

Ein integraler stochastischer Ansatz zur automatischen Bestimmung von Personentrajektorien aus Luftbildsequenzen

Zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

von der Fakultät für

Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Dipl.-Ing. Florian Schmidt

aus Potsdam

Tag der mündlichen Prüfung: 23.11.2012

Referent: Prof. Dr.-Ing. Stefan Hinz

Korreferent: Prof. Dr.-Ing. Peter Reinartz

Korreferent: Prof. Dr.-Ing. Peter Vortisch

Karlsruhe 2013

Kurzfassung

Luftbilder stellen eine bedeutende Quelle vielfältiger Informationen über unsere Umwelt dar. In der Vergangenheit wurden sie hauptsächlich zur Beschreibung der Topographie genutzt. Moderne Kamerasysteme mit einer Aufnahmefrequenz von wenigen Hertz ermöglichen es, nun auch dynamische Prozesse großflächig zu beobachten. Für eine effektive Auswertung dieser Luftbildsequenzen werden automatische Methoden benötigt, die jedoch oftmals noch entwickelt oder an die spezifischen Herausforderungen angepasst werden müssen. Hier liegt die übergeordnete Zielsetzung dieser Arbeit. Im Detail beschäftigt sie sich mit der Fragestellung, in wie weit es möglich ist, Informationen über das Bewegungsverhalten von Personen aus Luftbildsequenzen zu gewinnen. Diese ließen sich z. B. zur besseren Koordination von Großveranstaltungen oder zur Evaluation von weiträumigen Infrastrukturanlagen einsetzen.

In dieser Arbeit wird daher eine Strategie entwickelt, umgesetzt und evaluiert, die es erstmalig ermöglicht, automatisch Einzelpersonen im Luftbild zu erkennen und ihre Bewegung durch eine Sequenz hinweg zu verfolgen. Die Auswertung beginnt für jede Aufnahme mit der Suche nach potentiellen Standorten von Personen, wofür ein aussehensbasierter Ansatz mit implizitem Modell verwendet wird. Dieser nutzt für die Detektion sowohl weiterentwickelte Bildmerkmale als auch den optional vorhandenen Personenschatten, welcher direkt in das visuelle Objektmodell integriert wird. Der Trainingsprozess zum Lernen dieses Modells wird dahingehend verbessert, dass beim automatischen Sammeln von Hintergrundbeispielen nun auch Bereiche in Objektnähe sowie der jeweilige Konfidenzwert mit berücksichtigt werden.

Da Personen in Luftbildern mit einer Bodenauflösung von etwa 15 cm pro Pixel nur sehr schwach zu erkennen sind, lässt sich eine robuste Detektion oft nicht auf Basis eines einzelnen Bildes durchführen. Aus diesem Grund wird in dieser Arbeit erstmals die Detektion mit implizitem visuellem Modell in das Multi-Hypothesen-Tracking-Verfahren (MHT) integriert. Hierfür werden die Ergebnisse der Objekterkennung stochastisch modelliert und der MHT-Formalismus entsprechend erweitert. Die Detektion von Personen erfolgt somit erst während des Trackings, wenn mehr Informationen zur Verfügung stehen und zuverlässigere Entscheidungen getroffen werden können. Das MHT-Verfahren wird ebenfalls weiterentwickelt, so dass sich die Vorteile der hypothesen- und trajektorienorientierten Variante gleichzeitig nutzen lassen. Zusätzlich wird eine neue Methode zur automatischen Bestimmung der Auftrittswahrscheinlichkeiten von Falschalarmen und neuen Objekten integriert, sowie das besonders wichtige Clusterverfahren durch eine verbesserte Datenstruktur stark vereinfacht.

Das entwickelte System bestimmt in anspruchsvollen Testsequenzen die Trajektorie von etwa einem Drittel aller Personen annähernd vollständig. Die Ergebnisse sind besonders gut in Bereichen mit geringer Objektdichte und hoher Erkennbarkeit. Bilden sich jedoch Gruppen oder gar Menschenmassen, sind Einzelpersonen visuell nicht mehr unterscheidbar und das Detektionsverfahren stößt an seine Grenzen. Das Auswertesystem müsste um weitere Module zur Behandlung dieser Phänomene ergänzt werden, um die einheitliche und vollständige Analyse eines gesamten Luftbildes zu ermöglichen.

Abstract

Aerial images are an important source of information about our environment. They have been used in the past mainly to describe the topography. Modern camera systems with a frame rate of a few hertz make it possible now, to also observe dynamic processes. Automatic methods are required to analyze these aerial image sequences effectively. However most of those methods still have to be developed or adapted to the specific challenges. This is the higher objective of this work. It also tries to answer the question, if it is possible to gain information about the motion and the behavior of people from aerial image sequences. This knowledge could be used e. g. to manage major events more securely or to evaluate the quality of large infrastructure.

A strategy is therefore being developed, implemented and tested in this work, which makes it possible for the first time to automatically detect single persons in an aerial image and track their movement through a sequence. The processing starts for each image with the search for potential object locations with an appearance-based approach with implicit object model. It uses improved image features for detection as well as the optional shadow of person, which is integrated directly into the appearance model. The training procedure to learn this model is also being improved by advancing the way to collect samples of the background class. Now it considers samples in object proximity as well as their individual confidence score.

Since persons are hardly visible in aerial images with a ground sampling distance of about 15 cm, a reliable detection based on a single image is often not possible. For this reason the appearance-based detection method is being integrated into the Multiple Hypotheses Tracking (MHT) approach. Therefore the results of object detections are being modeled stochastically and the MHT formalism is augmented appropriately. In this way the detection of people takes place during tracking, when more information is available and more reliable decisions are possible. The MHT approach is also being enhanced. Now it incorporates the advantages of the hypotheses-oriented as well as of the track-oriented version. Additionally a new method for automatic determination of the occurrence probability of false alarms and new objects is being integrated. The very important clustering procedure is also being simplified considerably by means of an improved data structure.

The developed system estimates the trajectories of about one third of all persons mostly correctly in challenging test sequences. The results are especially good in places with low object density and high visibility. Yet the detection algorithms fails when there are groups or crowds in which single persons can not be identified as individuals anymore. Additional modules have to be developed to deal with these situations, too. In this way a complete and homogeneous analysis of an entire aerial image would become possible.

Inhaltsverzeichnis

1. Einleitung	11
1.1. Motivation und Relevanz	11
1.2. Zielsetzung und Beiträge der Arbeit	12
1.3. Gliederung	13
2. Grundlagen und Konzeption	15
2.1. Grundlagen	15
2.1.1. Modellbildung	15
2.1.2. Objekterkennung	17
2.1.3. Objektverfolgung	18
2.2. Detektion und Tracking in Luftbildsequenzen	22
2.2.1. Rahmenbedingungen und Herausforderungen	22
2.2.2. Stand der Forschung	23
2.3. Diskussion und Auswertestrategie	25
2.3.1. Detektion	26
2.3.2. Tracking	27
3. Objekterkennung	29
3.1. Einzelpersonen in Luftbildern	29
3.2. Vorverarbeitung	31
3.2.1. Radiometrische Anpassungen	31
3.2.2. Geometrische Anpassungen	32
3.3. Detektor	32
3.3.1. Lokale Bildmerkmale	32
3.3.2. Form des Detektors	34
3.4. Klassifikator	35
3.4.1. Wahl des Klassifikators und Merkmalsreduktion	36
3.4.2. Beispiele sammeln und Klassifikator trainieren	37
3.5. Lokalisierung	38
3.6. Stochastische Modellierung	39
4. Objektverfolgung	41
4.1. Veränderlichkeit des Objektmodells	41
4.1.1. Aussehen	41
4.1.2. Position und Bewegung	41
4.2. Stochastische Bewertungsfunktion	43
4.2.1. Hypothesenwahrscheinlichkeit	43
4.2.2. Hypothesenwert und Track Management	44
4.2.3. Bestimmung der Wahrscheinlichkeitsdichten	45
4.2.4. Integration des Detektionswertes	48
4.3. Multi-Hypothesen-Tracking	49
4.3.1. Datenstruktur	49
4.3.2. Ablauf	50
4.3.3. Hypothesengenerierung	53
4.3.4. Clustern	54

Inhaltsverzeichnis

5. Evaluierung	57
5.1. Grundlagen	57
5.1.1. Bilddaten	57
5.1.2. Referenzdaten	58
5.1.3. Evaluierungsmaße	58
5.2. Objekterkennung	60
5.2.1. Personenschatten	60
5.2.2. Detektor	61
5.2.3. Klassifikator	65
5.2.4. Bewertung der Detektion	67
5.3. Objektverfolgung	68
5.3.1. Hypothesenanzahl	68
5.3.2. Auftrittswahrscheinlichkeit	70
5.3.3. Integration des Detektionswertes	71
5.3.4. Laufzeit	71
5.3.5. Bewertung des Trackings	73
6. Zusammenfassung und Ausblick	77
6.1. Zusammenfassung	77
6.2. Ausblick	79
Literaturverzeichnis	81
A. Anhang	91
A.1. Umformung der Hypothesenwahrscheinlichkeit	91

Tabellenverzeichnis

3.1. Ausgangskonfiguration des Personendetektors	35
5.1. Ausgewählte Merkmalsarten im Experiment zu Haar-Merkmalen	62
5.2. Ausgewählte Merkmalsarten im Experiment zu Rechteckmerkmalen	64
5.3. Varianten des Experiments zum Sammeln von Hintergrundbeispielen	66
5.4. Ergebnisse des Experiments zur Auftrittswahrscheinlichkeit von Falschalarmen und neuen Objekten	70
5.5. Ergebnisse des Experiments zur Nutzung des Detektionswertes im Tracking . . .	71
5.6. Ergebnisse der Objektverfolgung für vier Testsequenzen	73

Abbildungsverzeichnis

1.1. Luftbildaufnahmen zweier repräsentativer Szenarien	12
2.1. Möglichkeiten der Objektrepräsentation im Bild	16
2.2. Herausforderungen beim Erkennen und Verfolgen von Personen in Luftbildern .	22
2.3. Strategie zum Erkennen und Verfolgen von Einzelpersonen in Luftbildsequenzen	26
3.1. Detektionsverfahren mit implizitem visuellem Objektmodell	30
3.2. Variation des Aussehens von Einzelpersonen in Luftbildern	30
3.3. Einfluss der Objektdichte und des Schattens auf die Erkennbarkeit von Personen	31
3.4. Übersicht gebräuchlicher Arten von Haar-Merkmalen	33
3.5. Bildungsvorschrift der verallgemeinerten Haar-Merkmale	34
3.6. Dichte und kumulative Verteilung des Detektionswertes von korrekten Detektio- nen und Falschalarmen	40
4.1. Wahrscheinlichkeitsdichten für neu erkannte Objekte und Falschalarme	46
4.2. Beispiel des Tracking-Graphs	50
4.3. Ablaufdiagramm des MHT-Verfahrens	51
4.4. Tracking-Graph vor und nach der Auswahl einer DT-Zuordnungsoption	51
4.5. Schematische Darstellung des Zuordnungsproblems im MHT-Verfahren	54
5.1. Veranschaulichung der Bewertungsmaße	59
5.2. Konfiguration des Experiments zur Nutzung des Personenschattens	60
5.3. Ergebnisse des Experiments zur Nutzung des Personenschattens	61
5.4. Testfehler des Klassifikators im Experiments zur Nutzung des Personenschattens	61
5.5. Ergebnisse des Experiments zu Haar-Merkmalen	62
5.6. Einfluss der Detektorgröße auf die Detektionsleistung mit Rechteckmerkmalen .	63
5.7. Ergebnisse des Experiments zu Kombination verschiedener Bildmerkmale	64
5.8. Einfluss der Anzahl verwendeter Basisklassifikatoren auf den Testfehler	65
5.9. Einfluss der Anzahl verwendeter Basisklassifikatoren auf die Detektion	66
5.10 Ergebnisse des Experiments zum Sammeln von Hintergrundbeispielen im Training	66
5.11 Detektionsergebnisse für drei Testsequenzen	67
5.12 Ergebnisse des Experiments zur Hypothesenanzahl	69
5.13 Hypothesenwert in Abhängigkeit von der Hypothesenanzahl	69
5.14 Ergebnisse des Experiments zur Laufzeit des MHT-Verfahrens	72
5.15 Vergleich der Detektionsergebnisse nach Detektion und Tracking	74
5.16 Beispiele für beim Tracking auftretender Probleme	75

1. Einleitung

Mit Hilfe von Luftbildern kann man sich in kurzer Zeit einen Überblick über weiträumige Gebiete verschaffen. Dieser Umstand erlaubt es, vielfältige Fragestellungen zu beantworten, welche sich vom Boden aus nicht oder nur mit sehr großem Aufwand klären ließen. Dieser als *Luftbildanalyse* bezeichnete Teilbereich der Fernerkundung liefert seit vielen Jahrzehnten sehr erfolgreich Ansätze und Lösungen für praktische und wissenschaftliche Problemstellungen. Der Schwerpunkt liegt hierbei oft auf einer Beschreibung bzw. Kartierung der Landschaft als Ganzes oder ausgewählter topographischer Objektklassen wie Straßen, Häuser, Wälder oder landwirtschaftliche Flächen. Neben der rein statischen Analyse von Luftbildern eines bestimmten Zeitpunktes, ermöglichen wiederholte Überflüge im Abstand von Monaten oder Jahren die Änderung der Topographie zu beschreiben und deren Ursachen zu ergründen.

1.1. Motivation und Relevanz der Arbeit

Moderne digitale Luftbildkameras wie Leicas *RCD30* oder Microsofts *UltraCam* werden seit der Jahrtausendwende operationell eingesetzt und besitzen eine maximale Aufnahmezeit von ein bis zwei Bildern pro Sekunde (Microsoft, 2011; Leica Geosystems, 2012). Neuartige Entwicklungen wie das 3K-Kamerasystem vom DLR erhöhen die Bildfrequenz sogar noch weiter auf wenige Hertz (Thomas u. a., 2008). Dadurch wird es möglich, Prozesse zu beobachten, die sich in sehr kurzen Zeiträumen in einem weiträumigen Gebiet abspielen. Von großem Interesse ist hier vor allem das Bewegungsverhalten von Objekten auf der Erdoberfläche, wie Fahrzeuge, Tiere oder Personen. Ließen sich diese bisher meist nur an bestimmten Orten und in einem relativ kleinen Bereich, z. B. mit Hilfe von Videokameras beobachten, so können diese Beschränkungen mit Hilfe von Luftbildsequenzen vollständig aufgehoben werden. Objektbewegungen lassen sich nun weiträumig und an beliebigen Orten erfassen.

Die große Anzahl an Bildern und beobachtbaren Objekten verlangt nach automatisierten Auswertungsverfahren (vgl. Abb. 1.1). Die hierfür notwendigen Methoden zum Erkennen und Verfolgen von Objekten in Luftbildsequenzen müssen jedoch meist noch entwickelt werden. Obwohl schon etliche Ansätze für die Auswertung von terrestrischen Videoaufnahmen existieren, lassen sich diese nicht ohne Weiteres auf die speziellen Anforderungen in der Fernerkundung übertragen. Vor allem das ungünstige Verhältnis von Objektgröße zu Bildauflösung und die hohe Komplexität der Bildinhalte stellen die Auswertung vor zusätzliche Herausforderungen. Indem in der vorliegenden Dissertation ein System zur automatischen Extraktion von Personenbewegungen aus Luftbildsequenzen entwickelt wird, leistet sie einen wichtigen Beitrag zum Schließen dieser Lücke.

