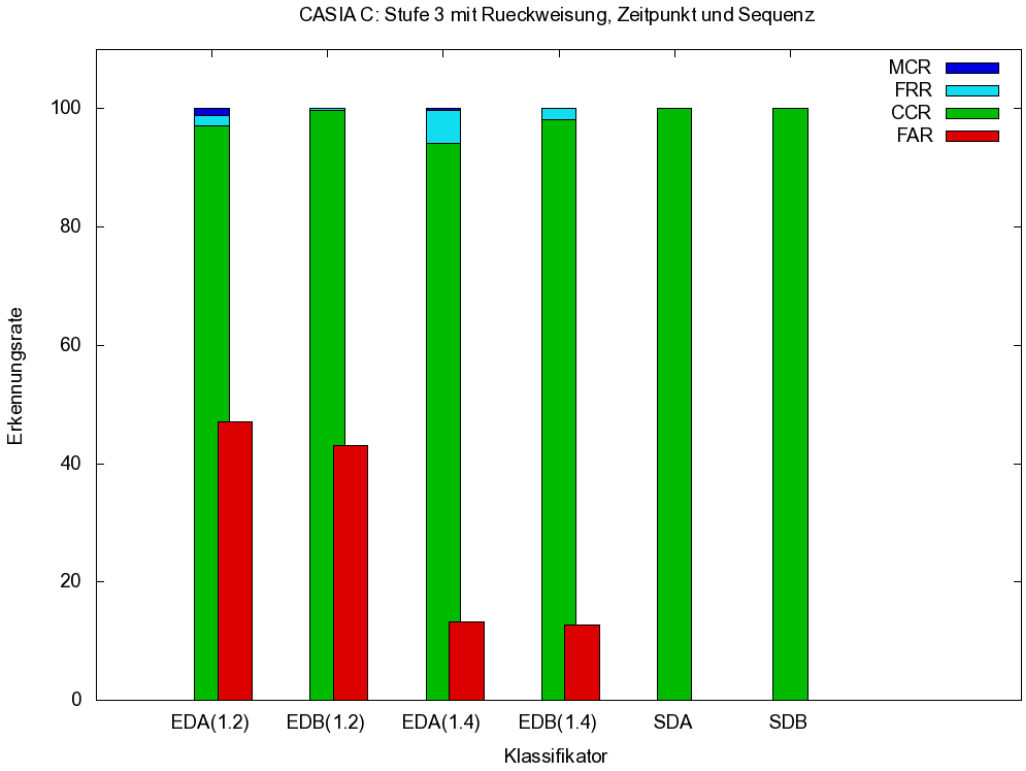


Abbildung 5.13: Klassifikationsraten mit Rückweisung auf Stufe 3 für den CASIA C Datensatz. Dargestellt sind die Erkennungsraten für unterschiedliche Klassifikationsarten. DA: Klassifikation auf Basis von Trackingmerkmalen, DB: Klassifikation auf Basis von Tracking- und zusätzlichen Merkmalen. Jeweils für (E)inzelzeitpunkt und (S)equenz. Der Klassifikationsschwellwert θ ist in Klammern angegeben.

Falschakzeptanzrate (FAR) signifikant um über 20% ab. Wie zu sehen ist, ist die Klassifikationsverbesserung von EDA zu EDB, also durch Hinzunahme der zusätzlichen Merkmale, nur marginal. Dies ist der Fall, da hier wenig bis keine spezifische Textur auf den Personen zu sehen ist und die Klassifikation fast ausschließlich auf Basis der Formbeschreibung erfolgt. Hierzu reichen die Merkmale, die während des Trackings gesammelt werden, aus, um sehr gute Performance zu erreichen.

Die Situation ist anders, wenn Daten aus dem sichtbaren Spektralbereich vorliegen und wenn es starke Ansichtsveränderungen bei der Wiedererkennung gibt. In diesem Fall sind nicht nur die Formmerkmale relevant, sondern insbesondere die auf dem Personenabbild gefundene Textur. Dies ist bei den Auswertungen in Abschnitt 5.4.2 und Kapitel 6 gegeben. Bei den dort durchgeführten Experimenten werden aus diesem Grund und auch aufgrund der hier gewonnenen Erkenntnis, dass die zusätzlichen Merkmale auch im infraroten Spektralbereich bei nur geringer Ansichtsveränderung bessere Performance erreichen, die zusätzlichen Merkmale verwendet.

Bei Klassifikation auf Basis der gesamten Sequenz wird, wie zu sehen ist, in beiden Fällen SDA und SDB perfekte Performance mit einer CCR von 100% und einer Falschakzeptanz von 0 erreicht. Die perfekte Performance in dieser Sequenz ist der Tatsache geschuldet, dass es keinerlei Ansichtsvariationen gibt und die Personen lediglich von links nach rechts gehen. Unter diesen Umständen ist die ISM-basierte Wiedererkennung durch die Kombination von visuellem Deskriptor und örtlicher Be-

schreibung durch die ISM-Ausprägung maximal distinktiv und auch zur Diskriminierung unter den hier schwierigen Bedingungen wie niedrigem Kontrast und kaum vorhandener Textur geeignet.

5.4.1.3 Vergleich mit Gangerkennung

In [153] wird die Personenwiedererkennung auf Basis einer Ganganalyse der gesamten Sequenz durchgeführt. Hierbei werden zwei Verfahren, HTI (Head Torso Image) und GEI (Gait Energy Image) zur Wiedererkennung getestet. Dabei werden abhängig von der Wahl der Trainings- und Testdaten (langsam, normal, schnell gehende Person, Person mit Rucksack) unterschiedliche Erkennungsraten festgestellt. Diese sind in Tabelle 5.3 unter A: Normal-Normal, B: Normal-Rucksack, C: Normal-Langsam und D: Normal-Schnell dargestellt.

Tabelle 5.3: CCR-Raten der ISM-Personenwiedererkennung und Gangerkennungsverfahren

Gangerkennung			**ISM-Wiederkennung**		
Typ	**GEI**	**HTI**	**Typ**	**Einzel**	**Sequenz**
A	96%	94%	**Stufe 1**	73%	86%
B	60%	51%	**Stufe 2**	76%	92%
C	74%	85%	**Stufe 3: A**	98%	100%
D	83%	88%	**Stufe 3: B**	100%	100%

Ebenfalls dargestellt sind die CCR-Erkennungsraten der ISM-Wiedererkennung für die Stufen 1-3, jeweils für die unabhängige Einzelklassifikation und für Klassifikation auf Basis der gesamten Sequenz.

Wie zu sehen ist, übersteigen die ISM-Klassifikationsraten auf Basis der gesamten Sequenz mit 100% die Raten der Gangerkennung bei weitem. Dies trifft selbst für die in Abschnitt 5.4.1.2 vorgestellten Raten bei Auswertung nach dem open-set Schema zu. Genauso sieht es für Stufe 3 ohne Nutzung der zeitlichen Kosistenz bei Klassifikation auf Einzelzeitpunktbasis mit 100% bzw. 98% aus. Die Sequenz-Raten auf Stufe 1 und 2 liegen mit 86% bzw. 92% nur hinter den Gangerkennungsraten von A und D für HTI. Dies ist bemerkenswert, da hier nur die niedrigdimensionale Aktivierung, bzw. die Aktivierung mit der Ortsverteilung genutzt wird. Selbst auf Basis von nur einzelnen Zeitpunkten sind die Raten auf Stufe 1 und 2 noch mit denen der Gangerkennung vergleichbar, übersteigen sogar die Raten der Gangerkennung für Szenario B.

Tests haben ergeben, dass die Performance der ISM-Wiedererkennung unabhängig von der Wahl der Trainings- und Testsequenzen ist[6]. Die Raten hierbei sind nahezu identisch mit den hier dargestellten Raten. Dies ist auch einleuchtend, da die erscheinungsbasierte Wiedererkennung unabhängig von der Bewegungsdynamik ist, da keine Bewegungsinformationen genutzt werden. Anderes ist es bei visuellen Veränderungen der Person, z.B. durch zusätzliche Accessoires wie einen Rucksack, zu erwarten. Hier wird die Erscheinung der Person verändert, was auch eine negative Beeinflussung der Wiedererkennung bedeuten kann. Experimente[7] hierzu haben allerdings ergeben, dass die Wiedererkennung nicht negativ beeinflusst wird und keinerlei Performanceeinbußen gegenüber den vorherigen Tests (Personen ohne Rucksack in Trainings- und Testdaten) zu verzeichnen sind.

Abschließend ist also zu sagen, dass die Performance der ISM-Wiedererkennung die der Gangerkennung auf diesem Datensatz bei weitem übersteigt. Dies ist insbesondere bemerkenswert, da es sich hier

[6]Hierzu wurde die Wiedererkennung für jeweils 15 Personen aus den Klassen „schnell“ und „normal“ getestet. Die Trainingsdaten bei diesen Tests blieben unverändert bei der Klasse „langsam“.
[7]Die Erkennungsraten von 20 Testpersonen sind unverändert geblieben.

um Sequenzen aus dem infraroten Spektralbereich handelt, bei denen nur sehr eingeschränkt distinktive visuelle Information zur Wiedererkennung vorhanden ist. Insbesondere übersteigt die Performance der ISM-Wiedererkennung selbst unter den wesentlich schwierigeren Bedingungen einer *open-set* Auswertung mit 100% die der Gangerkennung (ohne open-set Auswertung) bei weitem.

5.4.2 Sichtbarer Spektralbereich: Auswertung der Ansichtsinvarianz

Die Auswertung im sichtbaren Spektralbereich findet auf dem CASIA A Datensatz [1] statt. Dieser Datensatz enthält Sequenzen von 20 Personen, die mit einer Auflösung von 352x240 aufgezeichnet wurden. Für jede Person existieren jeweils 2 Sequenzen für unterschiedliche Winkel, in denen sich die Person relativ zur Kamera durch den Sichtbereich der Kamera bewegt. Sequenzen sind für 6 unterschiedliche Gehrichtungen (0°, 90°, 135°, 180°, 270°, 315°) vorhanden.

Die unterschiedlichen Gehrichtungen sind äquivalent zu durch Veränderung der Kameraperspektive generierten Ansichtsveränderungen. Dieser Datensatz eignet sich also gut, um die Ansichtsinvarianz der ISM-Wiedererkennung gezielt auszuwerten.

Zur Auswertung wird hier die in Abschnitt 5.4.1 genutzte zeitliche Konsistenzbedingungen insofern aufgelöst, als dass jeweils nur ein Zeitpunkt (der letzte an dem die Person verfolgt wurde – dieser enthält somit das gesamte Modell) aus dem aktuellen Anfragemodell genutzt wird. Dies ermöglicht eine Anfrage an die Datenbasis mit nur einem Modell als Anfrageschlüssel, was zusätzlich zur effizienten Vergleichsstrategie die Anfragezeit wegen des geringeren Vergleichsaufwands verringert.

Die Auswertung der Wiedererkennung findet auf 16 Personen des Datensatzes statt. Die restlichen 4 Personen werden zum Training des allgemeinen Personenmodells, auf dessen Basis auch die Ansichtsbestimmung (siehe Abschnitt 5.3.1.1) stattfindet, genutzt. Beispielbilder der Testpersonen sind in Abbildung 5.14 zu sehen.

Zur Durchführung der Experimente zur Wiedererkennung werden Modelle für alle Ansichten der 16 Personen durch ein Tracking der Personen in diesen Sequenzen generiert. Bei 6 Ansichten pro Person und 2 Sequenzen pro Ansicht ergibt dies prinzipiell 2 disjunkte Datensätze von jeweils 96 Modellen. Von diesen Datensätzen stellt einer die Datenbasis, der Andere die Anfragemodelle [8]. Zur Auswertung der Wiedererkennung werden alle Anfragemodelle mit allen Modellen in der Datenbasis verglichen. Obwohl die Datengrundlage lediglich 16 Personen umfasst, ergeben sich hieraus insgesamt 2x9216 Modellvergleiche, was die Experimente auf eine solide Datengrundlage stellt.

Der Vergleich der Anfragemodelle mit der Datenbasis erfolgt für alle Ansichten unabhängig voneinander. Dies ist zur genauen Bestimmung der Ansichtsinvarianz notwendig. Konkret heißt dies, dass die Datenbasis jeweils nur eine bestimmte Ansicht aller Personen enthält. Die Anfrage an diese Datenbasis findet nun für alle Modelle des Anfragedatensatzes, aber wiederum unabhängig für jede Ansicht statt. Somit kann die Ansichtsinvarianz verlässlich bestimmt werden.

Zur Auswertung werden die in Abschnitt 5.3 beschriebenen Methoden zur Ansichtsbestimmung und Modelltransformation (hier Spiegelung) genutzt.