Neben der zu entwickelnden Methodik besitzen auch die angestrebten Ergebnisdaten eine praktische und wissenschaftliche Relevanz. Die Möglichkeit homogene Informationen über das Bewegungsverhalten von sehr vielen Personen in einem großräumigen Gebiet zu erhalten, kann dazu dienen, Großveranstaltungen besser zu koordinieren und sicherer durchzuführen als dies bisher möglich war. In diesem Zusammenhang ist es auch denkbar, gefährliche Szenarien automatisch zu erkennen und lokale Einsatzkräfte darauf hinzuweisen. Des Weiteren sind die Daten auch geeignet, um die Qualität großräumiger Infrastrukturanlagen wie Fußballstadien, Messegelände oder ganze Innenstädte zu bewerten. Typische Laufwege oder ungünstige Engstellen können erkannt und die Aufenthaltsqualität und Sicherheit verbessert

1. Einleitung

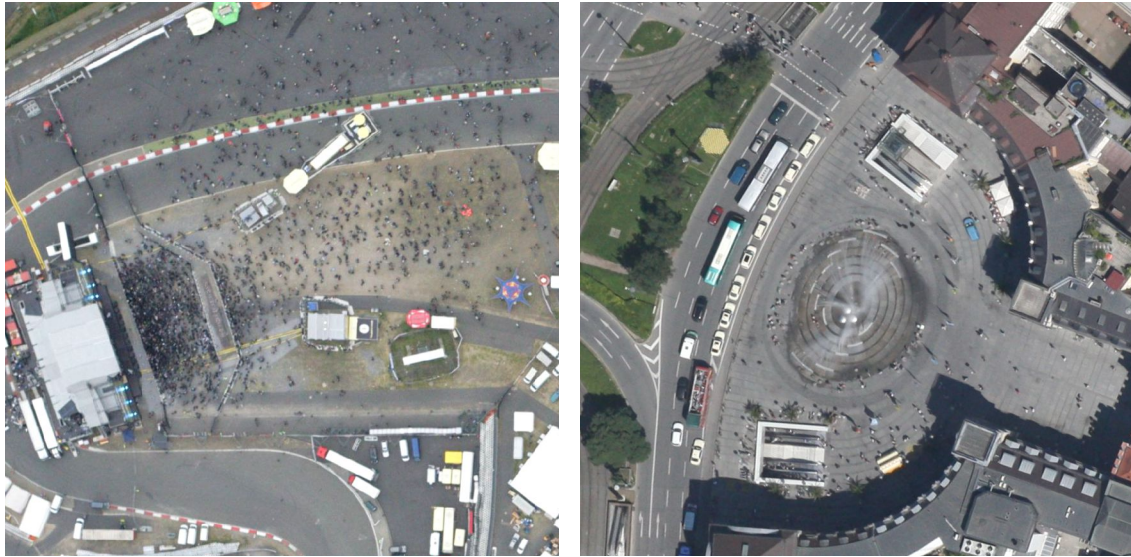


Abbildung 1.1.: Ausschnitte zweier Luftbildaufnahmen von repräsentativen Szenarien, wie sie in dieser Arbeit betrachtet werden. Das linke Bild zeigt den Bühnenbereich eines Freiluftkonzertes und das rechte einen Platz in einer Fußgängerzone.

werden. Darüber hinaus können auch andere wissenschaftliche Bereiche profitieren, die auf Informationen über das Bewegungsverhalten von Personen angewiesen sind. Die Möglichkeit automatisch eine Vielzahl von Bewegungsdaten zu gewinnen, lässt sich nutzen, um bestehende Modelle zu verbessern oder neue Erkenntnisse zu gewinnen.

1.2. Zielsetzung und Beiträge der Arbeit

Die vorliegende Arbeit behandelt die Frage, in wie weit es möglich ist, Informationen über das Bewegungsverhalten von Personen aus Luftbildsequenzen zu gewinnen. Hierfür soll eine Strategie entwickelt und umgesetzt werden, die automatisch einzelne Personen im Luftbild erkennt und deren Bewegung durch die Sequenz hinweg verfolgt. Als Ergebnis sollen möglichst vollständige Trajektorien aller sich in der Szene befindlichen Personen ausgegeben werden. Diese objektzentrierte, mikroskopische Auswertung ermöglicht eine detaillierte Analyse der Bewegung einzelner Personen und ihrer Interaktionen. Sie steht im Gegensatz zu einer flächenhaften, makroskopischen Auswertung, welche in dieser Arbeit nicht näher betrachtet wird. Hierbei werden Angaben über die Objektdichte und das allgemeine Bewegungsverhalten flächenhaft ermittelt, ohne dass Einzelobjekte explizit erkannt werden müssen (vgl. Zhan u. a. (2008); Jacques Junior u. a. (2010)).

Der Schwerpunkt dieser Dissertation liegt somit in der Konzeption, Umsetzung und Evaluierung einer effizienten Strategie zum Erkennen und Verfolgen von Einzelpersonen in Luftbildsequenzen. Hierfür werden die notwendigen Methoden entwickelt bzw. adaptiert unter Berücksichtigung der spezifischen Herausforderungen, wie einem ungünstigen Verhältnis von Objektgröße zu Bildauflösung, einer potentiell sehr großen Anzahl ähnlicher, merkmalsarmer Objekte und einer sich bewegenden Kameraplattform.

Der größte wissenschaftliche Beitrag liegt in der Entwicklung eines integralen, stochastischen Ansatzes für die Objekterkennung und -verfolgung. Dieser kombiniert im Rahmen einer Tracking-by-Detection-Strategie erstmals die beiden mächtigen Methoden der aussehensbasierten Detektion mit implizitem Objektmodell und des Multi-Hypothesen-Trackings (MHT). Bisher führte in vielen Verfahren der Detektionsschwellwert am Ende der Objekter-

kennung zu einer scharfen Trennung von Detektion und Tracking. Durch eine stochastische Beschreibung der Detektionsergebnisse wird der große Einfluss dieser binären, unumkehrbaren Entscheidung jedoch erheblich abgeschwächt und verschiebt die Objekterkennung in den Bereich des Trackings, wo mehr Informationen verfügbar sind und eine fundiertere Aussage getätigt werden kann. Von diesem Vorgehen können alle gleichartigen Tracking-Verfahren profitieren. Der Ansatz ist jedoch besonders in solchen Szenarien geeignet, in denen sich die gesuchten Objekte nur schwer vom Hintergrund abheben und es zu vielen Mehrdeutigkeiten kommt.

Darüber hinaus enthält die vorliegende Arbeit weitere wissenschaftliche Beiträge entlang der gesamten Prozesskette von den Luftbildern bis hin zu den Trajektorien. So wird gezeigt, wie Personen in Luftbildern effektiv mit Hilfe eines impliziten, visuellen Objektmodells erkannt werden können. Der hierfür entwickelte Detektor berücksichtigt ausdrücklich den optional vorhandenen Personenschatten und verwendet u. a. verallgemeinerte Haar-Merkmale, welche speziell auf die charakteristische Form der gesuchten Objektklassen anpassbar sind. Des Weiteren wird das Training des Detektors durch ein robusteres Verfahren zur automatischen Auswahl von Hintergrundbeispielen verbessert. Für die Objektverfolgung werden erstmals die Vorteile der hypothesen- und trajektorienorientierten Variante des MHT-Ansatzes kombiniert. Zusätzlich wird eine neue Methode zur integrierten, adaptiven Bestimmung der Falschalarm- und Objektauftrittswahrscheinlichkeit entwickelt und das für den MHT-Ansatz äußerst wichtige Clusterverfahren durch eine neue Datenstruktur zur Verwaltung der verschiedenen Hypothesen stark vereinfacht. Zudem wird gezeigt, dass der MHT-Ansatz auch für sehr viele Objekte in Echtzeit eingesetzt werden kann.

1.3. Aufbau der Arbeit

Diese Arbeit ist folgendermaßen gegliedert: In Kapitel 2 werden zuerst grundlegende Strategien und Methoden zur Objekterkennung und -verfolgung beschrieben. Anschließend werden die speziellen Herausforderungen für die Auswertung von Luftbildsequenzen dargelegt und ein umfassender Überblick gegeben, wie diese in bisherigen Veröffentlichungen zum Thema bzw. in anderen Bereichen mit ähnlichen Problemstellungen behandelt worden sind. Es folgen eine abschließende Diskussion der vorhandenen Verfahren und die Ableitung einer eigenen Auswertestrategie.

Kapitel 3 beschäftigt sich mit der Lokalisierung von Personen in einzelnen Luftbildern und erläutert alle Aspekte des Detektionsansatzes mit implizitem visuellem Objektmodell. Es beginnt mit Überlegungen zur äußeren Erscheinung von Personen, wichtigen Vorverarbeitungsschritten und geeigneten Bildmerkmalen. Es folgen Methoden zum robusten Trainieren des Detektors, zur subpixelgenauen Lokalisierung der Personen und zur stochastischen Modellierung der Detektionsergebnisse.

In Kapitel 4 wird gezeigt, wie sich auf Basis der Detektionen im Einzelbild mit Hilfe des MHT-Verfahrens Trajektorien erzeugen lassen. Nach einigen Anmerkungen zum Objektmodell wird die für das Tracking notwendige stochastische Bewertungsfunktion inklusive der vorgeschlagenen Anpassungen ausführlich behandelt. Anschließend wird dargelegt, wie sich der MHT-Ansatz effizient in Echtzeit ausführen lässt und welche Rolle hierbei die entwickelte Datenstruktur spielt.

Nach der vollständigen Darstellung der entwickelten Methodik, wird diese in Kapitel 5 umfassend validiert. Hierfür werden sowohl einzelne Komponenten als auch die Module Objekterkennung und -verfolgung als Ganzes in diversen Experimenten untersucht und deren Ergebnisse diskutiert.

Abschließend werden in Kapitel 6 Inhalt und Ergebnisse dieser Arbeit zusammengefasst und ein Ausblick über zukünftige Forschungsmöglichkeiten formuliert.

2. Grundlagen und Konzeption

In diesem Kapitel werden die Grundlagen der Arbeit dargelegt. Zuerst erfolgt eine allgemeine Einführung in die Bereiche Objekterkennung und -verfolgung. Die gesamte Thematik wird strukturiert dargestellt und unterschiedliche Strategien mit ihren zugehörigen Methoden werden benannt. Der zweite Teil dieses Kapitels beschäftigt sich dann konkret mit dem Thema Personenverfolgung in Luftbildsequenzen. Zuerst werden die speziellen Herausforderungen in diesem Bereich aufgeführt. Anschließend erfolgt ein ausführlicher Überblick über bisheriger Veröffentlichungen zum Thema und eine Einordnung der eigenen Arbeit. Darüber hinaus werden auch Arbeiten aus anderen Bereichen vorgestellt, bei denen sich ähnliche Herausforderungen stellen. Die Literatur- und Methodenrecherche mündet in der Entwicklung einer geeigneten Auswertestrategie.

2.1. Grundlagen zum Erkennen und Verfolgen von Objekten

In dieser Arbeit werden Luftbildsequenzen zur Gewinnung von Personentrajektorien genutzt. Von den drei grundlegenden, aufeinander aufbauenden Arbeitsschritten der objektbezogenen Videoanalyse: Objekterkennung, Objektverfolgung und Verhaltensanalyse (Yilmaz u. a., 2006), werden nur die ersten beiden behandelt. Eine genauere Betrachtung des Personenverhaltens auf Basis der gewonnenen Bewegungsinformationen ist nicht Gegenstand dieser Arbeit.

Nachfolgend werden die theoretischen Grundlagen der Bereiche Objekterkennung und Objektverfolgung zusammengefasst dargestellt. Der Fokus liegt auf einer strukturierten Übersicht vorhandener Methoden und Konzepte, losgelöst von einer konkreten Objektart. Einen guten Überblick und Einstieg in diese Thematik geben auch die folgenden Publikationen: Neumann (2003); Yilmaz u. a. (2006), sowie mit Schwerpunkt Objektverfolgung: Bar-Shalom und Fortmann (1988); Blackman und Popoli (1999).

2.1.1. Modellbildung

Die Bildanalyse wird in vielen Anwendungsgebieten für die unterschiedlichsten Objektarten erfolgreich eingesetzt. Grundvoraussetzung ist dabei immer, dass ein geeignetes Objektmodell vorliegt, welches auf die zur Verfügung stehenden Detektion- und Tracking-Methoden abgestimmt ist.

Grundsätzlich muss genau definiert sein, welche Art von Objekten der realen Welt überhaupt erkannt werden sollen und auf Basis welcher Informationen dies geschehen soll. Daraus ergibt sich das in der Bildanalyse genutzte Objektmodell. Je nach Schwierigkeit der Aufgabenstellung variieren Art und Umfang der benötigten Informationen, was sich direkt auf die Komplexität des Objektmodells und der verwendeten Methoden auswirkt.

In jedem Fall muss untersucht werden, wie sich die gesuchte Objektart unter der zu erwartenden Aufnahmekonfiguration (Sensormodell) in den Bilddaten darstellt. Form und Aussehen sowie deren mögliche Variation haben großen Einfluss auf das Detektionsverfahren und die Objektrepräsentation (s. Abb. 2.1). Sehr kleine Objekte mit starrer Form können z. B. al-

2. Grundlagen und Konzeption

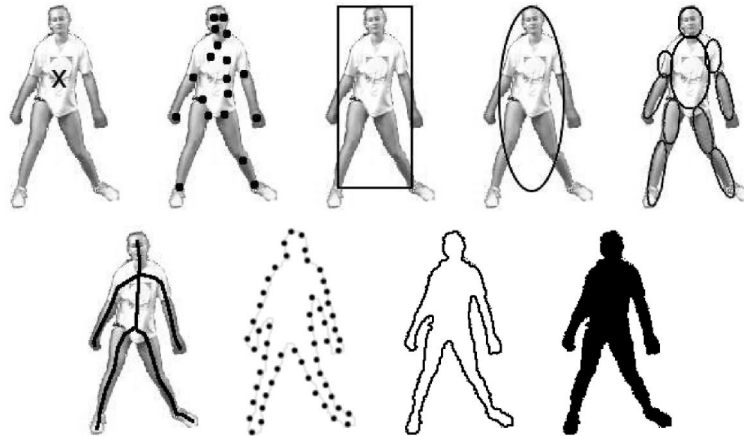


Abbildung 2.1.: In der Grafik aus (Yilmaz u. a., 2006) werden am Beispiel einer Person neun verschiedene Möglichkeiten aufgezeigt, wie ein Objekt im Bild repräsentiert werden kann.

lein durch ihren Mittelpunkt repräsentiert werden (Veenman u. a., 2001; Shafique und Shah, 2005), wohingegen für größere Objekte mit kompakter Struktur häufig einfache geometrische Formen wie Ellipsen oder Rechtecke genutzt werden (Viola u. a., 2005; Grabner u. a., 2008). Darüber hinaus lassen sich Objekte auch über ihren Umriss oder als Segment darstellen (Li u. a., 2008). Dieses Vorgehen bietet sich besonders dann an, wenn die Objektform in den Bildern stark variiert, sich jedoch gut vom Hintergrund abhebt. Ist das äußere Erscheinungsbild sehr komplex oder partitionierbar, so kann das gesuchte Objekte auch aus miteinander verbundenen Teilen dargestellt werden (Wu und Nevatia, 2007; Jüngling, 2011; Ulrich, 2003). Hierbei werden in einem hierarchischen Verfahren die Komponenten zuerst einzeln detektiert und in einer nachfolgenden Stufe entsprechend der erlaubten Konfiguration zusammengefügt.

Neben diesem rein auf das äußere Erscheinungsbild ausgerichtete Teil, kann das Objektmodell auch noch diverse weitere Eigenschaften bzw. Zustände umfassen, die für die jeweilige Anwendung interessant sind. Häufig lassen sich diese jedoch nur im Zuge einer Sequenzanalyse, wie etwa die Bewegungsrichtung oder in Relation zur Umgebung ermitteln.

Wenn das Objekt über einen längeren Zeitraum beobachtet werden soll, ist es notwendig festzuhalten, wie sich die Objekteigenschaften mit der Zeit verändern können (*dynamic model*). Ein Bewegungsmodell zur Beschreibung der Veränderung von Position, Geschwindigkeit und Richtung ist für die Objektverfolgung beispielsweise unerlässlich.