Dies wird auch bei Betrachtung von Abbildung 5.15 deutlich, bei der die CCR bei unterschiedlichen Transformationswinkeln abgebildet sind. Hier ist zu sehen, dass die Erkennungsrate bei gleicher Ansicht mit 97.7% natürlich am höchsten ist. Mit ansteigendem Transformationswinkel nimmt die Erkennungsrate bei 45° auf 46,3%, sowie bei 90° auf 23% ab. Bei einem Winkel von 90° ist die CCR somit am niedrigsten. Dies ist darauf zurückzuführen, dass die Erscheinung einer Person in Seitenansicht in der Realität visuell nur sehr wenig mit der Erscheinung der gleichen Person in Frontalansicht gemein hat. Anders ist es bei der Front- und Rückansicht einer Person. Zwar besteht auch hier keine direkte Verbindung in der Erscheinung (eine Person sieht von Vorne anders aus als von Hinten), allerdings bleibt der Umriss einer Person in beiden Fällen ähnlich (z.B. dicke im Gegensatz zu einer

[8] Jeder Datensatz stellt einmal die Datenbasis und einmal die Anfragemodelle.

Abbildung 5.14: Beispielbilder des CASIA A Testdatensatzes

dünnen Person, bzw. durch Kleidung definierte Form).

Dies trägt auch dazu bei, dass die Erkennungsraten bei Transformationen von 135° bzw. 180° wieder ansteigen. Der Hauptgrund für die hier trotz größerer Winkeldifferenzen wieder besseren Erkennungsraten von 36,8% bei 135° und 73,3% bei 180° ist aber, dass diese Ansichten durch Anwendung der ISM-Transformation, in diesem Fall einer Spiegelung, ineinander überführbar, bzw. im Fall von 135°, annäherbar sind.

Die detaillierte Auswertung für die jeweiligen Winkelkombinationen ist in Tabelle 5.4 zu sehen. Diese zeigt die *Correct Classification Rate* für die jeweiligen Kombinationen der Ansichten des Anfragemodells und der Modelle in der Datenbasis.

In [153] und [162] wurden Gangerkennungsverfahren zur Personenwiedererkennung auf dem gleichen Datensatz getestet. Hier wurde die Wiedererkennung nur in der jeweils gleichen Ansicht getestet, wobei die Auswertung unabhängig für die unterschiedlichen Ansichten erfolgte. Im Durchschnitt der Ansichten erreichten die Methoden dabei jeweils eine CCR von: HTI ([153]): 94,6%, NN+STC ([162]): 68,75%, NN+NED ([162]): 62.1%, ENN+NED ([162]): 83.3%. Da nur gleiche Ansichten getestet wur-

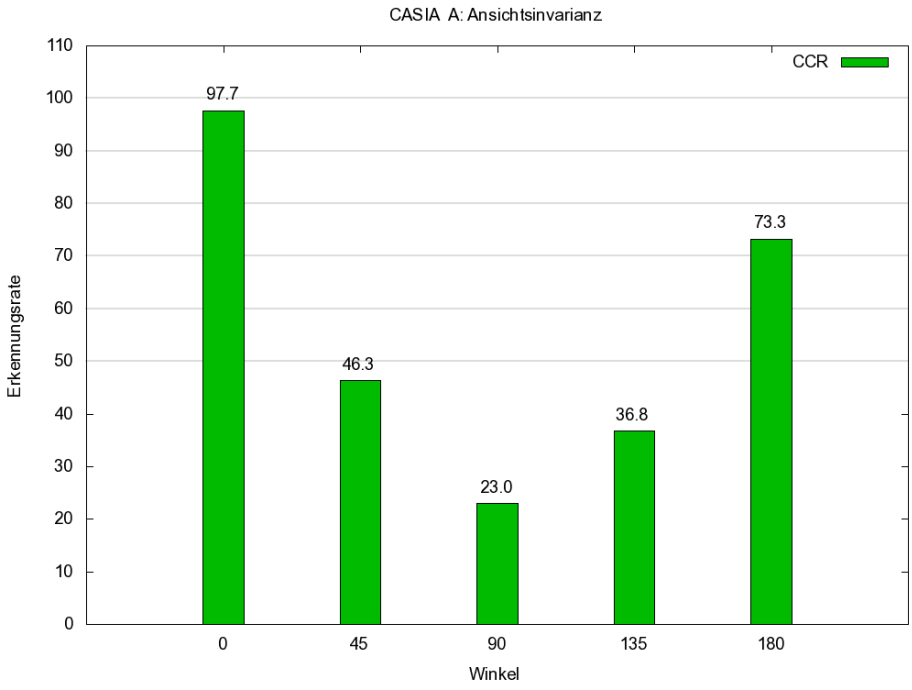


Abbildung 5.15: Ansichtsinvarianz der Personenwiedererkennung auf dem CASIA A Datensatz. Balkendiagramme zeigen Erkennungsrate für verschiedene Ansichtsveränderungen.

Tabelle 5.4: CCR-Raten bei verschiedenen Ansichtskombinationen des CASIA A Datensatzes.

Winkel	**0**	**90**	**135**	**180**	**270**	**315**
0	93	25	42	81	27	57
90		100	36	20	67	34
135			100	50	36	72
180				93	20	25
270					100	42
315						100

den sind die Raten nur mit den 97,7% der ISM-Wiedererkennung bei gleicher Ansicht vergleichbar. Es zeigt sich also, dass die ISM-Wiedererkennung eine bessere Performance aufweist als die Gangerkennungsverfahren. Interessant an den Ergebnissen der Gangerkennung ist, dass die besten Raten bei einem Winkel von 90°, also wenn die Person von vorne oder hinten zu sehen ist, erreicht werden. Dies ist unerwartet, wenn man bedenkt, dass die Gangart einer Person eigentlich am besten von der Seite einschätzbar sein sollte, und lässt darauf schließen, dass in der Gangerkennung implizit auch erscheinungsabhängige Merkmale eine Rolle spielen.

Zur Auswertung der Ansichtsinvarianz des Modells waren bei der jeweiligen Anfrage mit einem Modell nur Modelle einer bestimmten Ansicht in der Datenbasis. Dies war notwendig, um die Ansichtsinvarianz des Modells bewerten zu können. Unter realen Bedingungen sind allerdings häufig Modelle

verschiedener Ansichten in der Datenbasis vorhanden. Dies ist bei Benutzung der in Abschnitt 5.3.1.1 vorgestellten Methodik zur Ansichtsauswahl unproblematisch. Allerdings ist die Herausforderung für die Wiedererkennung in diesem Fall größer, da zwischen Modellen der gleichen Ansicht grundsätzlich eine höhere Ähnlichkeit als zwischen Modellen abweichender Ansichten zu erwarten ist. Liegt das richtige Modell in der Datenbasis nur in einer anderen Ansicht vor, viele falsche Modelle aber in der gleichen Ansicht, so ist die richtige Klassifikation eine größere Herausforderung. In Abschnitt 5.1.2 wurde bereits eine Herangehensweise eingeführt, die diese Problematik durch Ausfilterung ansichtsspezifischer Merkmale verringert. Insbesondere wurde durch eine eingeführte Ansichtstransformation des Modells ein Weg aufgewiesen – und für den Fall der Spiegelung exemplarisch gezeigt – die Problematik zu umgehen.

Für die Experimente unter den schwierigeren Bedingungen wird die Datenbasis mit Testmodellen aller Ansichten gefüllt. Das richtige Modell wird der Datenbasis in der jeweils zu testenden Vergleichsansicht hinzugefügt. Die Tests werden dabei für Modelle durchgeführt, die bei der ersten Auswertung korrekt klassifiziert wurden. In den Experimenten zeigt sich, dass die Erkennungsrate bei Modellen unterschiedlicher Ansichten, die nicht durch eine Transformation ineinander überführbar sind, stark abfällt. Viele der in den vorherigen Test korrekt klassifizierten Personen werden nicht mehr korrekt erkannt und weisen größere Ähnlichkeit zu anderen Personen auf, die in gleicher Ansicht in der Datenbasis vorliegen.

Da der Grund für die Fehlklassifikationen in der Grundähnlichkeit von Modellen gleicher Ansichten liegt, können diese durch eine Transformation auf Ebene des Ähnlichkeitswerts, auf Basis der bekannten Ansichten des Anfrage- und Datenbasismodells, behoben werden. Dazu wird nach der Ähnlichkeitsbestimmung eines Anfragemodells mit den Modellen der Datenbasis eine Transformation der jeweiligen Ähnlichkeiten in Abhängigkeit der Ansichtsunterschiede zwischen Anfrage- und Datenbasismodell durch Multiplikation mit einem Faktor ρ (Gleiche Ansicht: $\rho = 1.0$, durch Spiegelung ineinander überführbar: $\rho = 2.0$, sonst: $\rho = 3.0$) durchgeführt.

Die erneute Auswertung unter Nutzung dieser Transformation ergibt, dass alle Modelle, die bei den ersten Tests korrekt klassifiziert wurden, auch unter den schwierigeren Bedingungen korrekt klassifiziert werden. Es werden somit die in Tabelle 5.4 und Abbildung 5.15 dargestellten Erkennungsraten erreicht.

Kapitel 6

Auswertung des Gesamtsystems

In den letzten drei Kapiteln wurden die Verfahren zur Personendetektion, Personenverfolgung und Personenwiedererkennung vorgestellt. Hierbei wurde jeweils eine experimentelle Validierung der Verfahren vorgenommen. Da die drei Stufen aufeinander aufbauen, enthalten die Auswertungen implizit auch die Ergebnisse der vorherigen Stufen. D.h., die Auswertung der Wiedererkennung in Kapitel 5 enthält implizit auch die Performance der Personenverfolgung, da diese Grundlage für die Erstellung der Identitätsmodelle zur Wiedererkennung ist.

Allerdings wurde die Wiedererkennung in Kapitel 5 auf explizit für den Zweck der Personenwiedererkennung aufgezeichneten Daten getestet. Obwohl auch hier alle Teilschritte bis zur Wiedererkennung, sprich Detektion und Tracking, vollautomatisch abgelaufen sind, d.h. keine manuelle Annotation in irgendeiner Form erfolgt ist, ist der Schwierigkeitsgrad für Detektion und Tracking in diesen Sequenzen eher gering, da sich zu jedem Zeitpunkt nur eine einzelne Person im Sichtbereich der Kamera aufhält.

In realen Überwachungsszenarien ist dies natürlich nicht der Fall. Hier treten typischerweise mehrere Personen gleichzeitig auf, wobei sich dabei insbesondere die Problematik von Verdeckungen ergibt. Um das hier vorgestellte System auch unter dem Aspekt der realen Einsatzbedingungen zu betrachten, wird in diesem Kapitel eine Auswertung des Gesamtsystems, d.h. aller drei Stufen, auf realen Überwachungsdaten aus dem iLids-Datensatz [133] vorgenommen. Dieser besteht aus Videodaten mehrerer Kameras und wurde an einem Flughafen unter realen Bedingungen aufgenommen – d.h. er zeigt nicht-gestellte Aufnahmen. Hier treten also alle Herausforderungen auf, die in der Realität für ein Überwachungssystem bestehen. Beispielbilder dieses Szenarios sind in Abbildung 6.1 zu sehen.

Bei diesem Szenario treten Herausforderungen für alle Stufen der Verarbeitungskette bis zur Wiedererkennung auf[1]. Unzulänglichkeiten bei der Detektion pflanzen sich somit bis zur Wiedererkennung fort. D.h., wird eine Person nicht korrekt detektiert, so kann folglich auch kein konsistentes Identitätsmodell der Person zur Wiedererkennung erstellt werden. Gleiches gilt auch für das Tracking. Falls die Identitätszuordnung während des Trackings[2] nicht konsistent ist, d.h. es Identitätsvertauschungen zwischen den verfolgten Personen gibt, so wird auch das aufgebaute Identitätsmodell zur Wiedererkennung inkonsistent. Im Falle einer Identitätsvertauschung zwischen zwei Personen würde jedes Modell auch Merkmale der jeweils anderen Person enthalten.

Für die Wiedererkennung gelten im Vergleich mit der Auswertung in Abschnitt 5.4 nicht nur die erschwerten Bedingungen aufgrund der größeren Herausforderungen für Detektion und Tracking. Im Gegensatz zu 5.4 wird die Wiedererkennung hier in einem Multi-Kamera-Szenario getestet. D.h.,

[1] Es ist anzumerken, dass diese Schwierigkeiten auch bei der vorherigen Auswertung der Wiedererkennung bereits bestanden. Allerdings waren in diesen Fällen die Schwierigkeiten eher gering, da sich, wie bereits erwähnt, lediglich eine Person zu einem Zeitpunkt im Sichtbereich der Kamera aufhielt.