Auch kann ein Kontextmodell für die Detektion sehr hilfreich sein, vor allem bei komplexen Szenen und mehrdeutigen Objekten (Biederman u. a., 1982; Suetens u. a., 1992; Galleguillos und Belongie, 2010). Dieses beschreibt die Art und Wahrscheinlichkeit der lokalen und globalen Zusammenhänge (Relationen) in denen das gesuchte Objekt erscheint. Es kann z. B. die Wahrscheinlich enthalten, mit welcher ein Objekt in einer bestimmten Umgebung auftritt oder wie stark dessen Größe in Bezug zu anderen Objekten variieren darf. Obwohl die Modellierung des Kontextes sehr nützlich sein kann, so muss beachtet werden, dass die entsprechende Relation gegeben bzw. ebenfalls aus den Bilddaten ermittelt werden muss. In einigen Anwendungen sind Kontextinformationen in Form eines Szenenmodells vorhanden. Dieses ermöglicht die Nutzung von räumlichen Kontext für die Objekterkennung und erleichtert die Objektverfolgung, da Vorabinformationen über wahrscheinliche Bewegungsrichtungen verwendet werden können. Als Beispiel sei hier ein Straßennetz zum Erkennen und Verfolgen von Fahrzeugen genannt.

2.1.2. Objekterkennung

Das Ziel der Objekterkennung besteht darin festzustellen, ob und wenn ja wo sich das gesuchte Objekt im Bild befindet. Man unterscheidet hierbei zwischen Erkennen und Identifizieren. Im ersten Fall sucht man alle Instanzen einer Objektart, z. B. Fahrzeuge oder Personen, im zweiten Fall dagegen nur eine ganz bestimmte Instanz der Klasse. Die jeweils verwendeten Methoden sind jedoch identisch, für die Identifikation muss lediglich das Objektmodell individualisiert werden, so dass eine Unterscheidung innerhalb der Objektart möglich wird.

Für die Objekterkennung stehen zahlreiche Methoden zur Verfügung. Welche für die jeweilige Aufgabenstellung und Objektart am besten geeignet ist, hängt stark vom Modell und der Darstellung der Objekte im Bild ab. Die verschiedenen Verfahren lassen sich grob in zwei Klassen unterteilen. Je nach Schwerpunkt sind sie entweder daten- oder modellgetrieben. Im ersten Fall gehen sie von den Pixeln bzw. Messwerten aus (*bottom-up*). Benachbarte Pixel mit ähnlichen Eigenschaften im Merkmalsraum werden schrittweise zu größeren Strukturen, in der Regel flächenhafte Segmente, zusammengefasst. Anschließend erfolgt durch Klassifikation derselben die finale Objekterkennung. Bei modellgetriebenen Verfahren geht man den umgekehrten Weg (*top-down*). Hier liegt ein Modell des Objektes vor, welches direkt im Bild auf Basis der Pixel oder abgeleiteter Merkmale gesucht wird. Die Detektion erfolgt immer dann, wenn die Ähnlichkeit mit dem Modell ein gewisses Maß überschreitet. Häufig werden beide Strategien zusammen in einem hybriden Ansatz genutzt. So können z. B. datengetriebene Methoden genutzt werden, um den Suchbereich für modellgetriebene Verfahren einzuschränken. Nachfolgend werden beide Varianten inklusive einiger zugehöriger Methoden näher erläutert.

Datengetriebenen Verfahren

Zur Gruppierung von Pixeln in meist flächenhafte Segmente werden vielfältige Ansätze genutzt. Neben klassischen Bildverarbeitungsmethoden wie Regionenwachstumsverfahren oder dem Wasserscheidenalgorithmus werden auch modernere Segmentierungsverfahren basierend auf *Mean-Shift Clustering* (Comaniciu und Meer, 2002), *Graph-Cuts* (Rother u. a., 2004) oder *Markov Random Fields* (Jodoin u. a., 2007) eingesetzt.

Ein sehr häufig genutzter Ansatz zur datengetriebenen Objekterkennung basiert jedoch auf der Modellierung des Hintergrundes (Hu u. a., 2004). Pixel, welche diesem Modell nicht genügen, zählen zum Vordergrund und können als Segmente zusammengefasst weiter prozessiert werden. Dieses Verfahren ist besonders dann geeignet, wenn der Hintergrund einfacher zu beschreiben ist als die gesuchte Objektart oder alle bewegten Objekte in der Szene bestimmt werden sollen (*moving-object detection*). Statt den Hintergrund explizit zu modellieren, wird meist die Annahme getroffen, dass dieser für einen gewissen Zeitraum statisch ist bzw. sich lokal nur monoton verändert. Die Grauwerte der Hintergrundpixel lassen sich dann z. B. einzeln durch eine Gauß'sche Mischverteilung beschreiben. Bewegt sich nun ein Objekt durch die Szene, kann die Unterteilung in Vorder- und Hintergrundpixel einfach und schnell durchgeführt werden. Schwierigkeiten ergeben sich immer dann, wenn die gesuchten Objekte dem Hintergrund ähneln, wenn kein robustes Modell des Hintergrundes bestimmt werden kann und wenn durch eine bewegte Kamera oder Beleuchtungsänderungen der Hintergrund nicht mehr statisch erscheint. Neben der Modellierung der Grauwertverteilung kann die Trennung von Hintergrund und Objekten auch auf andere Weise geschehen. In (Xiao u. a., 2008a) wird z. B. ein Video vom Flugzeug aus genutzt, um Disparitätsbilder zu berechnen und erhöhte Bereiche von der Suche nach Fahrzeugen auszuschließen. Je nachdem wie vollständig das gewählte Modell den Hintergrund beschreibt, wird das jeweilige Verfahren allein oder als erste Stufe einer Klassifikationskaskade eingesetzt. In beiden Fällen sollte der Schlupf jedoch möglichst gering sein.

2. Grundlagen und Konzeption

Nach der Segmentierung werden die gebildeten Strukturen meist morphologisch nachbearbeitet und in einem Klassifizierungsverfahren bewertet. Hierfür stehen nun neben Pixelmerkmalen auch zusätzliche Form- und Texturmerkmale wie Flächeninhalt oder Grauwertvarianz zur Verfügung. Mit entsprechenden Verfahren aus der Mustererkennung (Jain u. a., 2000; Duda u. a., 2001) erfolgt die abschließende Einteilung der Segmente in Objekte oder Hintergrund.

Modellgetriebene Verfahren

Bei modellgetriebenen Verfahren zur Detektion steht die effektive Suche mittels eines Objektmodells im Vordergrund. Je nach Art des Modells wird die Suche direkt auf Basis der Pixel oder abgeleiteter Merkmale durchgeführt. Liegt eine ausreichend hohe Übereinstimmung vor, gilt das Objekt an der jeweiligen Stelle als gefunden.

Ist das Objektmodell explizit z. B. in Form eines Drahtgittermodells vorhanden (Suetens u. a., 1992; Hinz, 2003), muss dieses zuerst in das Bild projiziert und kann dann mit extrahierten Kanten verglichen werden. Je nach Anzahl und Variabilität der Modellparameter kann die Suche sehr aufwendig sein. Sie mündet daher häufig in einem Optimierungsverfahren, in welchem die besten Übereinstimmungen zwischen Modell und Bilddaten mit effektiven Methoden gesucht werden (Lafarge u. a., 2008; Vidal u. a., 2006; Ge und Collins, 2009).

Liegt das Objektmodell dagegen nicht parametrisiert vor, sondern implizit in Form von Pixelwerten oder Bildmerkmalen, so entfällt der sonst notwendige Projektionsschritt. Auch wird das Modell nicht wie bisher mittels Regeln und Bedingungen definiert, sondern ergibt sich direkt aus einer gewissen Anzahl von Beispielbildern. *Template Matching* ist z. B. eines der einfachsten impliziten Verfahren, bei dem das Objektmodell in Form eines repräsentativen Bildausschnittes vorliegt. Die Suche erfolgt hier, indem der Ausschnitt als gleitendes Fenster über das Bild geschoben wird. Mittels normalisierte Kreuzkorrelation lässt sich dann für jede Position eine Ähnlichkeit berechnen. Variieren Größe und Ausrichtung der gesuchten Objekte, so muss die Suche mehrfach mit transformierter Bildmaske durchgeführt werden. Aufwändiger aber deutlich robuster sind modernere Verfahren bei denen das implizite Modell nicht direkt auf Pixeln sondern auf lokalen Bildmerkmalen wie Histogrammen (Dalal und Triggs, 2005), Kanten- oder Texturmerkmalen (Ojala u. a., 2002; Oren u. a., 1997) basiert. Ein Klassifikator lernt dann anhand zahlreicher Beispiele, Objekte und Hintergrund von einander zu trennen. Die eigentliche Suche im Bild funktioniert dann ähnlich dem zuvor beschriebenen Verfahren, nur dass die zum Vergleich benötigten Merkmalswerte innerhalb des Detektionsfensters extrahiert und anschließend zur Bewertung dem Klassifikator zugeführt werden müssen (vgl. Viola und Jones (2001)). Mit dem *Implicit Shape Model* (Leibe u. a., 2008) oder dem hierarchischen Ansatz aus (Ulrich, 2003) stehen darüber hinaus weitere Verfahren zur Verfügung, welche die Konzepte der impliziten Modellierung auch auf komponentenbasierte Objektbeschreibungen erweitern.

Vergleicht man explizite und implizite Ansätze miteinander, so sind erstere besonders dann geeignet, wenn sich die gesuchte Objektart parametrisiert beschreiben lässt und vorhandenes Wissen in die Detektion mit eingebracht werden soll. Ist dies nicht der Fall, führen implizite Verfahren häufig zu besseren Ergebnissen. Die Schwierigkeit bei beiden Ansätzen besteht jedoch darin, die Objektart ausreichend genau zu modellieren, entweder durch geeignete Regeln und Parameter oder durch deskriptive Merkmale und repräsentative Beispiele.

2.1.3. Objektverfolgung

Bei der Objektverfolgung (Tracking) besteht die Hauptaufgabe darin, die Orte, die ein bestimmtes Objekt in allen Bildern einer Sequenz einnimmt, korrekt zu ermitteln. Neben der dabei entstehenden Trajektorie können je nach Anwendung weitere Objekteigenschaften und

-zustände von Interesse sein. Das Tracking kann auch als Spezialfall der Detektion verstanden werden, da auf Basis eines individualisierten Objektmodells eine bestimmte Instanz einer Objektart korrekt in aufeinanderfolgenden Aufnahmen identifiziert werden soll.

Es gibt zahlreiche Verfahren, welche diese Aufgabe auf unterschiedlichem Wege zu lösen versuchen. Sie können u. a. anhand folgender Eigenschaften charakterisiert werden: maximale Anzahl der gleichzeitig verfolgbaren Objekte, feste oder variable Objektanzahl, sequentielle oder abschnittsweise Prozessierung und der Toleranz gegenüber Falschalarmen und Schlupf. Häufig werden Tracking-Verfahren auch den Strategien *Tracking by Detection* oder *Tracking by Model Evolution* zugeordnet (Li u. a., 2008). Im ersten Fall laufen Detektion und Tracking nacheinander und meist unabhängig voneinander ab (vgl. Breitenstein u. a. (2011)). Zuerst werden mögliche Objekte im neuen Bild detektiert und anschließend als bekannt identifiziert oder nicht. Der Schwerpunkt liegt bei dieser Strategie auf der korrekten und effizienten Zuordnung von Detektionen zu bekannten Objekten. Bei der Variante *Tracking by Model Evolution* werden Detektion und Tracking dagegen gleichzeitig durchgeführt. Zu Beginn wird auf Basis der ersten Detektion ein individuelles Objektmodell erstellt. Die Position im nachfolgenden Bild wird dann ermittelt, indem mit diesem Modell nach der Stelle höchster Übereinstimmung gesucht wird (vgl. Comaniciu u. a. (2003)).

Unabhängig davon für welche Strategie man sich entscheidet, hat man laut Jähne (2005) beim Tracking generell mit zwei Schwierigkeiten zu kämpfen: visuell ähnliche Objekte müssen nicht in der realen Welt korrespondieren, da es viele gleichartige Objekte geben kann und umgekehrt, muss ein und dasselbe Objekte nicht visuell korrespondieren, wenn sich z. B. die Beleuchtung oder Ansicht verändert hat. Tracking ist somit ein schlecht gestelltes Problem, da zwar eine Lösung existiert, diese sich aber nicht zwingend aus den Eingangsdaten ergibt. Auf diesen Umstand kann reagiert werden, indem man die Aufnahmefrequenz soweit erhöht, dass der mittlere Verschiebungsvektor der Objekte deutlich kleiner ist als der Abstand zwischen ihnen (Jähne, 2005). Oft ist dies jedoch nicht möglich, so dass stattdessen zahlreiche Regularisierungen vorgenommen werden. Je stärker sich das Tracking-Problem durch angemessene Nebenbedingungen, Annahmen und Vorwissen einschränken lässt, umso besser sind die zu erwartenden Ergebnisse.

In den nachfolgenden Abschnitten werden die beiden beim Tracking zu lösenden Hauptaufgaben (Bar-Shalom und Blair, 2000) behandelt. Hierzu zählt zum einen die Art und Weise, in welcher sich die verschiedenen Teile des Objektmodelles mit der Zeit und dem Eintreffen neuer Detektionen verändern können, und zum anderen das Vorgehen zur Lösung des Korrespondenzproblems, da die Zuordnung von Detektionen zu bekannten Objekten in der Regel nicht eindeutig möglich ist.

Veränderlichkeit des Objektmodells

In der Regel setzt sich das Objektmodell aus mehreren Teilen zusammen, eines beschreibt das äußere Erscheinungsbild, andere diverse Objektzustände (s. Abs. 2.1.1). Im Rahmen der Objektverfolgung muss für jeden Teil definiert werden, welche Art von Anpassungen im Laufe der Zeit z. B. bei der Zuordnung von neuen Detektionen vorzunehmen sind (*dynamic model*). Sollte sich die zeitliche Veränderung des Objektmodells analytisch darstellen lassen, wird es möglich, eine Prädiktion durchzuführen. Dies vereinfacht in vielen Fällen das Zuordnungsproblem, da zwischen Detektion und prädizierten Objekteigenschaften die Ähnlichkeit deutlich höher ist.

Für einige Objekteigenschaften wird diese Aufgabe meist im Rahmen eines Bayes'schen Schätzproblems gelöst (vgl. De Laet (2010)). Hierbei gilt es die Zustände eines veränderlichen Systems über die Zeit auf Basis verrauschter Beobachtungen zu bestimmen. Dies geschieht mithilfe rekursiver Filter (Ristic u. a., 2004) wie dem *Kalman Filter* oder *Particle Filter*, welche sowohl die Prädiktion der Zustände, als auch eine Korrektur nach Zuordnung einer Beobachtung ermöglichen. Des Weiteren wird ein Beobachtungsmodell, welches die

2. Grundlagen und Konzeption

Verbindung zwischen Messungen und internen Zuständen schafft, ein Prozessmodell, das die zeitliche Veränderlichkeit der Zustände beschreibt und ein stochastisches Modell der Genauigkeiten von Zuständen, Beobachtungen und Prozessen benötigt.

Die Objekteigenschaften Position und Geschwindigkeit spielen beim Tracking eine besonders wichtige Rolle. Ihre zeitliche Variabilität wird mit einem Bewegungsmodell beschrieben, welches in der Regel auch eine Prädiktion zulässt. Ist die Objektbewegung im Vergleich zur Aufnahmefrequenz gering oder ist sie gleichförmig und glatt, so kann hierfür ein Modell geringer Komplexität genutzt werden. Ein Beispiel hierfür ist das häufig eingesetzte lineare Modell, welches die Veränderungsrate eines Zustandes als konstant annimmt. Bei abrupten Bewegungsänderungen oder geringer Detektionsfrequenz stoßen einfache Ansätze jedoch an ihre Grenzen und aufwändigere Bewegungsmodelle sind zu bevorzugen. Der IMM-Ansatz (*Interacting Multiple Models*, Blom (1984); Li u. a. (2008)) erlaubt es z. B., gleichzeitig mehrere Bewegungsmodelle für ein Objekt auszuwerten. Darüber hinaus gibt es weitere Modelle, welche versuchen, das spezifische Verhalten der Objektart mit zu berücksichtigen. Im Fall von Personen sind dies u. a. das *Social Force Model* (Helbing und Molnár, 1995), das *Linear Trajectory Avoidance Model* (Pellegrini u. a., 2009) und das *Discrete Choice Model* (Antonini u. a., 2006).