[2] Hier ist die Identitätszuordnung bei zeitlich zusammenhängendem Auftreten einer Person im Sichtbereich einer Kamera gemeint.

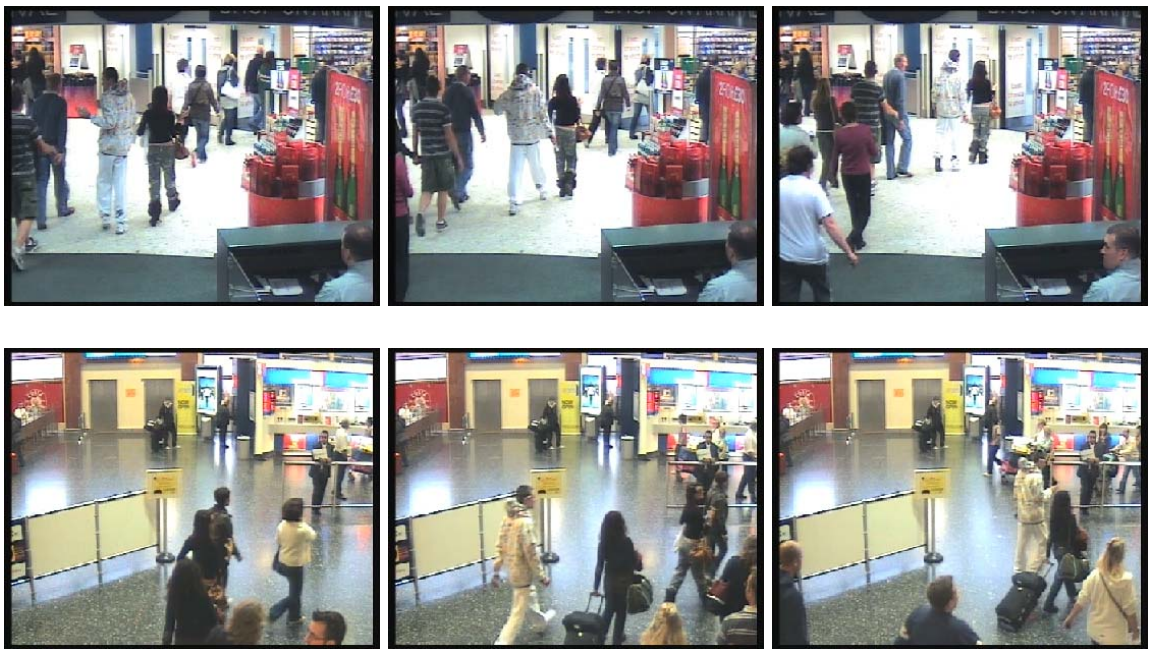


Abbildung 6.1: Beispielbilder des iLids-Szenarios. Obere Reihe: Kamera 1, untere Reihe: Kamera 3.

es geht nicht mehr darum, eine Person in der gleichen Kamera mit gleichem Blickwinkel auf die Szene wiederzuerkennen, sondern um den schwierigeren Fall, eine Person, die in einer Kamera gesehen wurde, in einer anderen Kamera wiederzuerkennen. Dies ist im Vergleich mit der Wiedererkennung in einer Kamera eine wesentlich größere Herausforderung, da die Personenwiedererkennung hier mit unterschiedlichen Blickwinkeln der Kameras, starken Abweichungen der Umgebungsvariablen wie z.B. der Beleuchtung, und auch mit durch die Verwendung unterschiedlicher Kameratypen induzierten Unterschieden wie unterschiedlichen Farbdarstellungen[3] der Kameras zurechtkommen muss.

Dieses Kapitel ist wie folgt aufgeteilt. In Abschnitt 6.1 erfolgt eine Beschreibung des Anwendungsszenarios. In den Abschnitten 6.2 bzw. 6.3 erfolgt die Auswertung des Personentrackings bzw. der Personenwiedererkennung.

6.1 Anwendungsszenario

6.1.1 Anwendungsfälle

Der iLids-Datensatz enthält Videodaten eines realen Multi-Kamera-Netzes, also eines typischen Szenarios, bei dem die hier vorgestellten Methoden zur Personenverfolgung und -wiedererkennung relevant sind.

Eine typische Aufgabe in solchen Kameranetzen ist die Extraktion der Laufwege von Personen im ganzen überwachten Bereich. D.h., man will ein umfassendes Gesamtbild erstellen, das angibt, wo sich welche Person zu welchem Zeitpunkt aufgehalten hat. Hierzu ist es zunächst notwendig, die Personen

[3]Der letzte Punkt betrifft vor allem Verfahren, welche zur Wiedererkennung über Kameragrenzen hinweg Farbmerkmale nutzen. Das hier vorgestellte Verfahren auf ISM-Basis ist von der unterschiedlichen Farbdarstellung in verschiedenen Kameras weitestgehend unabhängig.

in den einzelnen Kameras zu verfolgen. Die Personenwiedererkennung wird dann dazu eingesetzt, Verbindungen zwischen den Einzeltracks der Personen in den einzelnen Kameras aufzubauen. Auf Basis dieser Informationen ist es möglich, eine umfassende Situationsanalyse bzw. -interpretation durchzuführen.

Um ein vollautomatisches Multi-Kamera-Tracking von Personen durchzuführen, ist eine eindeutige Zuordnung von Tracks zwischen den Kameras notwendig. Dies ist in der Praxis bei einer Vielzahl von Personen in den jeweiligen Kameras oft nur schwer möglich. Um die Anzahl in Frage kommender Personen einzuschränken, können hierzu Informationen über Raum-Zeit-Abhängigkeiten zwischen den Kameras eingebracht werden [83, 84, 127]. Die rein erscheinungsbasierte Personenwiedererkennung muss in diesen Fällen also lediglich eine Untermenge der vorhandenen Personen unterscheiden (siehe Auswertung mit SDR-Kriterium in Abschnitt 6.3).

Da in den meisten insbesondere sicherheitskritischen Bereichen zur Zeit noch keine vollautomatischen Systeme zur Überwachung eingesetzt werden, ist ein in der heutigen Realität häufig vorkommendes Anwendungsszenario für ein Überwachungssystem die Unterstützung eines menschlichen Operators. Hier kann das System dazu genutzt werden, den Operator in bestimmten Situation zu alarmieren bzw. bei bestimmten Arbeiten zu unterstützen.

Gerade in Kameranetzen ist eine solche Unterstützung des menschlichen Auswerters durch ein automatisches System wünschenswert, da hier immense Datenmengen auflaufen, die für den Menschen nur schwer zu bewältigen sind. So kann ein System zur Personenwiedererkennung zur Unterstützung des Operators bei der Suche nach bestimmten Personen in den Daten eines Kameranetzes genutzt werden. Ein Anwendungsfall hierfür ist z.B. die Auswertung nach einem durch Videokameras beobachteten Diebstahl im Fall eines überwachten Kaufhauses. Hier kann die Wiedererkennung dazu genutzt werden, herauszufinden, woher der Dieb vor dem Diebstahl gekommen ist, bzw. auf welchem Weg er nach dem Diebstahl entkommen ist. Ein weiterer Anwendungsfall ist die Suche nach einem vermissten Kind in einem überwachten Areal (wie z.B. Flughäfen und Kaufhäuser). In diesem Fall könnte aufgrund des letzten bekannten Aufenthaltsort des Kindes ein Identitätsmodell aus der Datenbasis ausgewählt und anhand dieses Modells eine Suche in der Datenbasis durchgeführt werden, um den aktuellen Aufenthaltsort des vermissten Kindes zu bestimmen.

In beiden Fällen reicht es zur Unterstützung des Operators aus, wenn die durch den Operator zu überprüfende Personenmenge durch ein automatisches Verfahren reduziert wird. Hier kann das System dem Operator also eine Menge von N Personen aus der Datenbasis vorschlagen und es ist ausreichend, wenn die Anfrageperson in dieser Menge enthalten ist. Das Anfragemodell muss also nicht in 100% der Fälle als bestes Modell in der Datenbasis identifiziert werden.

Um solche Anwendungsfälle, bei denen anhand eines Beispielmodells nach Vorkommen dieses Modells in der Datenbasis gesucht wird, bearbeiten zu können, ist es notwendig, dass das System fortlaufend Tracking-Daten aller Personen in der Datenbasis ablegt. Insbesondere in diesem Fall kommt der Vorteil des hier vorgestellten Gesamtsystems zum Tragen. Es müssen nämlich nicht die gesamten Bilddaten abgelegt werden, sondern lediglich die während des Trackings aufgebauten Identitätsmodelle[4]. Da es bereits sehr effiziente Ablagemöglichkeiten für Bilddaten gibt, ist der geringere Speicherbedarf, je nachdem wieviele Personen sich im Durchschnitt in der überwachten Szenerie aufhalten, evtl. zu vernachlässigen, d.h. nur marginal von Vorteil. Vielmehr liegt der Vorteil dieses Systems darin, dass es besonders zur schnellen Indizierung in Datenbanken geeignet ist. Außerdem ist es in Multi-Kamera-Systemen durch das hier vorgestellte System prinzipiell möglich, nur noch Metadaten, d.h. Informationen über die getrackten Personen über das Kameranetzwerk zu einem zentralen Speicherort zu übertragen. Dies ermöglichen sogenannte „intelligente Kameras“, welche die Verarbeitungshardware bereits integriert haben.

[4]Unter der Annahme, dass alle relevanten Objektklassen im System modelliert sind.

Abbildung 6.2: Beispielbilder der Personen des iLids-Szenarios jeweils paarweise für Kamera 1 und 3.

6.1.2 Testdaten

Zum Test des Gesamtsystems wurde eine Teilmenge der Daten der Kameras 1 und 3 des iLids-Datensatzes [133] verwendet. Die Sichtbereiche der beiden Kameras sind disjunkt und unterscheiden sich in ihrem Blickwinkel auf die Szene. Insbesondere sind aufgrund der Standardlaufwege in diesen Sequenzen, die Personen nicht von vorne, und in den beiden Kameras auch aus unterschiedlichen Ansichten zu sehen, was die Personenwiedererkennung vor eine größere Herausforderung stellt.

Aus diesen Daten wurden Sequenzen von insgesamt 45 Personen, die in den Videodaten beider Kameras zu sehen sind, extrahiert. Die Länge der jeweiligen Sequenz richtet sich nach der Aufenthaltszeit der relevanten Person in der Szene. In den Sequenzen treten die Personen nur selten einzeln auf. Der Normalfall ist, dass sich mehrere Personen, teilweise in Gruppen durch den Sichtbereich der Kamera bewegen. Hierbei kommt es häufig zu Verdeckungen zwischen Personen, was das Tracking schwieriger macht. Insbesondere führen die meisten Personen auch Gepäck mit sich, was ebenfalls zu Teilverdeckungen der Personen führt, und insbesondere für die Wiedererkennung eine weitere Herausforderung darstellt. Weitere Herausforderungen für die Wiedererkennung sind die unterschiedlichen Umgebungsbedingungen in den beiden Kameras. So variiert die Beleuchtung und damit die Erscheinung der Personen zwischen den beiden Kameras stark.

Wie die Beispielbilder des Szenarios in Abbildung 6.1 zeigen, sind in diesem Szenario alle in der Realität vorkommenden Herausforderungen für ein Überwachungssystem enthalten. Beispielbilder der Personen in den Testdaten sind in Abbildung 6.2 zu sehen.

6.2 Tracking

Tabelle 6.1 zeigt die Ergebnisse der Detektions- und Trackingauswertung des iLids-Szenarios. Details hinsichtlich der Auswertungsmaße und -kriterien sind in Abschnitt 4.5.1 zu finden[5]. Für die Auswertung wurde eine Teilmenge der zur Auswertung der Wiedererkennung ausgewählten 90 Sequenzen (45 aus jeder Kamera) ausgewertet. Die ausgewerteten Sequenzen stellen hinsichtlich der Schwierigkeiten für das Tracking eine repräsentative Teilmenge aller Sequenzen dar. Auf eine Auswertung der gesamten Sequenzen wurde aufgrund des immensen Annotationsaufwands, der für eine Auswertung von Detektion und Tracking notwendig ist, verzichtet. Wie Tabelle 6.1 zeigt, ist die Performance des

Tabelle 6.1: Trackingperformance im iLIDS-Szenario.

	Frames	Objekte (#ids)	MOTP	$\overline{m}$	$\overline{fp}$	$\overline{mm}$	**MOTA**	Recall	Fp/Bild
iLids	1797	3842 (31)	0.76	0.19	0.11	0.001	**0.70**	0.81	0.26

Trackings bei Berücksichtigung der Schwierigkeit des Szenarios mit einer MOTA Rate von 0.7 sehr gut. Ein wichtiger Aspekt für die folgende Personenwiedererkennung ist die niedrige Vertauschungsrate $\overline{mm}$. Diese ist äußerst relevant, da es bei Vertauschungen zwischen Personen zu Verfälschungen der Identitätsmodelle kommt.