Meist lassen sich einige Teile des Objektmodells nicht in einem Zustandsvektor beschreiben. Verändert sich z. B. das äußere Erscheinungsbild, so werden Verfahren für Prädiktion und Korrektur benötigt, welche die spezielle Form des für die Detektion genutzten visuellen Objektmodells berücksichtigen. Liegt dieses beispielsweise implizit vor, so kann der zugrunde liegende Klassifikator im Laufe des Trackings mit zu- und abgewiesenen Detektionen weiter trainiert werden (Grabner u. a., 2008; Kalal u. a., 2012). Hierbei besteht die Schwierigkeit in der Wahl eines robusten Verfahrens zum unüberwachten Lernen, welches ein Verfälschen des Objektmodells durch Fehlzuordnungen verhindert.

Korrespondenzproblem

Neben der Modellierung der zeitlichen Variabilität des Objektmodells besteht die zweite Hauptaufgabe beim Tracking darin, alle im Laufe der Zeit gemachten Detektionen korrekt zuzuordnen, entweder als Falschalarm oder dem Objekt, welches sie ausgelöst hat. Die Schwierigkeit dieser als Korrespondenzproblem bezeichneten Aufgabe hängt größtenteils von der Anzahl und Unterscheidbarkeit korrekter und falscher Detektionen ab.

Um eine Lösung finden zu können, ist es notwendig, ein Bewertungsmaß (*matching score*) für mögliche Zuordnungen zu definieren. Dieses basiert im Allgemeinen auf einem Vergleich bestimmter Eigenschaften von Objekt und Detektion wie deren Position oder Aussehen. Die Funktion zur Berechnung der Ähnlichkeit (*similarity function*) kann stochastisch oder heuristisch motiviert sein (vgl. Wu und Nevatia (2007)). Bei der Lösung des Zuordnungsproblems werden dann Paarungen bevorzugt, die eine höhere Wahrscheinlichkeit bzw. Ähnlichkeit besitzen, oder entgegengesetzt formuliert, geringere Kosten verursachen.

In der Regel werden Detektionen nur Objekten zugeordnet, die auch in einer der kurz zuvor gemachten Aufnahmen erkannt worden sind. Prinzipiell kann dieser Zeitraum jedoch beliebig lang sein, weswegen man auch Verfolgen und Wiedererkennen unterscheidet. Im ersten Fall erfolgen nur Zuordnungen mit einem geringen zeitlichen Abstand, wobei die Detektionen meist alle vom selben Sensor stammen. Da zwischen den Beobachtungen eine hohe räumliche Korrelation besteht, spielt das Bewegungsmodell bei der Bewertung der Zuordnungen eine große Rolle. Beim Wiedererkennen ist der zeitliche Abstand dagegen deutlich größer. Objekte können z. B. zeitweise verdeckt sein oder ganz aus dem Bild verschwinden und später wieder auftauchen, eventuell auch in den Aufnahmen einer anderen Kamera. Da hier die räumliche Korrelation deutlich geringer ist, erfolgt die Zuordnung meist auf Grundlage der äußeren Erscheinung (vgl. Jüngling (2011)).

Die Anzahl möglicher Paarungen hängt quadratisch von der Zahl an Objekten und Detektionen ab. Zur Reduzierung der Komplexität des Zuordnungsproblems ist es daher üblich, sehr unwahrscheinliche Paarungen von vornherein auszuschließen, indem eine minimale Wahrscheinlichkeit bzw. Ähnlichkeit gefordert wird (*gating*). Da hierfür bereits alle potentiellen Zuordnungen überprüft werden müssen, werden meist effiziente, hierarchische Verfahren basierend auf mehrdimensionalen Suchbäumen und Näherungswerten eingesetzt (Collins und Uhlmann, 1992; Blackman und Popoli, 1999).

Im Vorfeld der Lösung des Korrespondenzproblems muss auch definiert werden, welche Art von Zuordnungen im jeweiligen Anwendungsfall möglich bzw. erlaubt sind. Werden z. B. nur 1:1-Zuordnungen zugelassen, bedeutet dies, dass jedem Objekt nur maximal eine Detektion zugeordnet werden darf bzw. jedes Objekt nur maximal eine Detektion hervorrufen kann (*uniqueness constraint*). Sind dagegen auch sog. Split- und Merge-Situationen erlaubt, kann ein Objekt auch mehrere Detektionen erzeugen (1:n) bzw. mehrere Objekte eine einzige gemeinsame Detektion (n:1). Des Weiteren ist es meist notwendig, Detektionen auch als Falschalarm oder als neues Objekt deklarieren zu können bzw. verschwundene Objekte vom weiteren Tracking auszuschließen.

Für die Lösung des Zuordnungsproblems stehen eine Vielzahl von Verfahren (*data association methods*) zur Verfügung (Pulford, 2005; Cox, 1993; Blackman und Popoli, 1999; De Laet, 2010). Das Spektrum reicht von simplen, heuristischen Ansätzen, die nur die lokale Objektnachbarschaft berücksichtigen, bis hin zu komplexen Methoden, welche versuchen, die global beste Zuordnung für alle Objekte und alle Detektionen für mehrere Zeitpunkte gleichzeitig zu ermitteln. *Local Nearest Neighbor* (LNN) ist bspw. eine der einfachsten und schnellsten Methoden, bei der jedem Objekt nacheinander die wahrscheinlichste, meist räumlich nächste Detektion zugewiesen wird. Etwas aufwändiger, dafür aber deutlich robuster sind *Global Nearest Neighbor* (GNN) Verfahren (Veenman u. a., 2001). Hierbei erfolgt die Zuordnung alle Detektionen und Objekte gleichzeitig für einen bestimmten Zeitpunkt, so dass z. B. die Summe der Ähnlichkeiten aller Paarungen maximal wird. Die optimale Lösung kann je nach Art des Zuordnungsproblems (vgl. Pentico (2007)) mit Hilfe von Standardverfahren wie der Ungarischen Methode (Kuhn, 1955) gefunden werden. LNN- und GNN-Verfahren sind in der Regel deterministisch, d. h. dass sie für jeden Zeitschritt eine eindeutige, unumkehrbare Zuordnung durchführen. Im Gegensatz dazu stehen probabilistische Verfahren (Cox, 1993; Bar-Shalom u. a., 2009), welche mehrere alternative Zuordnungen zulassen und so besser mit Falschalarmen und Mehrdeutigkeiten umgehen können. In schwierigen Situationen kann es auch von Vorteil sein, mehrere Zeitpunkte gleichzeitig auszuwerten (*multidimensional/multiframe assignment*). Treten Mehrdeutigkeiten auf, so lösen sich diese meist nach kurzer Zeit wieder auf. Das Korrespondenzproblem wird jedoch NP-schwer, sobald mehr als zwei Zeitpunkte gleichzeitig betrachtet werden. Um dennoch eine Lösung in akzeptabler Zeit zu finden, ist es immer notwendig, bestimmte Vereinfachungen zu treffen und mit Näherungslösungen zu arbeiten (vgl. Shafique u. a. (2008); Popp u. a. (2001)). Tracklet-Verfahren versuchen dies mit einem hierarchischen Ansatz (Kaucic u. a., 2005; Huang u. a., 2008; Li und Kanade, 2009; Jaqaman u. a., 2008). Auf der untersten Stufe werden mit einfachen und robusten Methoden kurze Trajektorien generiert. Anschließend werden diese mithilfe komplexerer Methoden und zusätzlichem Modellwissen zu immer längeren Einheiten zusammengefasst. Tracklet-Verfahren arbeiten daher immer nur abschnittsweise und nicht sequentiell. Das gleiche gilt für rasterbasierte Verfahren (Andriyenko und Schindler, 2010; Berclaz u. a., 2011). Diese können zwar eine optimale Lösung finden, brauchen aber viel Rechenzeit und schränken die Bewegungsmöglichkeiten der Objekte auf eine begrenzte Anzahl möglicher Positionen ein. *Multi Hypotheses Tracking* (MHT, Reid (1979); Blackman (2004)) ist ein weiteres Verfahren, welches versucht die optimale Zuordnung über mehrere Zeitpunkte zu finden. Es beruht darauf, gleichzeitig mehrere Lösungen für das Korrespondenzproblem zu ermitteln und weiterzuverfolgen bis sich im Laufe der Zeit die wahrscheinlichste durchsetzt. Im Gegensatz zu den meisten zuvor erwähnten Verfahren arbeitet es jedoch sequentiell und kann durch die maxi-

2. Grundlagen und Konzeption



Abbildung 2.2.: Ausschnitte veranschaulichen die typischen Herausforderungen, welche sich beim Erkennen und Verfolgen von Personen in Luftbildsequenzen stellen.

mal zulässige Anzahl alternativer Lösungen sehr gut in seiner Komplexität gesteuert werden (Cox und Hingorani, 1996; Miller u. a., 1997).

2.2. Detektion und Tracking in Luftbildsequenzen

Im letzten Abschnitt wurden allgemeine, theoretische Grundlagen der Bereiche Objekterkennung und -verfolgung behandelt. Der nachfolgende Text befasst sich nun wieder stärker mit der konkreten Aufgabenstellung dieser Arbeit. Zu Beginn werden deren spezifische Rahmenbedingungen und Herausforderungen aufgelistet. Anschließend wird der Stand der Forschung dargestellt und untersucht, für welche Aspekte der Aufgabenstellung bereits erfolgversprechende Lösungsansätze existieren.

2.2.1. Rahmenbedingungen und Herausforderungen

Die Aufgabe, Einzelpersonen in Luftbildsequenzen zu erkennen und zu verfolgen, unterscheidet sich in vielerlei Hinsicht von anderen Anwendungen aus dem Bereich Objektverfolgung und ist mit speziellen Herausforderungen verbunden (vgl. Abb. 2.2). Luftbilder werden in der Regel aus großer Höhe von mehr als 1000 m aufgenommen, um ein weites Gebiet abdecken zu können. Die Bodenauflösung ist daher häufig geringer als ein Dezimeter pro Pixel und im Verhältnis zur Objektgröße meist sehr gering. Aufgrund der großen Entfernung zwischen Kamera und Objekt können zudem Wolken, Dunst oder andere atmosphärische Einflüsse die Sichtbarkeit und damit das Signal-Rausch-Verhältnis weiter verschlechtern. Speziell bei Sonnenschein ist der Personenschatten häufig deutlicher zu erkennen als die Person selbst. Darüber hinaus hat auch die Objektdichte einen großen Einfluss auf die Erkennbarkeit von Einzelpersonen. Im Fall von Menschenmassen, aber auch schon bei kleineren, engstehenden Gruppen sind einzelne Personen häufig nicht mehr zu unterscheiden.

Da die Aufnahme von Luftbildern meist in Nadirrichtung erfolgt, werden Personen direkt von oben aufgenommen und können sich nicht gegenseitig verdecken. Sie erscheinen im Bild als kleine, kompakte Flecken (*blobs*) ohne visuelle Unterscheidungsmerkmale, was das Verfolgen und Wiedererkennen deutlich erschwert. Hinzu kommt noch, dass mit vielen Fehldetektionen zu rechnen ist. Die Ursache hierfür liegt in der schwachen Signatur der Personen und der hohen Komplexität der Bilddaten (Suetens u. a., 1992), welche besonders in urbanen Gebieten viele Objekte enthalten, die Personen visuell sehr ähnlich sind.

Luftbildsequenzen werden von einer sich bewegenden Aufnahmeplattform aus erfasst. Dies führt dazu, dass sich auch statische Objekte in den Bildern zu bewegen scheinen (Paralla-

xeneffekt) und ein und derselbe Ort meist nur für relativ kurze Zeit beobachtet wird. Obwohl zur Erleichterung der Auswertung meist eine Bewegungskompensation sowie andere Vorverarbeitungsschritte durchgeführt werden, muss im Vergleich zu Aufnahmen von statischen Kameras mit erhöhtem Rauschen gerechnet werden.

Da Luftbilder sehr groß sind, kann das Kamerasystem häufig nur mit einer im Vergleich zu Videoaufnahmen sehr geringen Frequenz von wenigen Hertz betrieben werden. Diese Tatsache erschwert die Objektverfolgung, da so der Bewegungsversatz zwischen zwei Bildern in vielen Situationen deutlich größer ist als der Abstand zwischen den Personen. Diese besitzen zusätzlich ein komplexes Bewegungsverhalten und können sich unabhängig voneinander und in beliebige Richtungen fortbewegen. Aufgrund der weiträumigen Abdeckung ist auch damit zu rechnen, dass gleichzeitig eine sehr großen Anzahl an Personen erkannt und verfolgt werden muss.

2.2.2. Stand der Forschung

In diesem Abschnitt werden Publikationen behandelt, die Aufnahmen von fliegenden Plattformen zum Erkennen und Verfolgen von Objekten nutzen. Der Forschungsschwerpunkt in diesem Bereich lag bisher verstärkt auf Fahrzeugen und weniger auf Personen. Es folgt daher eine Zusammenfassung aller relevanten Veröffentlichungen zu Personen und anschließend ein kurzer Überblick der für Fahrzeuge eingesetzten Methoden.

Personen

Obwohl es sehr viele Veröffentlichungen zum Thema Personenerkennung gibt (Dollar u. a., 2012), beschränken sich diese meist auf terrestrische Videoaufnahmen. Luftbildsequenzen oder ähnliche Bilddaten von fliegenden Plattformen werden nur in wenigen Arbeiten neueren Datums genutzt.

In (Reilly u. a., 2010b) werden Personen in Schrägluftbildern in einem zweistufigen Verfahren detektiert. Unter der Annahme, dass Personen aufrecht stehen und einen Schatten werfen, wird das Bild mit Filtermasken nach Orten durchsucht, die diese Bedingungen visuell erfüllen. Anschließend werden an diesen Stellen mit Wavelet-Filtern Bildmerkmale extrahiert, die dann einem SVM-Klassifikator zur endgültigen Entscheidung zugeführt werden. Das Verfahren liefert bei Sonnenschein gute Ergebnisse. Sobald diese Bedingung jedoch entfällt, treten sehr viele Fehldetektionen auf.

In einer anderen Arbeit (Miller u. a., 2008) sollen Personen mit Hilfe von Harris-Merkmalen in UAV-Sequenzen (*unmanned aerial vehicle*) erkannt werden. Hierbei wird angenommen, dass sich speziell im Schulter-Kopf-Bereich zuverlässig Ecken detektieren lassen. Die schlechten Ergebnisse widerlegen jedoch diesen Ansatz. Das in (Sirmacek und Reinartz, 2011) vorgestellte Verfahren geht in die gleiche Richtung, ist jedoch etwas fortschrittlicher. Auch hier wird vorausgesetzt, dass sich Bereiche mit Personen im Luftbild besonders durch Farbdiskontinuitäten auszeichnen. Nach Extraktion von markanten Ecken, werden in einem ersten Schritt Regionen mit besonders vielen dieser Merkmale als dichte Menschenmenge klassifiziert. Im restlichen Teil des Bildes werden anschließend Eckpositionen, die einen ähnlichen Farbwert besitzen, als Einzelpersonen ausgewiesen.

Das Verfahren von Oreifej u. a. (2010) basiert auf einem impliziten visuellen Modell von Personen. In UAV-Schrägaufnahmen werden mit Hilfe eines gleitenden Fensters und auf Basis von Gradientenhistogrammen (Dalal und Triggs, 2005) Personen detektiert und segmentiert. Nach Normalisierung der aktuellen Pose mittels eines einfachen Personenmodells, werden Farb- und Kantenmerkmale zur Wiedererkennung in nachfolgenden Bildern genutzt. Obwohl die Personen bereits sehr klein sein dürfen, erfordert das Verfahren jedoch eine gewisse Mindestgröße, die in echten Luftbilddaufnahmen nicht erreicht wird.

2. Grundlagen und Konzeption

In (Xiao u. a., 2008b) werden unter der Annahme eines statischen Hintergrundes und bewegter Objekte, Fahrzeuge und Personen in UAV-Videoaufnahmen erkannt. Die Berechnung des Optischen Flusses zwischen zwei Bildern ermöglicht sowohl die Kompensation der Kamerabewegung als auch die Detektion sich bewegender Objekte. Diese werden anschließend anhand ihrer Gradientenhistogramme von einem SVM-Klassifikator in Fahrzeuge und Personen unterschieden. Der Optische Fluss wird auch genutzt, um identische Objekte in aufeinanderfolgenden Bildern zu finden. Obwohl das System akzeptable Ergebnisse für Fahrzeuge liefert, scheitert es bei den kleineren, sich langsamer bewegenden Personen.

In einer frühen eigenen Arbeit (Burkert u. a., 2010) werden in einem ersten Schritt eine Reihe von einfachen Segmentierungsverfahren angewandt, um Hintergrundbereiche auszuschließen. Da Personen im Luftbild sich durch eine runde Form fester Größe auszeichnen, wird der übrige Teil mit einem Punktfiler gefaltet und anschließend segmentiert. Besonders kompakte Bereiche mit geringer Fläche werden schlussendlich als Position von Einzelpersonen erkannt. Das Tracking erfolgt anschließend rein datengetrieben, indem Segmente in aufeinanderfolgenden Bildern mit Hilfe des Optischen Flusses verbunden werden.

Fahrzeuge

Bereits seit längerer Zeit werden Aufnahmen von fliegenden Plattformen zur Detektion und Verfolgung von Fahrzeugen genutzt. Da in diesem Bereich ähnliche Probleme auftreten, werden nachfolgend einige Lösungsansätze vorgestellt.

Detektion Eines der ältesten und meist genutzten Verfahren basiert auf der Segmentierung sich bewegender Objekte (Kumar u. a., 2001; Medioni u. a., 2001; Hoogendoorn u. a., 2003). Da jedoch bei bewegter Kamera die Grundannahme eines statischen Hintergrundes nicht eingehalten wird, stellt die Bildregistrierung bzw. Bewegungskompensation in allen Arbeiten einen wichtigen Vorverarbeitungsschritt dar. Aufgrund des Parallaxeneffektes (s. Abs. 2.1.2) können trotzdem noch viele Fehldetektionen auftreten. Im Laufe der Zeit wurden daher etliche Vorschläge zur Lösung dieses Problems präsentiert (Yuan u. a., 2007; Xiao u. a., 2008a; Yu und Medioni, 2009). Die Notwendigkeit einer Objektbewegung bleibt jedoch eine grundsätzliche Schwäche des Ansatzes. Verfahren, die Fahrzeuge aufgrund ihrer Form und Farbe segmentieren (Hinz u. a., 2007; Sharma u. a., 2006; Eikvil u. a., 2009), benötigen meist nur ein Bild und können so auch statische Objekte erkennen. Einige Autoren kombinieren Informationen über Bewegung und Aussehen, um die Segmentierung weiter zu verbessern (Cheng und Butler, 2005; Benedek u. a., 2009). Neben diesen datengetriebenen Verfahren, gibt es auch etliche Arbeiten, die versuchen mit Hilfe eines Objektmodells mehr Vorwissen in den Detektionsprozess einfließen zu lassen. Während frühere Veröffentlichungen meist auf einer expliziten Modellierung der Fahrzeuge basierten (Zhao und Nevatia, 2003; Hinz, 2005), werden nun verstärkt implizite Verfahren genutzt (Yu u. a., 2006; Grabner u. a., 2008; Leitloff u. a., 2010).

Tracking Zum Verfolgen der Fahrzeuge kommen viele unterschiedliche Verfahren zum Einsatz. In (Nejadasl u. a., 2006) wird z. B. der Versatz eines Fahrzeuges allein auf Basis des Optischen Flusses bestimmt. In (Hinz u. a., 2007) und (Szottka und Butenuth, 2011) werden mögliche Positionen in nachfolgenden Bildern dagegen via *Shape-Based (Template) Matching* ermittelt. Anschließend werden, unter Annahme einer konsistenten Bewegung, verbleibende Mehrdeutigkeiten aufgelöst und die endgültige Objektposition gefunden. Viele Arbeiten setzen beim Tracking von Fahrzeugen auch auf die Zuordnung von Detektionen in aufeinanderfolgenden Bildern. In (Reilly u. a., 2010a) und (Xiao u. a., 2010) geschieht dies bspw. in einem sequentiellen GNN-Ansatz auf Basis von Abstand, Fahrverhalten, Straßenrichtung sowie der Beziehung zu benachbarten Fahrzeugen. Andere Autoren nutzen zur Generierung von Trajektorien den hierarchischen Tracklet-Ansatz (Kaucic u. a., 2005; Lin u. a., 2011).

Forschungsbereiche mit ähnlichen Problemstellungen

In den letzten beiden Abschnitten wurden Veröffentlichungen und Methoden vorgestellt, welche die Erkennung und Verfolgung von Personen und Fahrzeugen in Luftbildsequenzen oder ähnlichen Bilddaten zum Thema haben. Abstrahiert man jedoch die hier behandelte Problemstellung, so können auch Erkenntnisse aus anderen Forschungsbereichen zur Lösung beitragen. Allgemein geht es darum, winzige, gleichartige, schwach erkennbare Objekte in Bilddaten einzeln zu erkennen und eine sehr große Anzahl von ihnen unter Einfluss von Rauschen durch eine Bildsequenz zu verfolgen.

Detektion In Teilbereichen der Biologie ist die Bildanalyse bspw. ein wichtiges Werkzeug, um Experimente automatisch auswerten zu können. In Bildsequenzen, welche mit speziellen Mikroskopen aufgenommen werden, sollen primitive Organismen, Zellen oder Moleküle erkannt und verfolgt werden (Meijering u. a., 2009). Besonders interessant sind hier Arbeiten, welche sich mit subzellulären Strukturen beschäftigen, die nur wenige Pixel groß und schwach zu erkennen sind. In (Smal u. a., 2010) findet man einen Vergleich verschiedener Detektionsansätze. Es stellt sich heraus, dass Verfahren auf Basis eines impliziten visuellen Modells unter Nutzung von Haar-Merkmalen bei geringem Signal-Rausch-Verhältnis die besten Ergebnisse liefern (vgl. auch Jiang u. a. (2007)).

Tracking Verfahren zum Verfolgen kleiner, ähnlicher Objekte in großer Zahl werden in der Literatur häufig als *Particle/Point Tracking Methods* bezeichnet (Yilmaz u. a., 2006) und kommen in unterschiedlichsten Anwendungsfeldern wie der Strömungsmechanik, Tierbeobachtung oder Molekularbiologie zum Einsatz. Das in (Matov u. a., 2011) vorgestellte Verfahren basiert darauf, dass Detektionen in drei aufeinanderfolgenden Bildern zusammen analysiert werden. Alle so konstruierten Triplets werden in einem Graphen zusammengefasst, wobei sich die beste Zuordnung durch Lösen eines Minimum-Cost-Flow-Problems ergibt. In die Gewichtung der unterschiedlichen Varianten fließt sowohl die Gleichmäßigkeit der Bewegung als auch deren Konsistenz zu benachbarten Partikeln mit ein. In (Jaqaman u. a., 2008) wird ein zweistufiger Tracklet-Ansatz zum Verfolgen von Partikeln in Zellen genutzt. Da sich diese auch Teilen und Vereinen können, werden bei der Zusammenführung der Tracklets auch Split- und Merge-Situationen zugelassen. Die gleiche Strategie wird in (Wu u. a., 2011) zum Verfolgen von Fledermäusen in Infrarotaufnahmen angewandt. Hier führen häufige Verdeckungen dazu, dass Trajektorien sich aufteilen oder vereinigen. Viele der vorgestellten Verfahren versuchen mit heuristischen Ansätzen das stochastische MHT-Verfahren (Reid, 1979; Blackman, 2004) zu approximieren. Dieses wurde ursprünglich für die Luftraumüberwachung mittels Radar entwickelt und ist besonders geeignet, wenn kleine Objekte mit geringem Abstand, unter erhöhtem Rauscheinfluss verfolgt werden müssen (Blackman und Popoli, 1999).

2.3. Diskussion und Auswertestrategie

Die Aufgabe Personen in Luftbildsequenzen zu erkennen und zu verfolgen, impliziert zahlreiche Schwierigkeiten, die in anderen Anwendungsbereichen so nicht auftreten und stellt hohe Anforderungen an das zu entwickelnde System zur automatischen Bildauswertung.

Die Literaturrecherche hat ergeben, dass es nur sehr wenige Veröffentlichungen gibt, die sich mit dieser Aufgabenstellung beschäftigen. Keine von ihnen behandelt die Thematik jedoch in ausreichendem Umfang oder in zufriedenstellender Weise. Die eingesetzten Detektionsverfahren können entweder nicht ausreichend mit der hohen Komplexität der Bilddaten oder der geringen Größe der Objekte umgehen. Die Personenverfolgung wird meist gar nicht oder nur rudimentär behandelt. Luftbilddaten werden dagegen schon seit längerer Zeit zum Erkennen und Verfolgen von Fahrzeugen genutzt. Die Arbeiten in diesem Bereich sind daher

2. Grundlagen und Konzeption

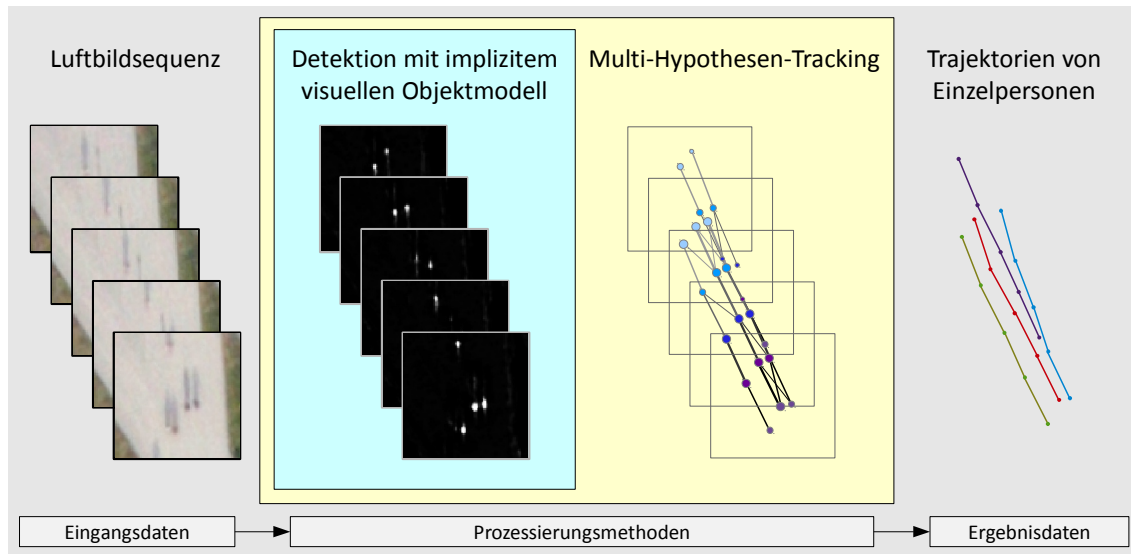


Abbildung 2.3.: Übersicht, der in dieser Arbeit verfolgten Strategie zum Erkennen und Verfolgen von Einzelpersonen in Luftbildsequenzen.

schon deutlich weiter fortgeschritten und liefern wichtige Ideen und Ansätze. Dies gilt auch für Veröffentlichungen aus anderen Forschungsbereichen, die zwar mit anderen Bilddaten und Objektarten zu tun haben, aber mit sehr ähnlichen Problemstellungen.

In dieser Arbeit werden daher Verfahren aus unterschiedlichen Bereichen kombiniert und weiterentwickelt, um die Herausforderungen der Aufgabenstellung zu bewältigen. Die gewählte Strategie (s. Abb. 2.3) wird in den nachfolgenden Absätzen grob umrissen und begründet. Die Details zur Umsetzung folgen in den Kapiteln 3 und 4.

2.3.1. Detektion

Die Detektion von Einzelpersonen erfolgt in dieser Arbeit mit Hilfe eines impliziten, visuellen Modells, da dieser Ansatz viele Vorteile bietet und bereits mehrfach erfolgreich in schwierigen Situationen eingesetzt wurde (vgl. Jiang u. a. (2007); Leitloff u. a. (2010)). Aufgrund der geringen Objektgröße und des geringen Signal-Rausch-Verhältnisses können datengetriebene Ansätze Einzelpersonen nicht zuverlässig segmentieren. Modellgetriebene Verfahren arbeiten in solchen Situationen wesentlich robuster. Der höhere Zeitbedarf für die Suche, lässt sich mit diversen Methoden reduzieren. Das äußere Erscheinungsbild der Personen wird mit einem impliziten Modell beschrieben, da diese Vorgehensweise viel Flexibilität bei den für die Detektion verwendeten Merkmalen erlaubt. Im Gegensatz zu Verfahren mit explizitem Modell, muss auch das Aussehen der Personen weniger exakt formuliert werden und der sonst notwendige Projektionsschritt entfällt vollständig. Da das Aussehen der Personen in den Bilddaten nur wenig variiert, lassen sich diese gut anhand visueller Formmerkmale beschreiben. Auf die Segmentierung von sich bewegenden Bildbereichen wird dagegen verzichtet, da diese Methode für sehr kleine Objekte und eine bewegte Kamera nicht zuverlässig genug arbeitet und stehende Personen gar nicht erkennt. Aufgrund der geringen Größe und kompakten Form werden detektierte Personen allein durch ihren Mittelpunkt beschrieben. Zusätzlich liefert der für die Objekterkennung eingesetzte Klassifikator einen Wert, der die Zuverlässigkeit jeder Detektion angibt.

2.3.2. Tracking

Die Personen werden mittels des Multi-Hypothesen-Tracking-Ansatzes verfolgt. Dieses Verfahren wurde speziell für schwierige Bedingungen mit kleinen, schwach erkennbaren Objekten und vielen Falschalarmen entwickelt und ist damit für die hier gestellte Aufgabenstellung besser geeignet als andere (vgl. Yilmaz u. a. (2006); Cox und Hingorani (1996); Blackman (2004)). Es ist zudem ausreichend flexibel, um mit einer variierenden Anzahl an Objekte umgehen zu können. Die sequentielle Formulierung in Zusammenspiel mit effektiven Methoden zur Begrenzung der Komplexität ermöglicht es, eine nahezu optimale Lösung für das Tracking-Problem in Echtzeit zu ermitteln. Da MHT zu den stochastischen Verfahren gehört und zu jedem Zeitpunkt mehrere alternative Lösungen verfolgt, kann es schwierige Zuordnungsentscheidungen in mehrdeutigen Situationen hinauszögern bis mehr Informationen verfügbar sind. Als nachteilig wird häufig gesehen (Veenman u. a., 2001), dass das MHT-Verfahren u. a. sehr rechenaufwändig ist und die benötigten Wahrscheinlichkeiten schwierig zu bestimmen sind. Auf beide Aspekte wird in dieser Arbeit entsprechend eingegangen.

Ein hierarchischer Tracklet-Ansatz wird dagegen nicht genutzt. Dieser beruht darauf, dass sich in einer ersten Phase ausreichend viele, zuverlässige Tracklets bilden lassen, was jedoch hier bei zahlreichen Fehldetektionen, vielen schwach erkennbaren Objekten und einer relativ geringer Aufnahmefrequenz nicht gegeben ist. Des Weiteren sind auch Verfahren nach der Methode *Tracking by Model Evolution* nicht geeignet, da sich für die einzelnen Personen kein individuelles visuelles Modell erzeugen lässt, mit dem im nachfolgenden Bild die neue Objektposition ermittelt werden könnte. Stattdessen ermöglicht es die ähnliche und gleichbleibende Form der Personen, einen gemeinsamen, konstanten Detektor für die gesamte Objektklasse zu nutzen. Dieser erzeugt in jedem neuen Bild Detektionen, die anschließend mittels des MHT-Ansatzes zugeordnet werden. Hierbei wird die Bedingung, dass eine Detektion nur von einem Objekt stammen kann, explizit berücksichtigt.

Des Weiteren geschieht die Zuordnung im MHT-Ansatz auf Basis von Wahrscheinlichkeiten, was eine einfache, theoretisch fundierte Integration unterschiedlicher Informationen erlaubt. In dieser Arbeit gehören hierzu sowohl die Ergebnisse der Schätzung der Personenzustände als auch die Konfidenz jeder einzelnen Detektion. Letzteres stellt einen besonderen wissenschaftlichen Beitrag dieser Arbeit dar, da so die Detektionsmethode mit implizitem visuellem Modell in den MHT-Ansatz integriert wird. Die finale Objekterkennung geschieht somit nicht mehr auf Grundlage eines einzelnen Bildes, sondern erst im Zuge des Trackings, wenn die initiale Hypothese durch zusätzliche, konsistente Beobachtungen über die Zeit bestätigt werden konnte. Diese Erweiterung ist besonders vorteilhaft für alle Anwendungen mit geringem Signal-Rausch-Verhältnis, da auch schwache Detektionen im Einzelbild über die Zeit betrachtet eine konsistente Trajektorie ergeben können.