6.3 Wiedererkennung

Die Basis der Wiedererkennung stellen die während des Trackings generierten 90 Modelle der 45 Personen dar. Die in einer Kameraansicht generierten Modelle werden als Datenbasis für die Wiedererkennung genutzt, die Modelle aus der anderen Kameraansicht als Anfragemodelle. Somit ist ein Test der Wiedererkennung über Kameraansichten hinweg möglich.

Die Wiedererkennung wird mit zwei aktuellen Verfahren [9] von 2010 verglichen, die ebenfalls eine Auswertung (mit 44 Personen in der Datenbasis) der Wiedererkennung auf dem iLids-Datensatz durchführen. Diese gehen von ähnlichen Vorbedingungen für die Wiedererkennung aus, was konkret bedeutet, dass die Wiedererkennung auch hier auf einer echten Detektion aufbaut und somit auch mit eventuellen Unzulänglichkeiten dieser zurechtkommen muss. Die beiden in [9] vorgestellten Verfahren nutzen *Haar-Merkmale*, bzw. *DCD (Dominant Color Descriptors)* [172] zur Personenwiedererkennung.

Zur Personendetektion wird bei diesen Ansätzen ein HOG-basierter Personendetektor [41] genutzt. Dieser stellt initiale bounding boxes der Personen bereit, wobei diese durch einen Farbsegmentierungsansatz [11] zu einer detaillierten Objektsegmentierung verfeinert werden. Diese Segmentierung wird zum Ausfiltern der Haar-Merkmale, bzw. DCD, die auf dem Hintergrund gefunden wurden, genutzt. Diese werden also nicht mit in die Wiedererkennung einbezogen. In einem Trainingsschritt werden mit dem AdaBoost-Algorithmus Modelle für die Personen trainiert. Die Detektionen der jeweiligen Person werden dabei als Positiv-, andere Personen aus der gleichen Szene als Negativbeispiele in den Trainingsalgorithmus eingebracht. An dieser Stelle erfolgt also ein manueller Eingriff in die Trainingsprozedur durch das Einbringen von Negativbeispielen. Diese können nicht automatisch vom

[5] Da für die Auswertung gezielt Sequenzen einzelner Personen, die in beiden Ansichten zu sehen sind, ausgewählt wurden, erhebt diese Auswertung keinen Anspruch auf Allgemeingültigkeit für das komplexe iLids-Szenario. Für die Auswertung wurden natürlich keine Massenszenen ausgewählt, da diese nicht, oder nur sehr vereinzelt, für die Auswertung der Wiedererkennung geeignet sind. Um eine umfassende Auswertung des Trackings auf den iLids-Daten vorzunehmen, müssten auch solche Szenen in großem Umfang betrachtet werden, was aber nicht Ziel dieser Auswertung ist.

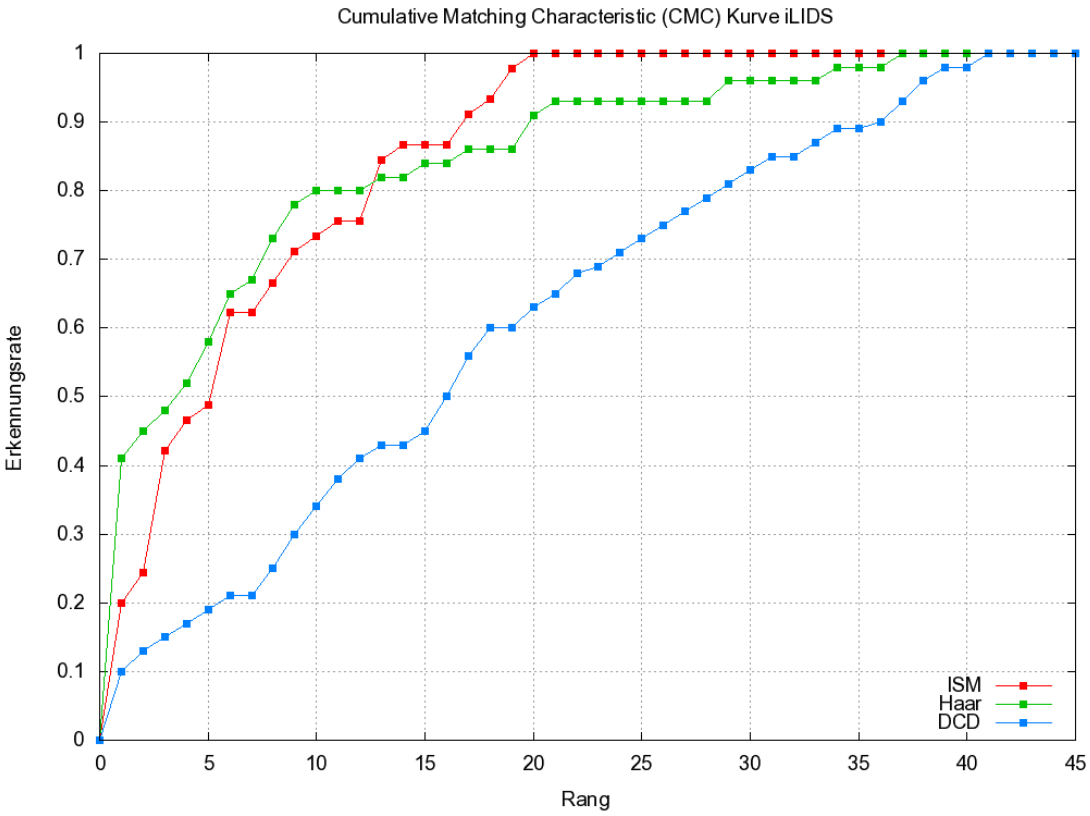


Abbildung 6.3: CMC-Kurve zur Personenwiedererkennung im iLids-Szenario. Dargestellt ist die Performance der hier vorgestellten „ISM“ Wiedererkennung, der „Dominant Color Descriptors (DCD)“ und „Haar“ Merkmale aus [9].

System ausgewählt werden, da sichergestellt sein muss, dass diese nicht die Person zeigen, für die das Modell trainiert wird. Insbesondere erfolgt der AdaBoost-Trainingsschritt offline und nicht während des Trackings. Dies ist aufgrund der langen Trainingsdauer[6], die der AdaBoost-Algorithmus benötigt, grundsätzlich nicht möglich.

Zur *Auswertung* der Wiedererkennung wird die *Cumulative Match Characteristic (CMC)* verwendet. Diese Charakteristik ist das Standardmaß für die Evaluierung von Personenwiedererkennung in komplexen Szenarien [72]. Das Wiedererkennungsproblem wird hierbei als Ranking-Problem betrachtet. Zum Aufbau der CMC-Kurve wird ein Histogramm mit N Fächern, wobei N die Anzahl von Personen in der Datenbasis ist, gebildet. Der Rang des richtigen Modells in der Datenbasis für ein Anfragemodell wird in das Histogramm eingetragen. Wenn das richtige Modell in der Datenbasis für ein Anfragemodell also die drittgrößte Ähnlichkeit aller Modelle in der Datenbasis aufweist, so hat dieses Modell Rang 3 und wird in das Histogrammfach 3 eingetragen. Durch Akkumulierung der Histogrammeinträge ergibt sich die CMC-Kurve. Diese Bewertungsstrategie passt zu einem Anwendungsszenario, bei dem die Wiedererkennung zur Unterstützung eines Operators bei der Suche nach Personen in großen Datenbeständen verwendet wird.

Die Ergebnisse der in dieser Arbeit vorgestellten *ISM-Wiedererkennung*, sowie die der beiden Verfahren „Haar“ und „DCD“ aus [9] sind in Abbildung 6.3 dargestellt. Wie zu sehen ist, liegt die Performance

[6] Abhängig von der Anzahl der Trainingsbilder benötigt das Training mehrere Stunden.

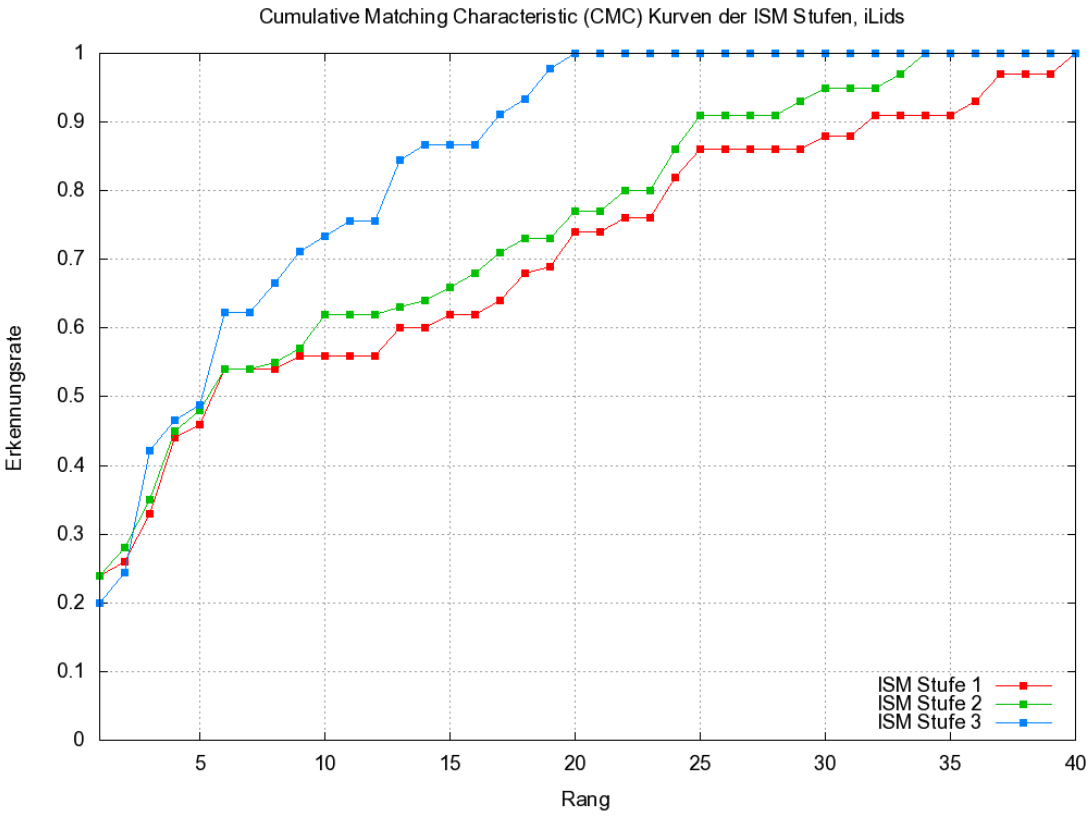


Abbildung 6.4: CMC-Kurven der verschiedenen Stufen der ISM-Wiedererkennung im iLids-Szenario.

des DCD-Ansatzes weit unter der Performance der beiden anderen Ansätze. Dies liegt daran [9], dass es starke, durch die Umgebungsbedingungen sowie den Kameratyp beeinflusste Farbunterschiede zwischen den beiden Kameraansichten gibt, die selbst durch eine durchgeführte Farbkalibrierung nicht zu beheben sind. Dies ist ein Beispiel für die in Kapitel 2 beschriebenen Schwierigkeiten, die bei der Nutzung von Farbe zur Wiedererkennung in Kameranetzen auftreten.

Wie zu sehen ist, ist die Performance des ISM-Ansatzes vergleichbar mit der Performance des Haar-basierten Ansatzes. Allerdings fällt auf, dass die Performance in der ersten Hälfte des Graphen etwas unter der der Haar-Wiedererkennung liegt. Dies liegt daran, dass die Distinktivität der Haar-Signatur etwas besser ist als die der ISM-Wiedererkennung. Diese Distinktivität wird bei der Haar-Signatur durch einen offline Trainingsschritt erkauft, in dem die Modelle auf Basis des AdaBoost-Algorithmus trainiert werden. Dies schränkt die Anwendbarkeit des Haar-Ansatzes im Vergleich mit der hier vorgestellten Wiedererkennung, bei der die Modelle online während des Trackings aufgebaut werden, ein. In der zweiten Hälfte liegt die Performance der ISM-Wiedererkennung über der der Haar-Signatur. Das bedeutet, dass es weniger absolut falsche Klassifikationen gibt, bei denen das Anfragemodell nicht unter den Top 50% der Modelle in der Datenbasis ist. Faktisch liegen bei der ISM-Wiedererkennung 100% der 45 Personen unter den Top 20 (Top 45%).

Eine detaillierte Auswertung der ISM-Wiedererkennung mit der Performance auf den verschiedenen Stufen ist in Abbildung 6.4 zu sehen. Wie hier zu sehen ist, ist die Performance auf Stufe 1 und 2 insbesondere in der linken Hälfte des Graphen vergleichbar mit der auf Stufe 3. Dies ist bemerkenswert, da die Modelldimension auf diesen Stufen wesentlich geringer (Stufe 1: Codebuchdimension. In diesem

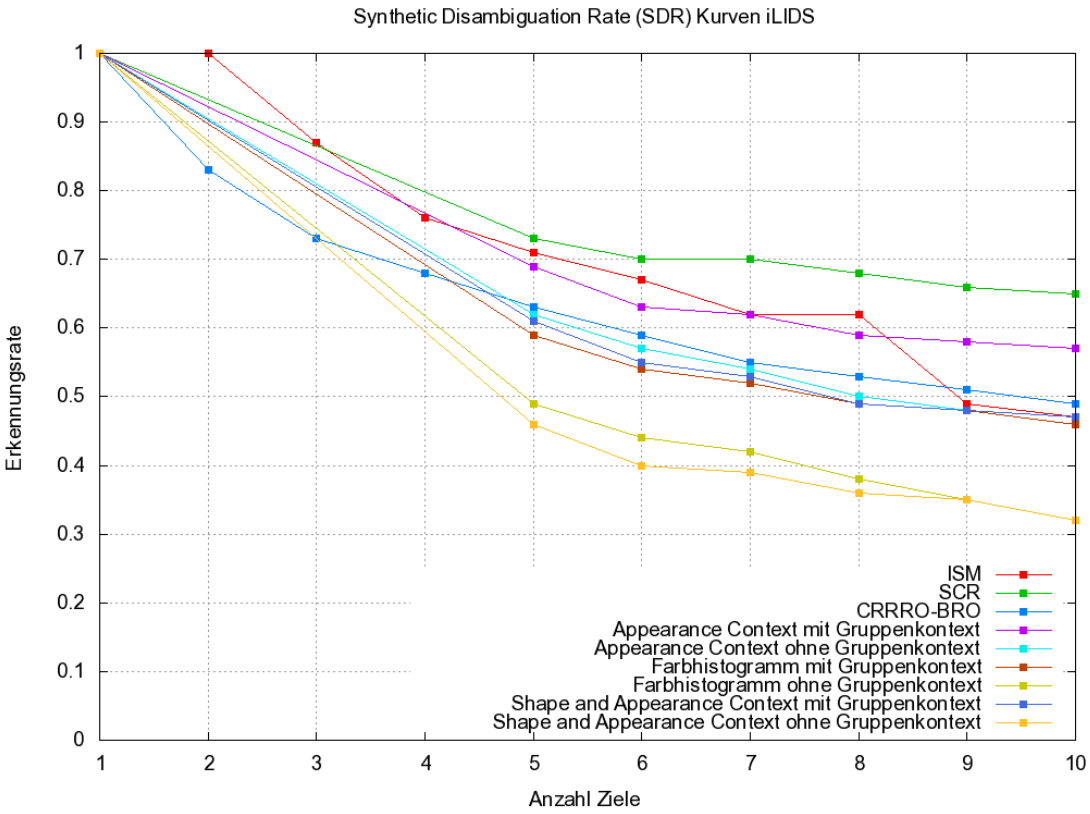


Abbildung 6.5: SDR Kurven veschiedener Wiedererkennungsverfahren auf dem iLids-Datensatz.

Fall 286) ist. Es zeigt also, dass die reine Codebuchsignatur auch bei stark unterschiedlichen Ansichten noch gute Ergebnisse bei der Wiedererkennung liefert. Die Performance verbessert sich durch Hinzunahme der Offsets auf Stufe 2 gegenüber Stufe 1 nur marginal. Dies zeigt, dass die Ortsverteilung aufgrund der starken Ansichtsunterschiede zwischen den Kameras nicht wesentlich zur Erhöhung der Distinktivität beitragen kann.

Eine umfangreiche Auswertung auf 65 (mit Gruppenkontext), bzw. 119 (Einzelpersonen) manuell ausgeschnittenen Personen des iLids-Datensatzes wird in [180] und [10] vorgenommen. Da die Schwierigkeit der Wiedererkennung mit größerer Anzahl von Personen in der Datenbasis ansteigt, ist ein Vergleich mit diesen Ansätzen auf Basis der für die ISM-Wiedererkennung generierten CMC nicht korrekt. Um auch mit diesen Methoden einen fairen Vergleich zu ermöglichen, kann ein Evaluierungsmaß genutzt werden, welches die Größe der verwendeten Datenbasis in die Erkennungsraten einbezieht.

Diese *Synthetic Disambiguation Rate (SDR)* gibt an, wie gut die Wiedererkennungsrate bei unterschiedlichen Datenbasisgrößen ist. Die SDR kann aus der CMC wie folgt [72] errechnet werden:

$$SDR(M) = CMC\left(\frac{N}{M}\right). \tag{6.1}$$

Hierbei ist N die Anzahl von Modellen in der Datenbasis, M die Anzahl von Anfragemodellen, d.h. infrage kommenden Modellen in der Datenbasis, und CMC(k) die CMC-Erkennungsrate für Rang k. Diese Auswertemethodik kann einen Eindruck geben, wie gut die Wiedererkennung für das Multi-

Kamera-Tracking von Personen geeignet ist, da hierbei typischerweise weitere, über die reine Erscheinung der Person hinausgehende Informationen, wie Raum-Zeit-Abhängigkeiten zwischen den Kameras eingebunden werden können. Diese zusätzlichen Informationen ergeben Bedingungen, welche die Anzahl der in Frage kommenden Modelle in der Datenbasis einschränken, wodurch sich die Anzahl von M auf die Anzahl von Personen reduziert, für die eine rein erscheinungsbasierte Wiedererkennung durchgeführt werden muss. In der Realität sind dies bei zwei Kameras typischerweise maximal 10 (es dürfen also 10 Personen gleichzeitig von dem Sichtbereich einer Kamera in den der anderen Kamera gehen).

Die SDR der ISM-Wiedererkennung sowie die der Vergleichsverfahren sind in Abbildung 6.5 dargestellt. Als Vergleichsverfahren sind die *Spatial covariance Regions (SCR)* aus [10], sowie die *Center Rectangular Ring Ratio-Occurrence Descriptor-Block based Ratio-Occurrence (CRRRO-BRO)*, *Appearance Context*, *Farbhistogramm* und *Shape and Appearance Context* aus [180] dargestellt. Die letzten drei Verfahren wurden jeweils mit und ohne die in [180] vorgestellte Gruppenkontexterweiterung ausgewertet. Die SDR-Raten dieser Verfahren wurden, falls in den Artikeln explizit ausgewiesen direkt übernommen, ansonsten anhand der ausgewiesenen CMC-Rate berechnet.

Der ISM-Ansatz liegt insgesamt in der Spitzengruppe der Verfahren. Wie zu sehen ist, ist die Performance des ISM-Ansatzes bei wenigen Personen besser, bzw. im Fall von SCR nur marginal schlechter, als die der anderen Ansätze. Ab 9 Personen ist die Performance dreier anderer Ansätze besser. Dies ist, wie bereits beim Vergleich mit [9] herausgestellt wurde, darin begründet, dass diese Ansätze einen offline Trainingsschritt durchführen. Ein solcher offline Trainingsschritt wird beim hier vorgestellten System aufgrund der online-Fähigkeitsanforderung nicht genutzt. Insbesondere sind auch die sonstigen Bedingungen bei Auswertung der ISM-Wiedererkennung von höherem Schwierigkeitsgrad, da hier, im Gegensatz zu allen anderen vorgestellten Verfahren, keine manuelle Annotation in Form von bounding boxes oder einer Segmentierung vorgenommen wird.

Insgesamt kann also gesagt werden, dass das hier vorgestellte System zur Detektion, Verfolgung und Wiedererkennung von Personen nicht nur durch seine allgemeine Anwendbarkeit und Generizität, sondern auch durch die gute Performance der einzelnen Stufen überzeugen kann. Hierbei wird die Performance von auf Einzelanwendungen spezialisierten state-of-the-art Methoden erreicht, bzw. sogar übertroffen.

Kapitel 7

Zusammenfassung und Ausblick

Zusammenfassung

In dieser Arbeit wurde ein System zur Detektion, -verfolgung und -wiedererkennung von Personen in Videosequenzen vorgestellt. Dieses System hat auf Ebene des Gesamtsystems zwei wesentliche Eigenschaften, die es von anderen Systemen abheben.

Dies ist zum Einen der *integrierte Systemansatz*, bei dem die drei Aufgaben als gemeinsame Problemstellung angegangen werden und das Implicit Shape Model als Basis des Gesamtsystems für die verschiedenen Ebenen erweitert wird.

Der weitere Aspekt ist die *Systemgenerizität*, die sich durch alle Systemebenen zieht und die in dieser Arbeit in verschiedenen experimentellen Validierungen nachgewiesen wurde. Dabei wird die *Sensorunabhängigkeit* durch ausschließliche Nutzung lokaler, auf Intensitätsgradienten basierender Bildmerkmale erreicht und durch Auswertung der Verfahren auf Daten aus dem sichtbaren und infraroten Spektralbereich validiert. Die *Unabhängigkeit von der Objektklasse* wurde anhand der Objektklasse „Schiff“ und durch die *hierarchische Objektdetektion und -verfolgung* demonstriert. Durch die Evaluierung in verschiedenen Anwendungsdomänen und -szenarien wurde die *Szenariogenerizität* gezeigt. Insbesondere wurde hier auch gezeigt, dass das System bei bewegter, bzw. durch die eingeführte Bewegungskompensation auch bei stark bewegter Kamera einsetzbar ist.

Über diese beiden wesentlichen Aspekte des Gesamtsystems hinaus wurden in dieser Arbeit neue Ansätze zur Personenverfolgung und -wiedererkennung eingeführt.

Zur *Personenverfolgung* wurde ein Verfahren vorgestellt, welches das zur Objektdetektion entwickelte Implicit Shape Model zur Nutzung zur Objektverfolgung weiterentwickelt. Dieses Verfahren *integriert Objektdetektion und -verfolgung* durch die Kombination von tracking-by-detection Strategien mit modellbasierten Herangehensweisen. In Auswertungen wurde gezeigt, dass dieses Verfahren die Objektdetektion stabilisiert und dessen Performance verbessert. Im Gegensatz zu reinen tracking-by-detection Verfahren ermöglicht dieses ISM-Tracking das *Tracking während Verdeckungen*, ohne dabei spezielle Heuristiken einzusetzen. Insbesondere macht der integrierte Ansatz separate Schritte zur Datenassoziation und Modellaktualisierung während des Trackings überflüssig. Die Evaluierung erfolgte in unterschiedlichen Anwendungsszenarien und -domänen sowie in Sequenzen mit verschiedenen Herausforderungen für die Personenverfolgung. Hierbei wurde gezeigt, dass das Verfahren unabhängig von Anwendungsszenario und Umgebungsbedingungen gute Performance erreicht. Durch die, zusätzlich zur Auswertung im sichtbaren Spektralbereich, durchgeführte Auswertung im infraroten Spektralbereich, wurde neben der Sensorunabhängigkeit des Systems auch die Anwendbarkeit des ISM-Trackings unter den im infraroten Spektralbereich schwierigeren Bedingungen unter Beweis gestellt. Die quantitative Auswertung hat gezeigt, dass das eingeführte Verfahren in Bezug auf die Performance mit spezialisier-

ten *state-of-the-art* Verfahren *vergleichbar* ist, bzw. diesen in einigen Fällen *überlegen* ist. Dies betrifft insbesondere auch Szenarien, in denen sensor-, anwendungs- oder umgebungsspezifische Information von anderen Verfahren genutzt wird, das ISM-Tracking aber die Systemgenerizität beibehält. Ebenso wurde gezeigt, dass ein *Tracking bei bewegter Kamera* sowie durch Einführung einer integrierten *Bewegungskompensation* auch bei *stark bewegter Kamera* möglich ist. Auch in diesen schwierigen Fällen ist durch das ISM-Tracking ein Tracking über Verdeckungen hinweg möglich. Dies ist Voraussetzung eines wesentlichen Aspekts bei der Objektverfolgung, dem korrekten Beibehalten von Objektidentitäten.