3. Objekterkennung

In diesem Kapitel wird der Prozess zum Erkennen von Einzelpersonen in Luftbildern detailliert ausgeführt. Dieser soll möglichst schnell ablaufen, robust gegenüber Störungen sein und Personen zuverlässig detektieren. Um dies zu erreichen, wird der aussehensbasierte Ansatz mit implizitem Objektmodell genutzt. Hierbei wird das äußere Erscheinungsbild der Personen durch einen Klassifikator anhand von lokalen Bildmerkmalen und Beispielbildern gelernt. Die eigentliche Detektion erfolgt dann, indem das gesamte Luftbild mit einer Filtermaske und dem trainierten Klassifikator nach Personen abgesucht wird. Einen Überblick der einzelnen Schritte des Detektionsprozesses gibt Abbildung 3.1.

Nachfolgend werden anwendungsspezifische Anpassungen und allgemeingültige Verbesserungen dieses Verfahrens beschrieben. Viele Teile dieses Kapitels sind bereits in (Schmidt und Hinz, 2011) veröffentlicht worden. Hier werden die Zusammenhänge jedoch wesentlich ausführlicher und in die Gesamtstrategie eingebettet präsentiert. Zudem stellt die stochastische Modellierung der Detektionsergebnisse in Abschnitt 3.6 eine wesentliche Neuerung dar.

3.1. Einzelpersonen in Luftbildern

Um Personen anhand ihres Aussehens in Luftbildern erkennen zu können, gilt es zuerst, alle möglichen Variationen und Einflussfaktoren zu bestimmen. Der gesamte Detektionsprozess wird anschließend darauf ausgerichtet. Die in Abbildung 3.2 enthaltenen Beispiele geben einen guten Überblick über das Spektrum möglicher Erscheinungsformen. Wie bereits im Abschnitt 2.2.1 erwähnt, fallen besonders die geringe Größe und der Einfluss des Schattens auf.

Geht man davon aus, dass eine durchschnittliche Person eine Fläche von etwa 30 cm x 50 cm benötigt (vgl. Abb. 30 in (Still, 2000)), so nimmt sie direkt von oben betrachtet, bei einer für Luftbilder typischen Bodenauflösung von 10 cm bis 20 cm, nur etwa 5 bis 10 Pixel im Bild ein. In den Beispielbildern kann man gut erkennen, dass die geringe Auflösung wie eine Glättung wirkt und die exakte Kontur der Personen verschwinden lässt. Grundsätzlich sind einzelne Personen jedoch als kleiner, kompakter Fleck erkennbar, wobei ihre Ausrichtung bzw. Blickrichtung nur einen geringen Einfluss auf die Form hat. Ihre Intensität variiert hauptsächlich zwischen hell und dunkel und wird von der aktuellen Aufnahmekonfiguration sowie dem Stand der Sonne beeinflusst.

Aufgrund der Gesetzmäßigkeiten der Zentralprojektion werden Personen auch in Nadiraufnahmen im Allgemeinen nie genau von oben, sondern leicht schräg von der Seite abgebildet. Da sich Kamera und Personen relativ zueinander bewegen, ändert sich zudem die Form der Personen kontinuierlich im Lauf der Bildsequenz. In dieser Arbeit wird jedoch immer von einer Draufsicht ausgegangen, was in den meisten Fällen eine akzeptable Approximation darstellt und die Modellierung deutlich vereinfacht. Zu einer Verschlechterung der Detektionsergebnisse kommt es nur in schräg aufgenommenen Luftbildern und in abgeschwächter Form am Rand von Nadiraufnahmen. Sollte sich diese Einschränkung für bestimmte Anwendungsfälle zu stark negativ auswirken, muss der Detektionsprozess entsprechend angepasst werden (vgl. Reilly u. a. (2010b)).

3. Objekterkennung

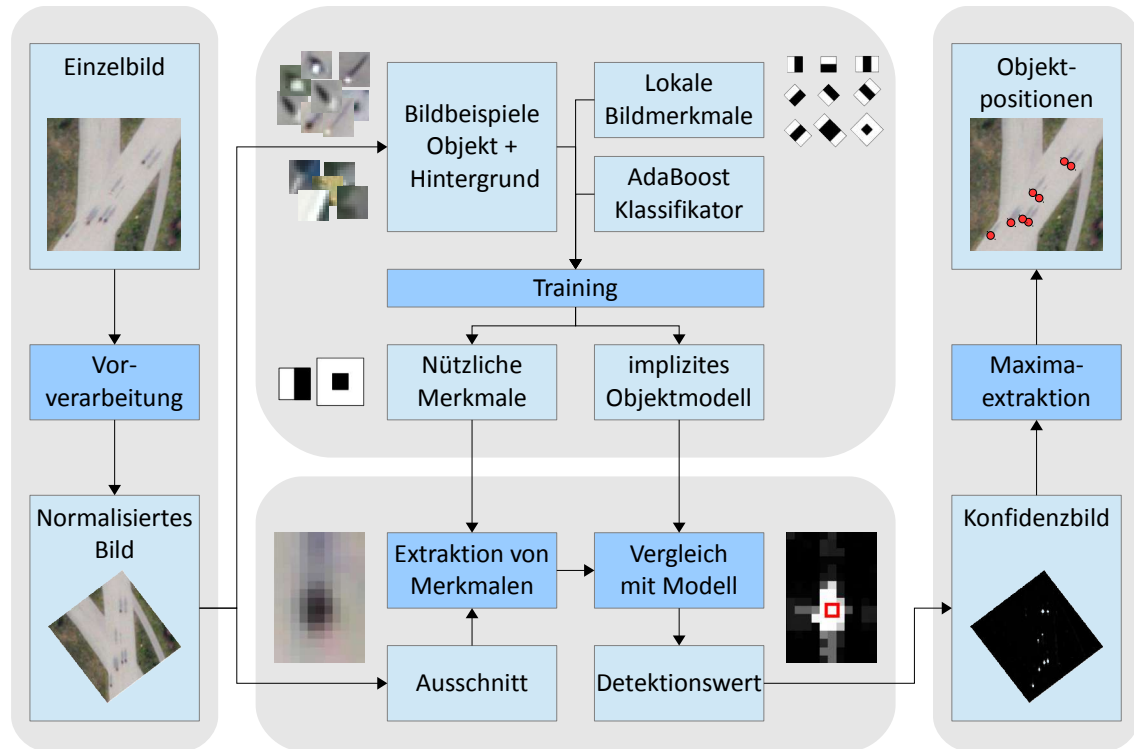


Abbildung 3.1.: Übersicht des Detektionsverfahrens mit implizitem visuellem Objektmodell.

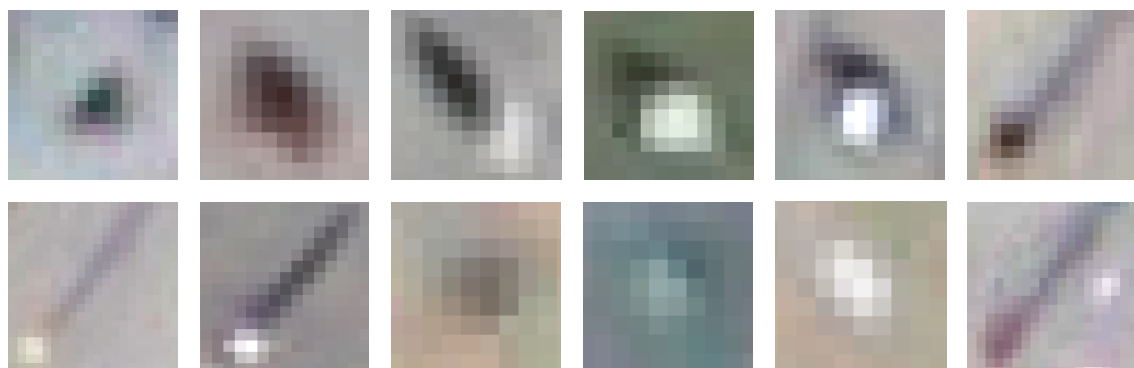


Abbildung 3.2.: Beispiele für Einzelpersonen in Luftbildern bei einer Bodenauflösung von 12 cm bis 17 cm pro Pixel, variierendem Sonnenstand und unterschiedlichem Kontrast.



Abbildung 3.3.: Einfluss der lokalen Objektdichte und des Personenschattens auf die Erkennbarkeit von Einzelpersonen.

Scheint die Sonne, so ist üblicherweise auch ein Personenschatten erkennbar, der je nach Sonnenstand in Länge und Ausrichtung variiert. Da der Schatten sich meist gut vom Untergrund abhebt und auch die spektrale Signatur der Personen vergrößert, erleichtert er deren Erkennbarkeit in der Regel deutlich. Unter bestimmten Umständen verschmelzen jedoch Personen mit ihrem Schatten und der Körpermittelpunkt lässt sich nur noch erahnen. Auch kann es vorkommen, dass eine Person im Schatten einer anderen verschwindet. Diese negativen Einflüsse lassen sich gut in Abbildung 3.3 beobachten. Hier wird zudem dargestellt, wie die Erkennbarkeit von Einzelpersonen mit zunehmender lokaler Objektdichte abnimmt und wie der Personenschatten diesen Effekt zusätzlich verstärkt.

3.2. Vorverarbeitungsschritte

Die Vorverarbeitung der Bilddaten stellt eine wichtige Voraussetzung für gute Detektionsergebnisse dar. Hierbei sollen sowohl zufälliges Rauschen als auch systematische Effekte reduziert werden, so dass es anschließend zu weniger Falschalarmen und weniger Schlupf kommt. Vorverarbeitungsschritte werden meist global auf das ganze Bild angewandt, wobei in radiometrische und geometrische Korrekturen unterschieden wird. Im ersten Fall werden die Farbwerte der Pixel verändert, während es im zweiten zu einer geometrischen Transformation kommt. Diese ist in der Regel mit einer Neuabtastung (*resampling*) verbunden und hat daher stets einen leichten Informationsverlust zur Folge. Aus diesem Grund sollten radiometrische Korrekturen immer zuerst durchgeführt werden. Mehrere geometrische Transformationen sollte zudem nicht sequentiell sondern gemeinsam ausgeführt werden, um Zeit zu sparen und den Informationsverlust zu minimieren. Nachfolgend werden die zur Verbesserung der Detektion von Personen in Luftbildern sinnvollen Vorverarbeitungsschritte diskutiert.

3.2.1. Radiometrische Anpassungen

Durch Wolken, Dunst, Sonne und Schatten wird die Sichtbarkeit der Personen beeinträchtigt. Da diese Phänomene meist lokal stark variieren, wird jedoch von einer globalen Verbesserung des Kontrastes z. B. durch eine einfache Grauwertspreizung abgesehen. Stattdessen wird im Zuge der Detektion auf Bildmerkmale zurückgegriffen, die invariant gegenüber Beleuchtungsunterschieden sind bzw. bei ihrer Berechnung explizit eine lokale Farbraumnormalisierung durchführen (vgl. Parks und Levine (2010)).

Des Weiteren wäre es möglich, ein durch den Aufnahmeprozess hervorgerufenen Bildrauschen durch geeignete Glättungsverfahren zu reduzieren. Aufgrund des niedrigen Signal-Rausch-Verhältnisses in den vorliegenden Bilddaten bestünde jedoch die Gefahr, dass für die Detektion von Personen wichtige Bildinformationen verloren gehen. Ein besserer Ansatz ist es daher auch hier, zur Beschreibung der Personen auf Bildmerkmale zurückzugreifen, die robust gegenüber Rauschen sind.

3. Objekterkennung

Eine nützliche radiometrische Korrektur, die auch global angewendet werden kann, ist die Transformation der Bilddaten in einen anderen Farbraum. Die Luftbilder liegen standardmäßig im RGB-Format vor, welches jedoch aufgrund seiner stark korrelierten Kanäle für die Berechnung von Bildmerkmalen weniger gut geeignet ist. In dieser Arbeit wird daher der $i1i2i3$ -Farbraum genutzt (Ohta u. a., 1980), welcher Farbe und Intensität gut voneinander trennt, leicht zu berechnen ist und keine Sprungstellen aufweist. Denkbar sind auch andere Farbräume mit ähnlichen Eigenschaften. Welcher die besten Ergebnisse liefert, hängt auch von der Wahl der Bildmerkmale ab und sollte daher während des Trainings des Klassifikators ermittelt werden (vgl. Liu und Liu (2010)). Bei jeder Farbraumtransformation ist stets zu berücksichtigen, dass sich die beiden Ziele, hohe Robustheit gegenüber Störeinflüssen und hohe Trennfähigkeit nicht gleichzeitig erreichen lassen (Geusebroek u. a., 2001).

3.2.2. Geometrische Anpassungen

Die Bodenauflösung der Luftbilder hängt von der Flughöhe und dem Kamerasensor ab und liegt üblicherweise zwischen 10 cm und 20 cm pro Pixel. Um die damit verbundene Variation der Objektgröße zu eliminieren, werden die Bilder einheitlich auf eine Pixelgröße von 15 cm skaliert. Dieser Wert stellt einen Kompromiss dar, bei dem der Vorteil einer höheren Auflösung nicht komplett verloren geht und auch schlechter aufgelöste Bilddaten noch mit dem selben Detektor ausgewertet werden können. Würde die Auflösung noch stärker schwanken, hätte dies eine stärkere Variation des Aussehens von Personen zur Folge und es müssten eventuell mehrere Detektoren eingesetzt werden.

Der Personenschatten kann dazu beitragen, dass Einzelpersonen leichter zu erkennen sind. Je nach Sonnenstand verändern sich jedoch dessen Länge und Richtung, was die Varianz der äußeren Erscheinung von Personen erhöht und sich daher negativ auf die Qualität der Detektion auswirkt. Um diesen Nachteil zu eliminieren, wird bei Sonnenschein das gesamte Bild so gedreht, dass der Schatten immer nach Norden zeigt. In die Berechnung des Drehwinkels fließen die geographischen Koordinaten des Bildes sowie Datum und Uhrzeit der Aufnahme ein (Reda und Andreashinz, 2004).

3.3. Detektor

In der Objekterkennung mit implizitem visuellem Modell fungiert die Filtermaske als Detektor. Angewandt auf eine bestimmte Stelle im Bild, berechnet dieser bestimmte Merkmalswerte und gibt sie an den Klassifikator zum Abgleich mit dem vorab gelernten Objektmodell weiter (s. Abb. 3.1). Die Form des Detektors sowie die Art der lokalen Merkmale müssen speziell auf den jeweiligen Anwendungsfall und die Objektart zugeschnitten werden.

3.3.1. Lokale Bildmerkmale

Es gibt eine Vielzahl möglicher Merkmale, welche innerhalb des Detektor extrahiert werden können (vgl. Gerónimo u. a. (2010)). Um jedoch für die Objekterkennung geeignet zu sein, sollten sie eine hohe Toleranz gegenüber zufälligen Variationen besitzen, eine zuverlässige Trennung von Objektklasse und Hintergrund erlauben sowie schnell zu berechnen sein. Häufig ist die Form das wichtigste visuelle Merkmal zum Erkennen bestimmter Objekte, noch vor Farbe oder Textur. Meist werden jedoch unterschiedliche Arten von Merkmalen extrahiert, um für die folgende Klassifikation komplementäre Informationen zur Verfügung zu haben (vgl. Grabner u. a. (2008)).

Weit verbreitet sind Histogrammmerkmale (vgl. Abs. 2.2.2). Hierbei werden die Eigenschaften aller Pixel innerhalb des Detektors in einem oder mehreren Histogrammen akkumuliert.

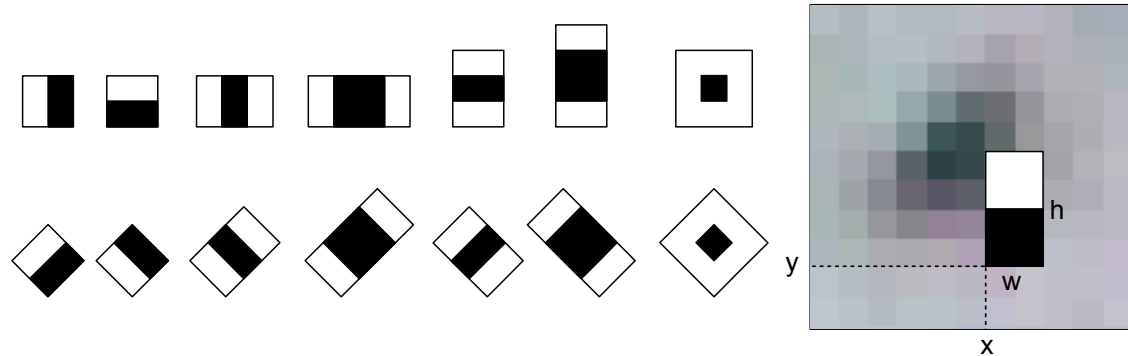


Abbildung 3.4.: Übersicht der gebräuchlichsten Arten von Haar-Merkmalen. Diese werden zusätzlich über Größe (w , h) und Lage (x , y) innerhalb des Detektors instanziiert.

Genutzt werden unterschiedliche Eigenschaften wie Farbwerte, Textur (*Local Binary Pattern*, Ojala u. a. (2002)), Gradienten (*Local Edge Orientation Histogram*, Levi und Weiss (2004); *Histogram of Oriented Gradients*, Dalal und Triggs (2005)) oder Bewegung (*Histogram of Optical Flow*, Dalal u. a. (2006)). Im Allgemeinen wird das gebildete Histogramm normalisiert und die relativen Häufigkeiten als Merkmalswerte verwendet. Histogrammmerkmale haben den Vorteil, dass sie robust gegenüber Variationen der Form des gesuchten Objektes sind, da bei ihrer Berechnung die Lage der Pixel innerhalb des Detektors keine bzw. nur eine untergeordnete Rolle spielt. Sollten im Histogramm die Eigenschaften von besonders vielen Pixeln enthalten sein, so ist es auch robust gegenüber kleineren Störungen wie partieller Verdeckung. Histogrammmerkmale können jedoch nicht zuverlässig für sehr kleine Objekte eingesetzt werden. Der Detektor umfasst hier nur wenige Pixel, was dazu führt, dass Störungen einen zu großen Einfluss haben und das Histogramm nicht zuverlässig normalisiert werden kann.

Verallgemeinerte Haar-Merkmale

Die markanteste visuelle Eigenschaft von Personen in Luftbildern ist ihre kompakte Form. Diese lässt sich sehr gut mit Haar-Merkmalen beschreiben, welche zudem robust gegenüber Beleuchtungsänderungen und geringen Formvariationen sind, sich mit Hilfe von Integralbildern (Crow, 1984; Viola und Jones, 2001) sehr schnell berechnen lassen und auch für sehr kleine Objekte eingesetzt werden können (Jiang u. a., 2007; Leitloff u. a., 2010).

Haar-Merkmale wurden erstmals von Oren u. a. (1997) zum Erkennen von Personen in Bildern eingesetzt und besitzen eine starke Ähnlichkeit mit Haar-Wavelets. Der ursprünglich kleine Satz von drei Merkmalen ist im Laufe der Zeit deutlich erweitert worden, um bessere Detektionsergebnisse zu erzielen (*Haar-like features*, Viola und Jones (2001); Lienhart und Maydt (2002)). Er umfasst nun Eck-, Linien- und Punktmerkmale jeweils in waagerechter, senkrechter und diagonaler Ausführung (s. Abb. 3.4). Ein konkretes Haar-Merkmal wird durch seine Art, Größe und Lage innerhalb des Detektors definiert.

Die Vorschrift zur Berechnung des Merkmalswertes ist für alle Arten gleich. Dieser ergibt sich, indem die Summe aller Pixelwerte innerhalb des schwarzen Rechtecks von der Summe aller Pixelwerte innerhalb des gesamten Merkmals subtrahiert werden, wobei die unterschiedliche Anzahl beteiligter Pixel durch einen Faktor kompensiert wird. Hierdurch ist der Merkmalswert invariant gegenüber additiven Beleuchtungsänderungen. Eine zusätzliche Robustheit wird erreicht, indem durch die Standardabweichung aller Pixelwerte innerhalb des Detektorfensters dividiert wird (Varianznormalisierung). Da es in homogenen Bereichen hierbei zu numerischen Problemen kommen kann, muss eine untere Schranke für die Varianz festgelegt werden.

3. Objekterkennung



Abbildung 3.5.: Bildungsvorschrift der verallgemeinerten Haar-Merkmale und Auswahl einiger speziell für die Detektion von Personen in Luftbildern entwickelter Varianten.

Unter Beachtung der grundlegenden Bildungsvorschrift werden die Haar-Merkmale in dieser Arbeit noch weiter verallgemeinert (s. Abb. 3.5). Die Proportionen des Merkmals und das Verhältnis der beiden Flächen können nun individuell auf den jeweiligen Anwendungsfall zugeschnitten werden. Zudem stehen zusätzliche Merkmale für Ecken und Linienenden zur Verfügung. Obwohl sich diese Formen auch durch eine Kombination einfacher Haar-Merkmale abbilden lassen, so erlauben es die verallgemeinerten Haar-Merkmale, komplexere Strukturen mit einem einzigen Merkmal zu repräsentieren. Dies hat den Vorteil, dass bei der Detektion weniger Merkmale und folglich auch weniger Zeit benötigt werden. Zudem reduzieren höherwertiger Merkmale im Allgemeinen die Anzahl an Beispielen, die für das Training des Klassifikators gesammelt werden müssen (Levi und Weiss, 2004). Abbildung 3.5 stellt einige Merkmale vor, welche auf die spezielle Form von Personen in Luftbildern angepasst wurden. Da die Farbintensität der Personen zwischen hell und dunkel schwankt, wird auch mit Haar-Merkmalen experimentiert, bei denen der Betrag des Merkmalswertes genutzt wird.

Rechteckmerkmale

Obwohl die Form das dominierende Merkmal von Personen in Luftbildern ist, sollen für die Detektion noch Farbe und Textur als komplementäre Informationen genutzt werden. Daher werden in dieser Arbeit Rechteckmerkmale eingeführt. Deren Wert ergibt sich aus allen Pixeln innerhalb eines beliebig geformten Rechtecks innerhalb des Detektors. Genutzt werden Mittelwert und Varianz, die sich analog zu den Haar-Merkmalen sehr schnell mit Hilfe von Integralbildern berechnen lassen. Der Mittelwert repräsentiert Farbe oder Helligkeit, während die Varianz die Homogenität innerhalb des Rechtecks wiedergibt. Da mehrere Pixel in das Merkmal einfließen, ist es weniger rauschanfällig als einzelne Pixelwerte. Mittelwert und Varianz sind jedoch an sich nicht invariant gegenüber bestimmten Beleuchtungsänderungen. Dieser Nachteil lässt sich reduzieren, indem der L^2 -Farbraum zur Berechnung der Merkmalswerte genutzt wird (van de Sande u. a., 2010).

3.3.2. Form des Detektors

Als Filtermaske eingesetzt besitzt der Detektor einen Bezugspunkt und eine auf Objektgröße und Bildmerkmale abgestimmte Form. Diese ist meist rechteckig, was Implementierung und Handhabung vereinfacht und für viele Objektarten geeignet ist. Grundsätzlich kann der Detektor jedoch beliebig geformt sein. Allgemein ergibt sich seine Gesamtfläche aus der Schnittmenge aller für die Merkmalsberechnung benötigten Teilflächen. Werden, wie in dieser Arbeit, unterschiedliche Merkmalsarten genutzt, so ist es sinnvoll, für jede von ihnen den optimalen Bereich und Farbkanal zu definieren. Aus der Festlegung von Merkmalsart und Bereichsgröße ergibt sich die Anzahl möglicher Merkmale. Welche davon tatsächlich für die Detektion genutzt werden, wird erst im Zuge des Klassifikatortrainings und der darin enthaltenen Merkmalsreduktion entschieden.

Es ist zu beachten, dass die Anzahl potentieller Merkmale meist stark von der Größe des Auswertebereiches abhängt. So ergeben sich für ein einziges 2×2 großes Haar-Merkmal in

Merkmalsart	Farbkanäle	Auswertebereich	Anzahl möglicher Merkmale
Haar	i1	9 x 15 Pixel	9477
Rechteck	i1 + i2 + i3	9 x 9 Pixel	3 x 3969

Tabelle 3.1.: Konfiguration des in dieser Arbeit genutzten Detektors vor der Merkmalsreduktion.

einem 5 x 5 großen Bereich durch diverse Skalierungs- und Verschiebungsmöglichkeiten allein 33 und in einem 6 x 6 großen Bereich schon 65 mögliche Ausprägungen. Um das Training des Klassifikators und die Merkmalsauswahl nicht durch zu viele Merkmale zu überlasten, sollte daher der Auswertebereich so klein wie möglich gewählt werden. Für sehr große Objekte wird jedoch auch ein entsprechend großer Detektor benötigt. In solch einem Fall wird die Anzahl potentieller Merkmale meist durch heuristische Methoden eingeschränkt, z. B. indem jedes zweite von vornherein verworfen wird. Aufgrund der geringen Ausdehnung von Personen in Luftbildern ist dieses Vorgehen jedoch nicht notwendig.

In einem speziellen Vorverarbeitungsschritt wird das Luftbild stets so gedreht, dass der Personenschatten nach Norden zeigt (s. Abs. 3.2.2). Dieses Vorgehen reduziert die Varianz der äußeren Erscheinung, ermöglicht es aber auch, den Personenschatten direkt in das implizite, visuelle Objektmodell mit aufzunehmen und ihn so zur Verbesserung der Detektion zu nutzen. Hierfür wird die Form des Detektors bzw. des Auswertebereiches der Merkmale so ausgelegt, dass neben der Person auch ein Teil des Schattens mit abgedeckt ist. Damit nicht je nach Sonnenschein zwei unterschiedliche Detektoren eingesetzt werden müssen, wird der zugehörige Klassifikator mit Personenbeispielen mit und ohne eigenen Schatten trainiert. Ein alternativer, jedoch aufwändigerer Ansatz zur Nutzung des Personenschattens stellt die komponentenbasierte Detektion dar. Wie in (Reilly u. a., 2010b) gezeigt, müssen hierfür Person und Schatten separat modelliert, erkannt und ihre relative Lage auf Konsistenz überprüft werden.

In Tabelle 3.1 ist die in dieser Arbeit genutzte Konfiguration des Detektors dargestellt. Sein Bezugspunkt fällt mit dem Mittelpunkt einer Person zusammen. Für die verallgemeinerten Haar-Merkmale wird ein Bereich gewählt, welcher die Person und einen Teil des Hintergrundes abdeckt. Dieser darf nicht zu klein gewählt werden, damit die Form gut zu erkennen ist und auch die Varianznormalisierung gut funktioniert. Die Berechnung der Merkmalswerte wird nur auf dem Intensitätskanal des i1i2i3-Farbraumes durchgeführt, da hier die meisten Forminformationen enthalten sind. Der Auswertebereich der Rechteckmerkmale ist kleiner und beschränkt sich im wesentlichen auf den Körper der Person. Die Berechnung erfolgt auf allen drei Kanälen des i1i2i3-Farbraumes. Welche Kombination aus Merkmalen, Farbkanälen und Auswertebereichen die optimalen Detektionsergebnisse liefert, lässt sich aufgrund der Vielzahl an Kombinationsmöglichkeiten nur näherungsweise durch Austesten verschiedener Varianten bestimmen (s. Experimente in Abs. 5.2.2).

3.4. Klassifikator

Im Gegensatz zur expliziten Modellierung, bei der das Objektmodell durch Regeln, Zugehörigkeitsfunktionen und Parameter definiert wird, muss es bei der impliziten Modellierung durch einen Klassifikator anhand von Beispielen gelernt werden. Die damit verbundenen Aspekte werden nachfolgend näher behandelt.

3. Objekterkennung

3.4.1. Wahl des Klassifikators und Merkmalsreduktion

Da die Objekterkennung als Zwei-Klassen-Problem aufgefasst werden kann (Objekt vs. Hintergrund), werden hierfür diskriminative Klassifikationsverfahren eingesetzt. Diese modellieren nicht die Verteilung der beiden Klassen im Merkmalsraum, sondern direkt die Entscheidungsgrenze (Jain u. a., 2000). Als besonders geeignet haben sich Verfahren auf Basis von SVM (*Support Vector Machines*, Vapnik (2000); Burges (1998)) oder Boosting herausgestellt.

In dieser Arbeit wird der AdaBoost-Klassifikator eingesetzt (Freund und Schapire, 1997), welcher wie alle Boosting-Verfahren zur Gruppe der Ensemble-Methoden gehört und somit die Antworten mehrerer Klassifikatoren zur Entscheidungsfindung nutzt. Im Detail werden für AdaBoost mehrere schwache Basisklassifikatoren (*weak learner*) in einem adaptiven Verfahren zu einem leistungsfähigen Klassifikator kombiniert. Ein Basisklassifikator muss dabei nur etwas besser als der Zufall sein und kann z. B. aus einem Schwellwert oder einem Entscheidungsbaum mit wenigen Knoten bestehen.

Das Training läuft nun folgendermaßen ab. Zu Beginn liegen Beispiele für beide Klassen in Form von Merkmalsvektoren \mathbf{x} vor, welche alle das selbe Gewicht erhalten. Dann wird derjenige Basisklassifikator h^{weak} ausgewählt, welcher auf den gewichteten Beispielen die beste Klassifikationsleistung erzielt. Die Anzahl der hierbei auftretenden Fehler bestimmt dessen Stimmengewicht α im finalen Klassifikator. Anschließend erfolgt eine Neubewertung aller Beispiele, so dass die aktuell falsch klassifizierten ein größeres Gewicht und die korrekt klassifizierten ein niedrigeres erhalten. Diese Prozedur wird so oft wiederholt, bis entweder der Test- oder Trainingsfehler konvergiert oder eine manuell gesetzte Anzahl an Iterationen erreicht wurde. Die Umgewichtung der Trainingsbeispiele führt in jeder Iteration zu einer Fokussierung auf besonders schwierige Beispiele und zu einer schnellen Steigerung der Klassifikationsleistung.

Die Bewertung eines bestimmten Beispiels erfolgt anschließend auf Basis des gewichteten Mittels aller ausgewählten Basisklassifikatoren. Da jeder von ihnen entweder -1 oder $+1$ zurückgibt¹, lässt sich am Vorzeichen der gemeinsamen Antwort die Klassenzugehörigkeit ablesen. Normiert man die Antwort zusätzlich mit dem Gesamtgewicht aller Basisklassifikatoren, so erhält man einen deutlich nützlicheren, von -1 und $+1$ begrenzten Wert, der als Konfidenz für eine Klasse interpretiert werden kann:

$$conf(\mathbf{x}) = \frac{\sum \alpha_n \cdot h_n^{weak}(\mathbf{x})}{\sum \alpha_n} \quad (3.1)$$

Die bisherige Beschreibung hat bereits einige Vorteile des AdaBoost-Verfahrens verdeutlicht. Es ist sehr leicht zu verstehen, kann gleichzeitig mit unterschiedlichen Arten von Merkmals-typen umgehen und besitzt keine sensiblen Parameter. Darüber hinaus neigt das Verfahren nicht zur Überanpassung an die Trainingsdaten und führt während des Trainings eine Merkmalsreduktion durch. Vor allem der letzte Punkt ist in dieser Arbeit besonders wichtig. Bricht der Trainingsprozess bspw. bereits nach 20 Iterationen ab und wurden nur Schwellwerte als Basisklassifikatoren verwendet, so müssen während der Objekterkennung auch nur maximal 20 unterschiedliche Merkmale innerhalb des Detektors berechnet werden. AdaBoost führt also eine Merkmalsreduktion durch sequentielle Auswahl der nützlichsten Merkmale durch (sog. *Wrapper-Verfahren*, Blum und Langley (1997)). Diese Fähigkeit erlaubt es, bei der Entwicklung des Detektors nicht exakte Merkmale, sondern nur Merkmalsarten und Auswertebereiche festlegen zu müssen. In dieser Arbeit wird die Anzahl an Merkmalen von 21.000 möglichen auf die 100 nützlichsten reduziert (s. Experiment in Abs. 5.2.3).