Diese korrekte Identitätswahrung während des Trackings ist wesentliche Grundlage für die *Personenwiedererkennung*, da hier die während des Trackings aufgebauten Merkmalsmodelle von Personen genutzt werden. Diese *online-Generierung von Merkmalsmodellen* trägt wesentlich dazu bei, dass das Gesamtsystem bis zur Wiedererkennung in realen Systemen einsetzbar ist – im Gegensatz zu den meisten anderen state-of-the-art Systemen, die insbesondere bei der Wiedererkennung von manuell annotierten Personen ausgehen. Ebenso wird hierdurch die Nutzung des Implicit Shape Model und SIFT auf die Personenwiedererkennung ausgedehnt, wodurch auch auf dieser Ebene die Systemgenerizität beibehalten wird. Dies wurde durch eine *Evaluierung* der Wiedererkennung sowohl im *sichtbaren*, als auch im *infraroten* Spektralbereich gezeigt. Für die Wiedererkennung selbst wurde ein effizienter *dreistufiger Ansatz* vorgestellt, der sowohl die *Codebuchsignaturen*, als auch die *ISM-Ausprägung* von Personen nutzt, um eine Wiedererkennung auf Stufen ansteigender Komplexität durchzuführen. In Auswertungen wurde gezeigt, dass die Stufen ansteigender Komplexität auch mit ansteigender Distinktivität verbunden sind und somit durch Anordnung in einer *Klassifikationshierarchie* am effizientesten eingesetzt werden können. Der Vergleich mit anderen Verfahren hat gezeigt, dass selbst die ersten beiden Stufen der Hierarchie, auf denen die Komplexität sehr gering ist, mit der Performance anderer Wiedererkennungsverfahren konkurrieren können, bzw. diese sogar übertreffen. Außerdem wurde gezeigt, dass mit der ISM-Wiedererkennung eine *open-set-Klassifikation*, welche die schwierigste Aufgabe für die Wiedererkennung darstellt, durchzuführen ist. Selbst bei diesen schwierigeren Bedingungen übersteigt die Performance der ISM-Wiedererkennung die anderer Verfahren bei weitem.

Zur Erhöhung der Ansichtsinvarianz der Wiedererkennung wurde ein Ansatz vorgestellt, der den *Aufbau von ansichtsspezifischen Personenmodellen* durch *automatische Bestimmung der Ansicht* einer Person während des Trackings ermöglicht. Diese ansichtsspezifischen Modelle bilden die Basis einer *Transformation von Personenmodellen*, welche Modelle unterschiedlicher Ansichten ineinander überführt. Anhand des Beispiels der Spiegelung wurde gezeigt, dass diese Transformation zur Erweiterung der Ansichtsinvarianz der Wiedererkennung beiträgt. Dies wurde in einer Auswertung auf Daten, bei denen Ansichtsveränderungen durch Laufrichtungsveränderungen der Personen in der Szene verursacht werden und die somit eine exakte Bewertung der Ansichtsinvarianz des Models erlauben, demonstriert. Hier wurde auch gezeigt, dass die Performance der ISM-Wiedererkennung die anderer Methoden zur Personenwiedererkennung auf diesem Datensatz übertrifft und selbst unter Ansichtsveränderungen noch teilweise bessere Performance aufweist als andere Verfahren ohne Ansichtsveränderungen.

Die Auswertung des Gesamtsystems in Multi-Kamera-Daten eines realen Flughafenszenarios hat gezeigt, dass das System bis hin zur Wiedererkennung unter schwierigen, realen Bedingungen einsetzbar ist. Diese schwierigen Bedingungen, die insbesondere für die Wiedererkennung eine Herausforderung darstellen, umfassen hier Verdeckungen zwischen Personen, große Veränderungen der Umgebungsvariablen zwischen den Kameras, sowie das Vorhandensein von die Erscheinung der Person beeinflussenden Objekten wie etwa mitgeführten Taschen. Der Vergleich mit anderen erscheinungsbasierten Methoden auf diesen Multi-Kamera-Daten hat gezeigt, dass die ISM-Wiedererkennung wie die Personenverfolgung mit spezialisierten Methoden in den jeweiligen Spezialfällen konkurrieren kann. Insbesondere ist anzumerken, dass die ISM-Wiedererkennung als einzige Methode direkt in das Tracking integriert ist und online Modelle aus dem Tracking verwendet, wohingegen bei anderen Methoden ein manueller Eingriff in Form der Nutzung manuell annotierter Daten oder eines offline-Trainingsschrittes erfolgt. Des weiteren ist der hier vorgestellte Ansatz die einzige erscheinungsbasierte Methode überhaupt, die sowohl die Funktionalität im infraroten, als auch im sichtbaren Spektralbereich unter Beweis stellt.

Ausblick

Das hier vorgestellte System ist vom Systemaufbau so gestaltet, dass es in realen Anwendungen vollautomatisch einsetzbar ist. Das bedeutet unter anderem, dass kein manueller Eingriff oder offline-Schritt im System erfolgt und das insbesondere das Tracking schritthaltend ausgelegt ist, d.h., es werden zu jedem Zeitpunkt nur Informationen aus der Vergangenheit genutzt.

Ein weiterer wesentlicher Aspekt beim realen Einsatz des Systems sind die *Laufzeitanforderungen*. Um wirklich einsetzbar zu sein, muss ein System auf aktueller Hardware in Echtzeit nutzbar sein. Dies ist im aktuellen Entwicklungsstand des Systems noch nicht gegeben aber Schwerpunkt aktueller Arbeiten. Da für die rechenintensiven Teilaspekte des Systems, wie die Merkmalsextraktion und den Merkmalsvergleich aber bereits Echtzeitrealisierungen existieren und verschiedene andere Teilaspekte wie die Personenwiedererkennung auf Algorithmenebene explizit effizient gestaltet wurden, kann angenommen werden, dass eine Echtzeitumsetzung des Systems möglich ist.

Durch die Verwendung des ISM ergibt sich in der Praxis bei der Anwendung in einem typischen Multi-Kamera-Szenario die Möglichkeit, die Verarbeitung aufzuspalten, um diese nicht für eine Vielzahl von Kameras an einem zentralen Punkt durchführen zu müssen. Dies wird durch aktuelle Hardwareentwicklung in Form sogenannter *intelligenter Kameras* möglich. Diese Kameras haben Verarbeitungshardware in Form eines FPGA (Field Programmable Gate Array) oder eines voll funktionsfähigen PCs integriert. Somit wird ermöglicht, bestimmte Verarbeitungen direkt auf der Kamera durchzuführen, und lediglich Metadaten – nicht die gesamten Bilddaten – über das Netzwerk zu einem zentralen Punkt zu übertragen. Beim hier verwendeten System bedeutet dies, dass die Personenverfolgung auf die einzelnen Kameras ausgelagert werden und nur noch Informationen über die Personentrajektorien in den einzelnen Kameras und eine Beschreibung der Personen durch die niedrigdimensionalen Codebuchsignaturen, bzw. ISM-Aktivierungen über das Netzwerk übertragen werden müssen. Diese ISM-Aktivierungen können an einem zentralen Punkt direkt in eine Datenbasis integriert werden. Neben dem Vorteil, dass keine Bilddaten mehr übertragen und gespeichert werden müssen, besteht der spezielle Vorteil der ISM-Beschreibung darin, dass diese direkt zur effizienten *Datenbankindizierung* verwendet werden kann. Auf Basis der Informationen in der Datenbasis kann dann die weitere Analyse durchgeführt werden. Diese muss dabei nicht nur die bereits erwähnte Multi-Kamera-Analyse auf Basis der durch die Wiedererkennung generierten Multi-Kamera-Trajektorien beinhalten, sondern kann auch weitere Formen annehmen.

Wie in [128] gezeigt wurde, ist es ebenfalls möglich, eine *2D-Posenrekonstruktion* auf Basis der ISM-Aktivierungen durchzuführen. Diese kann dann dazu genutzt werden, ausgeführte Aktionen zu bestimmen und eine Situationsanalyse durchzuführen. Wie in [96] bereits in Ansätzen gezeigt wurde, können neben den ISM-Aktivierungen auch die klassifizierten Körperteile zur Posenrekonstruktion genutzt werden. Hierbei bietet sich neben der 3D-Posenrekonstruktion auch die Möglichkeit, Basisaktionen wie „gehen“, „stehen“ und „laufen“ zu erkennen. Es sind allerdings auch noch Erweiterungsmöglichkeiten gegeben. So wurde in dieser Arbeit ein Verfahren zur Körperteildetektion vorgestellt, welches aber bisher nur Hinweise auf Körperteilpositionen in Einzelbildern liefert. Diese sind nicht in jedem Fall vollständig und können auch teilweise widersprüchlich sein – so können für eine Person z.B. mehrere Köpfe gefunden werden. An dieser Stelle sind also Erweiterungen sinnvoll, die z.B. durch das Einbringen von Wissen über den Aufbau von Personen Widersprüchlichkeiten beseitigen. Insbesondere sind auch hier, ähnlich wie bei der Personenverfolgung auf Objektebene, bei einer zeitlichen Betrachtung große Verbesserungen bzgl. der Stabilität und Zuverlässigkeit der Körperteilklassifikation zu erwarten. Eine solche zeitliche Betrachtung ist auch für weitere Interpretationsverfahren nützlich, da somit die Interpretation direkt auf Basis der *2D-Körperteiltrajektorien* durchgeführt werden kann.

Neben der Posenrekonstruktion und Aktionserkennung bieten sich noch weitere Nutzungsmöglichkeiten der ISM-Personenmodelle. So kann die ISM-Aktivierung dazu genutzt werden, *Objekte im Kontext anderer Objekte* zu erkennen. Dies geht über die bereits vorgestellte hierarchische Objekterkennung insofern hinaus, als dass hier auch Objekte betrachtet werden können, die nicht direkt im Bild als

solche zu identifizieren sind. Dies können z.B. von Personen geführte Waffen sein, die nicht als solche im Bild erkennbar sind, auf die aber im Kontext der Person aufgrund bestimmter Körperhaltungen geschlossen werden kann. Gleiches gilt z.B. für mitgeführte Taschen. Auch hier ist es möglich, dass diese im Bild nicht direkt detektierbar sind, eine Betrachtung der ganzen Person aber Rückschluss auf diese zulässt. Eine Möglichkeit, dies auf Basis der ISM-Aktivierung durchzuführen wurde in einer vom Autor betreuten Diplomarbeit [18] aufgezeigt. Hier werden Aktivierungsmodelle von Personen mit und ohne die relevanten Objekte aufgezeichnet um auf Basis dieser eine Klassifikation aktuell sichtbarer Personen vorzunehmen.

Neben den auf der ISM-Personenverfolgung aufbauenden Anwendungsmöglichkeiten gibt es auch direkte Erweiterungsmöglichkeiten bei den hier vorgestellten Algorithmen. So könnte der zur Ansichtsbestimmung durchgeführte Vergleich des Kurzzeit-Modells der aktuell verfolgten Person mit den Trainingsmodellen durch Einbringen dieser Information in den Trainingsschritt ersetzt werden. Hier könnten die bekannten Ansichten einer Person zur Annotation der Trainingsmerkmale genutzt werden. Die Bildmerkmale könnten so bei der Objektdetektion direkt für eine bestimmte Ansicht einer Person stimmen – somit wäre eine *direkte Ansichtsbestimmung* ohne zusätzlichen Vergleich mit Trainingsmodellen möglich.

Eine andere Möglichkeit, die auch gleichzeitig die Ansichtsinvarianz der Wiedererkennung verbessern könnte, wäre die Nutzung *ansichtsspezifischer Codebücher*. Obwohl dies zunächst wie ein Nachteil erscheint, da hier mehrere Codebücher für eine Objektklasse trainiert werden müssten und somit auch bei der Objektdetektion mit mehreren Codebüchern verglichen werden müsste, kann dies zu einer Verbesserung der Wiedererkennung beitragen. Hierzu können die in [155] beschriebenen Techniken zur Herstellung von Verbindungen zwischen Codebüchern und die Zusammenführung der Codebücher in ein globales Codebuch genutzt werden. Auf Basis der Verbindungen von Codebucheinträgen in den einzelnen Codebüchern könnte somit bei Aktivierung eines Eintrags in einer Ansicht sofort auf die Aktivierung in einer anderen Ansicht geschlossen werden. Hierdurch könnte über die Spiegelung hinaus eine Transformation auf Ebene der ISM-Aktivierungen eingeführt werden. Darüberhinaus ist es auch möglich, eine über die Spiegelung hinausgehende Transformation auf Basis der SIFT-Merkmale durchzuführen. Hierzu könnte die in [123] vorgestellte *affine Normalisierung* genutzt werden. Hierbei werden SIFT-Deskriptoren auf normalisierten Bildausschnitten berechnet. Die Informationen über die Transformationen der an einer Personenhypothese beteiligten Merkmale könnten dann auch zur Transformation des ISM genutzt werden.

Eine Möglichkeit, die Sensorunabhängigkeit des Systems weiter zu evaluieren, wäre die Übertragung des Systems auf 3D-Daten. Aktuelle Forschungsarbeiten beschäftigen sich mit dieser Thematik im Zusammenhang mit dem ursprünglichen ISM, wobei hier auf echte 3D-Modelle von Objekten abgezielt wird. Da diese in der Praxis nicht durch Sensoren zu akquirieren sind, ist eine anwendungsnähere Alternative die Nutzung von Tiefendaten einer *time-of-flight* Kamera. Es ist anzunehmen, dass das hier vorgestellte System ohne größere Adaptionen, bei Wahl geeigneter Merkmalsdeskriptoren, bzw. Adaption der vorhandenen, auch bei diesen Daten nutzbar ist, da auch hier Gradienteninformationen akquirierbar sind.

Das Gesamtsystem wurde so ausgelegt, dass es Generizität bzgl. verschiedener wichtiger Aspekte aufweist. Es wurde gezeigt, dass die Performance der einzelnen Systemebenen trotz dieser Generizität mit spezialisierten state-of-the-art Methoden, die diese Generizität nicht aufweisen, vergleichbar ist. Eine Möglichkeit, die Performance in spezifischen Szenarien weiter zu verbessern, wäre, das System auf diese Szenarien anzupassen und weiteres Wissen aus den Anwendungsbereichen zu nutzen. Sofern dieses Wissen in Form von austauschbaren, nicht systemrelevanten Einzelmodulen eingebracht werden kann, wird die Generizität des Basissystems nicht beeinträchtigt. Ein Beispiel für den Einsatz solchen speziellen Wissens ist z.B. die Nutzung eines bekannten Szenenhintergrunds bei stationären Kameras. Hier könnten neben den Codebüchern der relevanten Objektklassen auch Codebücher für den spezifischen Hintergrund aufgebaut werden und dazu dienen, die Detektionsperformance und damit auch die Performance des Gesamtsystems durch eine Reduzierung der Falschalarme weiter zu verbessern.

Literaturverzeichnis

[1] Casia gait database, http://www.sinobiometrics.com. obtained from http://www.cbsr.ia.ac.cn/english/gait 22, 79, 85, 86, 91

[2] Ec funded caviar project, url: http://homepages.inf.ed.ac.uk/rbf/caviar/, 2001. 60

[3] Pets 2006. 9th ieee international workshop on performance evaluation of tracking and surveillance. New York, USA, June 2006. (see http://www.cvg.rdg.ac.uk/PETS2006/index.html). 11, 60

[4] Winter-pets 2009. 12th ieee international workshop on performance evaluation of tracking and surveillance. New York, USA, December 2009. (see http://winterpets09.net/). 62

[5] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, USA, June 2008. 9, 14, 15, 48

[6] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, San Francisco, USA, 06/2010 2010. 9, 14

[7] D. Arsic, A. Lyutskanov, G. Rigoll, and B. Kwolek. Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2009. 62

[8] C. Arth, C. Leistner, and H. Bischof. Object reacquisition and tracking in large-scale smart camera networks. In *International Conference on Distributed Smart Cameras*, pages 156–163. IEEE, September 2007. 14

[9] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using haar-based and dcd-based signature. In *Proc. Advanced Visual and Signal based Surveillance*, pages 1528–1535, 2010. 13, 14, 99, 100, 101, 103

[10] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. Advanced Video and Signal based Surveillance*, pages 435–440, 2010. 13, 14, 102, 103

[11] S. Bak, S. Suresh, F. Bremond, and M. Thonnat. Fusion of motion segmentation with online adaptive neural classifier for robust tracking. In *Proc. Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 1–8, 02 2009. 99

[12] S. Baker, R. Gross, and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 3. *International Journal of Computer Vision*, 56:221–255, 2002. 6

[13] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Readings in computer vision: issues, problems, principles, and paradigms*, pages 714–725, 1987. 8, 29

[14] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proc. International joint conference on Artificial intelligence*, pages 659–663, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc. 8

[15] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. European Conference on Computer Vision*, pages 404–417, 2006. 18

[16] L. Bazzani, D. Bloisi, and V. Murino. A comparison of multi hypothesis kalman filter and particle filter for multi-target tracking. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2009. 62

[17] L. Bazzani, M. Farenzena, A. Perina, V. Murino, and M. Cristani. Multiple-shot person re-identification by hpe signature. In *Proc. International Conference on Pattern Recognition*, pages 1–8, 2010. 12, 14

[18] S. Becker. Erkennen von Objekten im Personenkontext. Master's thesis, Karlsruhe Institut für Technologie, 2010. 107

[19] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *Proc. European Conference on Computer Vision*, pages 45–58, 1997. 11

[20] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Transactions on Pattern Analysis and Machine Intelligence*, 24(24):509–522, April 2002. 18

[21] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *Proc. Computer Vision and Pattern Recognition*, pages 744–750, 2006. 6, 15

[22] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2009. 62, 63

[23] J. Berclaz, A. Shahrokni, F. Fleuret, J. M. Ferryman, and P. Fua. Evaluation of probabilistic occupancymap people detection for surveillance systems. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 55–62, 2009. 62

[24] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal of Image and Video Processing*, 2008:1–10, 2008. 54

[25] S. A. Berrabah, G. De Cubber, V. Enescu, and H. Sahli. Mrf-based foreground detection in image sequences from a moving camera. In *Proc. International Conference on Image Processing*, pages 1125–1128, 2006. 6

[26] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke. Pedestrian detection in infrared images. In *Proc. Intelligent Vehicles Symposium*, pages 662–667, June 9–11 2003. 10

[27] K. W. Bowyer, K. Hollingsworth, and P.J. Flynn. Image understanding for iris biometrics: A survey. *Computer Vision and Image Understanding*, 110(2):281–307, 2008. 11

[28] J. E. Boyd and J. J. Little. Biometric gait recognition. In *Biometrics School 2003, LNCS 3161*, pages 2–6, 2005. 11

[29] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Proc. International Conference on Computer Vision*, pages 1–8, October 2009. 9, 13, 15, 62

[30] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool. Markovian tracking by detection from a single uncalibrated camera. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 71–78, 2009. 62, 63

[31] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *Transactions on Pattern Analysis and Machine Intelligence. To appear.*, 2010. 9, 13, 15, 62

[32] F. Burkert, F. Schmidt, M. Butenuth, and S. Hinz. People tracking and trajectory interpretation in aerial image sequences. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS)*, volume XXXVIII, Part 3A, pages 209–214. ISPRS Commission III, September 1-2 2010. 3

[33] P.J. Burt and E.H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31(4):532–540, 1983. 21

[34] K. I. Chang, K. W. Bowyer, and P. J. Flynn. An evaluation of multi-modal 2d+3d face biometrics. *Transactions on Pattern Analysis and Machine Intelligence*, 27:619–624, 2005. 11

[35] C.H.Chuang, S.S. Huang, L.C. Fu, and P.Y. Hsiao. Monocular multi-human detection using augmented histograms of oriented gradients. In *Proc. International Conference on Pattern Recognition*, pages 1–4, 2008. 7

[36] Y. Cheng. Mean shift, mode seeking, and clustering. *Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995. 8, 30

[37] A. Colombo, J. Orwell, and S. Velastin. Colour Constancy Techniques for Re-Recognition of Pedestrians from Multiple Surveillance Cameras. In *Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008*, 2008. 12

[38] D. Comaniciu and P. Meer. A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 30

[39] C. O. Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. Smearson. Background modelling in infrared and visible spectrum video for people tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 20–25, june 2005. 6, 15

[40] T. Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *Image Analysis and Processing*, volume 5716 of *Lecture Notes in Computer Science*, pages 179–189. Springer Berlin, 2009. 11

[41] E. Corvee and F. Bremond. Combining face detection and people tracking in video sequences. In *International Conference on Imaging for Crime Detection and Prevention*, pages 151–161, 12 2009. 99

[42] C. Dai, Y. Zheng, and X. Li. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision Image Understanding*, 106(2-3):288–299, 2007. 10

[43] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 886–893, June 25–25, 2005. 7, 10, 18

[44] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, pages 7–13, 2006. 7

[45] J. Daugman. New methods in iris recognition. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1167–1175, September 2007. 11

[46] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2–3):162–182, 2007. 6, 10, 38, 55

[47] J. W. Davis and M. A. Keck. A two-stage template approach to person detection in thermal imagery. In *WACV/MOTION*, pages 364–369, 2005. 10

[48] J.W. Davis and V. Sharma. Robust background-subtraction for person detection in thermal imagery. In *Proc. Computer Vision and Pattern Recognition Workshop*, pages 128–136, 2004. 6, 15

[49] Y. Dufournaud, C. Schmid, and R.P. Horaud. Matching images with different resolutions. In *Proc. Computer Vision and Pattern Recognition*, pages 612–618. IEEE Computer Society Press, June 2000. 19

[50] A. Ellis, A. Shahrokni, and J.M. Ferryman. Pets2009 and winter-pets 2009 results: A combined evaluation. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2009. 62

[51] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *Transactions on Pattern and Machine Intelligence*, 31(12):2179–2195, December 2009. 6

[52] A. Ess, B. Leibe, and K. Schindler L. Van Gool. Robust multi-person tracking from a mobile platform. *Transactions on Pattern Analysis and Machine Intelligence*, 31:1831–1846, 2009. 9, 13, 15

[53] A. Ess, K. Schindler, B. Leibe, and L Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *International Journal of Robotics Research*, pages 1–8, 2010. 9, 15

[54] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki. Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection. In *Proc. Intelligent Vehicles Symposium*, pages 505–510, 2003. 10

[55] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, 2010. 12, 14

[56] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, 2008. 8

[57] P. F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010. 8, 9, 10

[58] P. F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 7, 9

[59] X. Fengliang and F. Kikuo. Pedestrian detection and tracking with night vision. In *Proc. Intelligent Vehicle Symposium*, volume 1, pages 21–30, June 17–21, 2002. 10

[60] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *Transactions on Pattern Analysis and Machine Intelligence*, 30(1):36–51, January 2008. 8

[61] V. Ferrari, T. Tuytelaars, and L. Van Gool. Integrating multiple model views for object recognition. In *Proc. Computer Vision and Pattern Recognition*, pages 105–112, 2004. 8

[62] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *Transactions on Computers*, 22(1):67–92, 1973. 7

[63] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2009. 6, 15

[64] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. European Conference on Computational Learning Theory*, pages 23–37, London, UK, 1995. Springer-Verlag. 7

[65] Junfeng G., Yupin L., and Gyomei T. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *Transactions Intelligent Transportation Systems*, 10(2):283–298, June 2009. 10

[66] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Proc. Conference Computer Vision and Pattern Recognition*, pages 1–8, 2009. 8

[67] S. Gammeter, A. Ess, T. Jäggli, K. Schindler, B. Leibe, and L. Van Gool. Articulated multi-body tracking under egomotion. In *Proc. European Conference on Computer Vision*, pages 816–830, 2008. 9, 14, 15

[68] T. Gandhi and M.M. Trivedi. Person tracking and reidentification: Introducing panoramic appearance map (pam) for feature representation. *Machine Vision Applications*, 18(3):207–220, 2007. 12

[69] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007. 9, 15

[70] N. Gheissari, T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, Los Alamitos, CA, USA, 2006. IEEE Computer Society. 12, 14, 70

[71] G.Kaur, A. Girdhar, and M. Kaur. Enhanced iris recognition system - an integrated approach to person identification. *International Journal of Computer Applications*, 8(1):1–5, October 2010. Published By Foundation of Computer Science. 11

[72] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. International Workshop on Performance Evaluation for Tracking and Surveillance*, pages 1–8, 2007. 12, 70, 100, 102

[73] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. European Conference on Computer Vision*, pages 262–275, Berlin, Heidelberg, 2008. Springer-Verlag. 12, 14

[74] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *International Conference on distributed smart cameras*, pages 1–6, September 2008. 12, 14

[75] I. Haritaoglu, D. Harwood, and L.S. Davis. W4s: A real-time system for detecting and tracking people in 2.5 d. In *Proc. European Conference on Computer Vision*, pages 877–886, 1998. 6

[76] C. Harris and M. Stephens. A combined corner and edge detection. In *Proc. Alvey Vision Conference*, pages 147–151, 1988. 22

[77] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979. 25

[78] S. Hinz. Density and motion estimation of people in crowded environments based on image sequences. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 38, Part 1-4-7/W5, on CD, 2009. 3

[79] P.V.C. Hough. Method and means for recognizing complex patterns. U.S. Patent 3069654, 1962. 29

[80] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 34(3):334–352, 2004. 6

[81] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proc. European Conference on Computer Vision*, volume 5303 of *Lecture Notes in Computer Science*, pages 788–801. Springer Berlin / Heidelberg, 2008. 9, 14, 15

[82] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise de Sciences Naturelles*, 4(3):223–370, 1908. 36

[83] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. *International Conference on Computer Vision*, 2:952–958, 2003. 12, 97

[84] O. Javed, K. Shafique, Z. Rasheed, and M Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008. 12, 97

[85] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. Computer Vision and Pattern Recognition*, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society. 11

[86] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. In *Proc. Int. Conference on Computer Vision, ICCV Workshops*, pages 1129–1136, 2009. 18

[87] K. Jüngling and M. Arens. Feature based person detection beyond the visible spectrum. In *Proc. Computer Vision and Pattern Recognition, CVPR Workshops*, pages 30–37, 2009. 17, 18, 33

[88] K. Jüngling and M. Arens. *Local Feature based Person Detection and Tracking Beyond the Visible Spectrum*. Springer, 2010. 18, 33, 59

[89] K. Jüngling and M. Arens. Local feature based person reidentification in infrared image sequences. In *Proc. International Conference on Advanced Video and Signal based Surveillance*, pages 448–454, 2010. 86

[90] K. Jüngling and M. Arens. Pedestrian tracking in infrared from moving vehicles. In *Intelligent Vehicles Symposium*, pages 470–477, 2010. 18

[91] K. Jüngling, M. Arens, M. Hanheide, and G. Sagerer. Fusion of perceptual processes for real-time object tracking. In *Proc. International Conference on Information Fusion*, pages 1139–1146, 2008. 15

[92] A. Kale, A.N. Rajagopalan, N. Cuntoor, V. Krueger, and R. Chellappa. Identification of humans using gait. *Transactions on Image Processing*, 13:1163–1173, 2002. 11

[93] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 48

[94] S.M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Proc. European Conference on Computer Vision*, pages 1–8, 2006. 6, 15

[95] J. Kittler and M. S. Nixon, editors. *Audio- and Video-Based Biometric Person Authentication*. Number 2688 in Lecture Notes in Computer Science. Springer-Verlag, 2003. 11

[96] V. Klinger and M. Arens. Ragdolls in action - action recognition by 3d pose recovery from monocular video. In *Proc. Computer Graphics, Visualization, Computer Vision and Image Processing*, pages 219–223, june 2009. 32, 106

[97] H.W. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistic Quarterly*, pages 83–97, 1955. 43

[98] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *Pattern Analysis and Machine Intelligence*, 31(12):2129–2142, dec. 2009. 7

[99] M. Lantagne, M. Parizeau, and R. Bergevin. Vip: Vision tool for comparing images of people. *Proc. Vision Interface*, pages 1–8, 2003. 12

[100] O. Lanz. Approximate bayesian multibody tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 28:1436–1449, 2006. 6, 15

[101] A. Lehmann, B. Leibe, and L. Van Gool. Fast prism: Branch and bound hough transform for object class detection. *International Journal of Computer Vision*, pages 1–8, 2010. 8, 18

[102] A. Lehmann, B. Leibe, and L.J. Van Gool. Prism: Principled implicit shape model. In *Proc. British Machine Vision Conference*, pages 1–8, 2009. 8, 18

[103] N. Lehment, D. Arsic, A. Lyutskanov, B. Schuller, and G. Rigoll. Statistical filters for crowd image analysis. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2009. 62

[104] B. Leibe. *Interleaved Object Categorization and Segmentation*. PhD thesis, 2004. 8

[105] B. Leibe, A. Ettlin, and B. Schiele. Learning semantic object parts for object categorization. *Image and Vision Computing*, 26(1):15–26, January 2008. 8, 14, 27, 32

[106] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on statistical learning in computer vision*, pages 17–32, Prague, Czech, May 2004. 8

[107] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008. 8, 10, 14, 17, 18, 25, 29, 41

[108] B Leibe, K Mikolajczyk, and B Schiele. Efficient clustering and matching for object class recognition. In *Proc. British Machine Vision Conference*, pages 1–8, September 2006. 25

[109] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1683–1698, 2008. 9, 10, 13, 14, 15

[110] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *Proc. International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brasil, October 2007. 9, 14, 15

[111] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. Computer Vision and Pattern Recognition*, pages 878–885, San Diego, USA, June 2005. 8, 17

[112] A. Leykin and R. Hammoud. Robust multi-pedestrian tracking in thermal-visible surveillance videos. In *Proc. Computer Vision and Pattern Recognition Workshop*, pages 136–142, June 17–22, 2006. 6, 10

[113] Z. Li, W. Bo, and R. Nevatia. Pedestrian detection in infrared images based on local shape features. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, June 17–22, 2007. 10, 14

[114] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, pages 224–270, 1994. 19, 20

[115] L.-F. Liu, W. Jia, and Y.-H. Zhu. Survey of gait recognition. In *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence*, volume 5755 of *Lecture Notes in Computer Science*, pages 652–659. Springer Berlin / Heidelberg, 2009. 11

[116] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157, Los Alamitos, CA, USA, August 1999. IEEE Computer Society. 19

[117] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 7, 18, 20, 21, 84

[118] R. Ma, J. Chen, and Z. Su. Mi-sift: mirror and inversion invariant generalization for sift descriptor. In *Proc. ACM Image and Video Retrieval*, pages 228–235, New York, NY, USA, 2010. ACM. 83

[119] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *Transactions on Pattern Analysis and Machine Intelligence*, 26:810–815, 2004. 6

[120] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *Proc. International Conference on Computer Vision*, volume 2, pages 1792–1799, October 17–21, 2005. 7, 18, 19

[121] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proc. Computer Vision and Pattern Recognition*, pages 26–36, 2006. 8

[122] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conference Computer Vision*, pages 128–142. Springer Verlag, 2002. 20

[123] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 18, 107

[124] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 19

[125] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. European Conference on Computer Vision*, pages 69–82, 2004. 8

[126] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005. 18, 19, 84

[127] E. Monari and K. Kroschel. Dynamic sensor selection for single target tracking in large video surveillance networks. In *Proc. Advanced Video and Signal Based Surveillance*, volume 0, pages 539–546, Los Alamitos, CA, USA, 2010. IEEE Computer Society. 12, 97

[128] J. Müller and M. Arens. Human pose estimation with implicit shape models. In *Proc. Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, pages 1–6, 2010. 106

[129] S. Munder, C. Schnörr, and D. M. Gavrila. Pedestrian detection and tracking using a mixture of view-based shape-texture models. *Transactions on Intelligent Transportation Systems*, 9(2):333–343, 2008. 9, 15

[130] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, pages 32–38, 1957. 43

[131] H. Nanda and L. Davis. Probabilistic template based pedestrian detection in infrared videos. In *Proc. Intelligent Vehicle Symposium*, volume 1, pages 15–20, June 17–21, 2002. 10

[132] C. Nandini and C.N. Kumar. Comprehensive framework to gait recognition. *International Journal of Biometrics*, 1(1):129–137, 2008. 11

[133] UK Home Office. ilids multiple camera tracking scenario definition, 2008. 95, 98

[134] K. Okuma, A. Taleghani, N. De Freitas, O. De Freitas, J.J. Little, and D.G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. European Conference on Computer Vision*, pages 28–39, 2004. 62

[135] S. Paisitkriangkrai, C. Shen, and J. Zhang. An experimental evaluation of local features for pedestrian classification. In *Proc. Digital Image Computing Techniques and Applications*, pages 53–60, December 3–5, 2007. 7

[136] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise:visual search engine using multiple networked cameras. In *Proc. International Conference on Pattern Recognition*, pages 1204–1207, August 2006. 11

[137] T.V. Pham, M. Worring, and A.W.M. Smeulders. A multi-camera visual surveillance system for tracking of reoccurrences of people. In *Proc. International Conference on In Distributed Smart Cameras*, pages 164–169, 2007. 11

[138] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition*, pages 947–954, 2005. 11

[139] F. Porikli. Inter-camera color calibration by correlation model function. In *Proc. International Conference on Image Processing*, pages 133–136, 2003. 12

[140] V. Prisacariu and I. Reid. Fasthog - a real-time gpu implementation of hog. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009. 7

[141] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106. 8

[142] Ying Ren, Chin-Seng Chua, and Yeong-Khing Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 24(1–3):183–196, 2003. 6

[143] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K.W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 27:162–177, 2005. 11

[144] C. Schmid, G. Dorko, S. Lazebnik, K. Mikolajczyk, and J. Ponce. *Pattern Recognition with Local Invariant Features*. 2004. 19

[145] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007. 8, 9

[146] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *Proc. British Machine Vision Conference*, pages 1–8, 2004. 7, 18

[147] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *Proc. Computer Vision and Pattern Recognition*, pages 1582–1588, New York, USA, June 2006. 8, 9

[148] P. K. Sharma, C. Huang, and R. Nevatia. Evaluation of people tracking, counting and density estimation in crowded environments. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 39–46, 2009. 62

[149] S. V. Sheela and P. A. Vijaya. Article:iris recognition methods - survey. *International Journal of Computer Applications*, 3(5):19–25, June 2010. Published By Foundation of Computer Science. 11

[150] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003. 14, 72

[151] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. *Proc. Computer Vision and Pattern Recognition*, 2:252 Vol. 2, August 1999. 6

[152] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In *Proc. Intelligent Vehicles Symposium*, pages 206–212, 2006. 10

[153] D. Tan, K. Huang, S. Yu, and T. Tan. Efficient night gait recognition based on template matching. In *International Conference on Pattern Recognition*, volume 3, pages 1000–1003, 2006. 11, 85, 90, 92

[154] S. Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. 7

[155] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, June 2006. 8, 84, 107

[156] D.-N. Truong Cong, L. Khoudour, C. Achard, and L. Douadi. People detection and re-identification in complex environments. *IEICE Transactions on Information and Systems*, 93:1761–1772, 2010. 11

[157] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 11

[158] T. Tuytelaars and K. Mikolajczyk. *Local Invariant Feature Detectors: A Survey*. Now Publishers Inc., Hanover, MA, USA, 2008. 19

[159] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *Proc. Workshop on Applications of Computer Vision*, pages 1–8, Snowbird, Utah, December 2009. 12

[160] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518 vol.1, 2001. 7, 10, 13

[161] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. Ninth IEEE International Conference on Computer Vision*, pages 734–741, October 13–16, 2003. 7

[162] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 25:1505–1518, 2003. 11, 85, 92

[163] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P.H. Tu. Shape and appearance context modeling. In *Proc. International Conference on Computer Vision*, pages 1–15, 2007. 12

[164] J. L. Wayman, A. K. Jain, D. Maltoni, and D. Maio. *Biometric Systems: Technology, Design and Performance Evaluation*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2004. 11

[165] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 51–56, October 14–16, 1996. 6

[166] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. International Conference on Computer Vision*, pages 90–97, Washington, DC, USA, 2005. IEEE Computer Society. 8

[167] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, November 2007. 9, 10, 13, 14, 15

[168] T. Xiaoyang, C. Songcan, Z. Zhi-hua, and Fuyan Z. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39:1725–1745, 2006. 11

[169] F. Yajun, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki. A shape-independent method for pedestrian detection with far-infrared images. *Transactions on Vehicular Technology*, 53(6):1679–1697, November 2004. 10

[170] J. Yang, Z. Shi, P. Vela, and J. Teizer. Probabilistic multiple people tracking through complex situations. In *Proc. International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 79–86, 2009. 62, 63

[171] M. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, Jan 2002. 11

[172] N-C. Yang, W.-H. Chang, C.-M. Kuo, and T. Li. A fast mpeg-7 dominant color extraction with new similarity measure for image retrieval. *Journal of Visual Communication and Image Representation*, 19(2):92–105, 2008. 99

[173] M. Yasuno, N. Yasuda, and M. Aoki. Pedestrian detection and tracking in far infrared images. In *Proc. Computer Vision and Pattern Recognition*, pages 125–132, 2004. 10

[174] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13–28, 2006. 6

[175] S. Yoon, S.S. Choi, S. Cha, Y. Lee, and C. Tappert. On the individuality of the iris biometric. *ICGST International Journal on Graphics, Vision and Image Processing*, V5:63–70, 2005. 11

[176] Y. Yu, D. Harwood, K. Yoon, and L. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications*, 18:139–149, 2007. 11

[177] Qiang Z., Mei-Chen Y., Kwang-Ting C., and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 1491–1498, 2006. 7

[178] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73:1–9, 2007. 7, 18

[179] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computating Surveys*, 35(4):399–458, December 2003. 11

[180] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. British Machine Vision Conference*, pages 1–8, 2009. 12, 14, 15, 70, 102, 103

[181] H. Zhou, Y. Yuan, and C. Shi. Object tracking using sift features and mean shift. *Computer Vision and Image Understanding*, 113(3):345–352, March 2009. 6