Dort wo der Detektor zu Objekterkennung eingesetzt wird, müssen alle für die Klassifizierung benötigten Merkmalswerte aus dem Bild extrahiert werden. Dieser Prozess lässt sich

¹Häufig stehen auch die Werte 1 und 0 für Objekt und Hintergrund. Dann liegt die Entscheidungsgrenze entsprechend bei 0.5.

deutlich beschleunigen, indem eine Kaskadenstruktur aus mehreren unterschiedlich komplexen Klassifikatoren aufgebaut wird (Viola und Jones, 2001). Der Nachteil dieses Ansatzes ist jedoch der Umstand, dass sich das Training erschwert. In dieser Arbeit wird daher auf die Implementierung einer Kaskadenstruktur verzichtet und stattdessen ein einzelner Klassifikator eingesetzt.

3.4.2. Beispiele sammeln und Klassifikator trainieren

Für das Training des Klassifikators werden Beispiele sowohl der Objekt- als auch der Hintergrundklasse benötigt. Da mit AdaBoost ein diskriminatives Lernverfahren eingesetzt wird, müssen die Beispiele nicht die gesamte Verteilung der Klassen im Merkmalsraum beschreiben. Es würde ausreichen, repräsentative Beispiele entlang der Entscheidungsgrenze bereitzustellen, was jedoch a priori nicht ohne Weiteres möglich ist.

Aus diesem Grund wird zuerst damit begonnen, die Klasse mit der geringeren Variation durch ausreichend viele Beispiele repräsentativ zu beschreiben. Ein Operateur markiert daher in Trainingsbildern die Position des Mittelpunktes von Personen immer auf die gleiche Art und Weise. Der Merkmalsvektor jedes Beispiels ergibt sich, indem der Detektor mit seinem Bezugspunkt an der gleichen Stelle ausgewertet wird. Da die gesamte Variabilität der Objektklasse abgebildet werden muss, sollte der Satz an Beispielen ausreichend groß sein und auch Personen mit und ohne Schatten umfassen. Zudem sollten auch schwierige Beispiele enthalten sein, jedoch keine, in denen die Position der Einzelperson nicht eindeutig erkennbar ist wie z. B. in dichtem Gedränge (vgl. Abb. 3.2 und 3.3).

Die unendlich vielfältige Hintergrundklasse vollständig mit Beispielen zu beschreiben, ist nicht möglich. Da dies bereit für die Objektklasse geschehen ist, lassen sich nun die relativ wenigen, zur Modellierung der Entscheidungsgrenze notwendigen Hintergrundbeispiele in einem iterativen Verfahren automatisch bestimmen (Sung und Poggio, 1998; Grabner u. a., 2008). Hierfür wird der Klassifikator zuerst mit allen Objektbeispielen und wenigen manuell ausgewählten Hintergrundbeispielen trainiert. Anschließend wird der Detektor auf Bilder ohne Objekte angewandt. Aus allen Fehldetektionen werden dann zufällig einige wenige ausgewählt und als zusätzliche Hintergrundbeispiele in einem erneuten Training genutzt. Das Verfahren wird so lang wiederholt, bis die Anzahl der Fehldetektionen in den Bildern ohne Objekte niedrig genug ist. Da in jeder Iteration ausschließlich wenige, schwierige Hintergrundbeispiele entlang der Entscheidungsgrenze gesammelt werden, verbessert sich der Klassifikator mit jedem Mal und die Anzahl an Beispielen steigt nur langsam an.

Diese Methode zum unüberwachten Lernen hat jedoch einen Nachteil. Sind in den genutzten Bildern Objekte mit großer visueller Ähnlichkeit zur gesuchten Objektklasse enthalten, werden diese automatisch zur Menge der Hintergrundbeispiele hinzugefügt und verschlechtern die Leistung des Klassifikators. Dieses Problem tritt in allen Anwendungsbereichen auf, es äußert sich jedoch bei der Detektion von Personen in Luftbildern besonders stark, da diese sehr viele personenähnliche Objekte enthalten. Obwohl bereits die *Gentle*-Variante von AdaBoost (Friedman u. a., 2000) genutzt wird, welche weniger anfällig gegenüber Fehlern in den Trainingsdaten ist, so lassen sich bessere Ergebnisse erzielen, wenn die Beispiele beider Klassen nicht vermischt werden (s. Experiment in Abs. 5.2.3).

In dieser Arbeit wird das Problem mit Hilfe der vom AdaBoost-Klassifikator zurückgegebenen Konfidenz gelöst. Werte von +1 oder -1 klassifizieren das jeweilige Beispiel eindeutig als Person bzw. Hintergrund. Je näher der Wert Richtung Null geht, umso unsicher ist die Entscheidung. Sollen nun während des iterativen Trainingsprozesses neue Beispiele der Hintergrundklasse zugeordnet werden, geschieht deren Auswahl nicht mehr zufällig, sondern auf Basis ihrer jeweiligen Konfidenz. Fehldetektionen mit positiven Werten nahe Null werden bevorzugt selektiert. Da die Konfidenz personenähnlicher Detektionen näher bei +1 liegt, wird

3. Objekterkennung

so die Wahrscheinlichkeit deutlich reduziert, dass diese als Hintergrundbeispiele ausgewählt werden und die Klassentrennbarkeit reduzieren.

Nutzt man zum automatischen Sammeln von Hintergrundbeispielen nur Bilddaten ohne die gesuchten Objekte, würde der Detektor viele Falschalarme erzeugen. In dem hier betrachteten Anwendungsfall würden diese vor allem in der Nähe von Personen z. B. in deren Schatten auftreten. Da der Detektor diese Bereiche während des Trainings nicht zu sehen bekommen hat, kann er sie nicht eindeutig als Hintergrund ausweisen. Diese Problem wird gelöst, indem auch Luftbilder mit Personen im Trainingsprozess genutzt werden. Zuvor müssen jedoch alle enthaltenen Personen manuell markiert und anschließend ausmaskiert werden.

3.5. Lokalisierung

In den vorhergehenden Abschnitten dieses Kapitels wurde ein implizites visuelles Objektmodell erstellt. Nun soll es darum gehen, wie sich damit Personen in Luftbildern lokalisieren lassen. Wie bei allen modellgetriebenen Verfahren erfolgt auch hier die Objekterkennung, indem das Bild nach Orten abgesucht wird, die eine hohe Ähnlichkeit mit dem vorliegenden Modell aufweisen.

Im Detail wird der Detektor als gleitende Filtermaske über das gesamte Luftbild geschoben und im Abstand von einem Pixel ausgewertet (*Brute-Force-Ansatz*). An jeder Stelle werden alle notwendigen visuellen Merkmale extrahiert und an den Klassifikator übergeben. Dieser berechnet einen Konfidenzwert, welche an der Stelle des Bezugspunktes in ein Ergebnisbild eingetragen wird (vgl. Abb. 3.1). Eine Drehung oder Skalierung des Detektors ist aufgrund der Vorverarbeitungsschritte und des einheitlichen Aussehens der Personen nicht notwendig. Es reicht daher aus, das Luftbild ein einziges Mal vollständig abzusuchen.

Als Ergebnis der Filterung erhält man ein Konfidenzbild, in welchem lokale Maxima zwischen Null und Eins potentielle Objektpositionen markieren. Um diese subpixelgenau ermitteln zu können, wird das Konfidenzbild zuerst mit einem 3 x 3 Pixel großen Gaußfilter geglättet. Dieser Schritt ist notwendig, um aus den einzelnen Konfidenzwerten, welche an diskreten Stellen bestimmt worden sind, eine kontinuierliche, räumliche Verteilung abzuleiten. Anschließend lassen sich lokale Maxima subpixelgenau bestimmen, indem das Konfidenzbild in jedem Punkt durch ein zweidimensionales Polynom approximiert wird. Besitzt das jeweilige Maximum sowohl eine ausreichende Krümmung als auch eine Konfidenz über dem Detektionsschwellwert, wird an dieser Stelle eine Einzelperson erkannt. Wurde das Luftbild in einem Vorverarbeitungsschritt geometrisch verändert, so müssen abschließend die Detektionen noch entgegengesetzt transformiert werden.

Die vollständige Suche mit gleitendem Detektor ist relativ zeitaufwändig, besonders wenn ganze Luftbilder analysiert werden sollen. Es gibt jedoch zahlreiche Ansätze, diesen Prozess zu beschleunigen. Eine der effektivsten Methode besteht darin, den Suchbereich von vornherein einzuschränken z. B. durch Vorgabe einer bestimmten Region oder durch Ausschluss unwahrscheinlicher Orte auf Basis eines geographischen Informationssystems. Soll eine Sequenz ausgewertet werden, kann auch das vorhergehende Konfidenzbild wertvolle Hinweise liefern. Die Anzahl der Stellen, an denen der Detektor ausgewertet werden muss, ließe sich weiter reduzieren, wenn erst in einem groben Raster auffällige Stellen identifiziert und anschließend nur diese vollständig untersucht würden. Des Weiteren ließen sich auch eindeutige Hintergrundbereiche mit simplen Segmentierungsverfahren schnell ermitteln und für weitere Analysen ausschließen. Ein ähnliches Resultat, jedoch auf eleganterem Wege, erreicht man, indem für die Objekterkennung eine Kaskade aus Klassifikatoren eingesetzt wird (s. Abs. 3.4.1). Deutlich einfacher realisieren lässt sich dagegen die Parallelisierung der Suchaufgabe auf mehrere Prozessoren.

3.6. Stochastische Modellierung

In diesem Abschnitt wird gezeigt, wie sich die Detektionsergebnisse stochastisch modellieren lassen. Dies ist eine notwendige Voraussetzung, damit im nachfolgenden Kapitel 4 die Objekterkennung mit implizitem Modell in das MHT-Verfahren integriert werden kann. Da die Modellierung jedoch unabhängig vom Tracking-Verfahren ist und auch zur Verbesserung der Objekterkennung im Einzelbild genutzt werden kann, wird sie bereits in diesem Kapitel behandelt.

Ein wichtiges Argument für die stochastische Modellierung der Detektionen ist, dass sich eine optimale Entscheidung im Sinne der Bayes'schen Schätztheorie treffen lässt (Duda u. a., 2001), statt Objekte und Hintergrund nur anhand eines manuell gesetzten, festen Schwellwertes zu unterscheiden. Im Zuge dessen lassen sich auch Vorwissen über unterschiedliche Klassenwahrscheinlichkeiten und Kontextinformationen leichter und theoretisch fundiert in das Detektionsverfahren integrieren (vgl. Perko und Leonardis (2010)). Des Weiteren wird es einfacher, eine Rückweisungsklasse zu definieren, welche Detektionen enthält, die anhand der Informationen aus einem einzigen Bild nicht eindeutig zugeordnet werden können.

In der Objekterkennung unterscheidet man die beiden disjunkten Ereignisse H_1 , das gesuchte Objekt liegt vor und H_0 , das Objekt liegt nicht vor, mit:

$$P(H_1) + P(H_0) = 1 \quad (3.2)$$

Tritt eine Detektion D mit dem Wert s auf, lässt sich mit Hilfe der Maximum-a-posteriori-Methode (MAP) eine optimale Entscheidung für eines der beiden Ereignisse treffen. Hierfür wird jenes ausgewählt, welches die größte a posteriori Wahrscheinlichkeit besitzt:

$$\hat{H}_{MAP} = \operatorname{argmax}_{H_i} P(H_i | s, D), \quad i \in \{0, 1\} \quad (3.3)$$

Mit Hilfe des Bayestheorems und den Rechenregeln für bedingte Wahrscheinlichkeiten lässt sich die a posteriori Wahrscheinlichkeit eines Ereignisses folgendermaßen darstellen:

$$\begin{aligned} P(H_i | s, D) &= \frac{P(s, D | H_i) \cdot P(H_i)}{P(s, D)} \\ &\propto P(s, D | H_i) \cdot P(H_i) \\ &\propto P(s | D, H_i) \cdot P(D | H_i) \cdot P(H_i) \end{aligned} \quad (3.4)$$

Da die Wahrscheinlichkeit $P(s, D)$ für alle Ereignisse gleich ist, hat sie keinen Einfluss auf die MAP-Entscheidung und wird nicht weiter betrachtet. Übrig bleiben die a priori Wahrscheinlichkeit des Ereignisses $P(H_i)$, die Wahrscheinlichkeit in dieser Situation eine Detektion zu erhalten $P(D | H_i)$ und die Wahrscheinlichkeit $P(s | D, H_i)$, dass diese den Wert s hat. Alle drei Wahrscheinlichkeiten müssen für beide Ereignisse definiert werden, um eine optimale MAP-Entscheidung treffen zu können. Dies ist anders als in den Arbeiten von (Platt, 1999) und (Niculescu-Mizil und Caruana, 2005). Hier wird die Verteilung der a posteriori Wahrscheinlichkeit direkt modelliert, jedoch unter der Annahme, dass die a priori Klassenwahrscheinlichkeiten konstant sind und sich im Training bestimmen lassen.

Diese Voraussetzungen sind in den meisten Anwendungen der Objekterkennung nicht gegeben. Die a priori Klassenwahrscheinlichkeit hängt stattdessen von Vorwissen und Kontextinformationen ab und kann sich je nach Situation ändern. Sind diese Informationen nicht vorhanden, wird der Wert meist auf 1 gesetzt oder weggelassen. Die Detektionswahrscheinlichkeit ist abhängig von den für die Objekterkennung genutzten Methoden und der Schwierigkeit der beobachteten Szene. Ihr Wertebereich lässt sich im Zuge einer Evaluation des Detektionsprozesses ermitteln (s. Abs. 5.2.4).

Die Wahrscheinlichkeitsdichte des Detektionswertes $P(s | D, H_i)$ lässt sich dagegen analytisch beschreiben. Wie in Abbildung 3.6 zu sehen ist, sind die Werte für beide Ereignisse annähernd normalverteilt. Diese Beobachtung lässt sich mit dem *Zentralen Grenzwertsatz der*

3. Objekterkennung

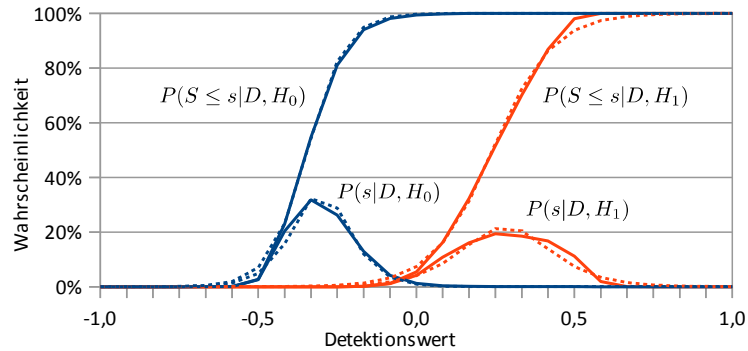


Abbildung 3.6.: Dichte und kumulative Verteilung des Detektionswertes s von korrekten Detektionen (D, H_1) und Falschalarmen (D, H_0) basierend auf dem beschriebenen Verfahren zum Erkennen von Einzelpersonen in Luftbildsequenzen. Die wahre Verteilung (durchgezogene Linie) wird mit einer Sigmoidfunktion (gestrichelte Linie) modelliert.

Wahrscheinlichkeitsrechnung (Niemeier, 2008) erklären. Werden vom AdaBoost-Klassifikator nur ausreichend viele Basisklassifikatoren genutzt, so nähert sich deren gemeinsame Antwort immer einer Normalverteilung an. Statt jedoch deren Parameter direkt zu bestimmen, wird in dieser Arbeit zuerst die kumulative Verteilungsfunktion $P(S \leq s|D, H_i)$ ermittelt. Diese lässt sich mittels einer Sigmoidfunktion beschreiben:

$$P(S \leq s|D, H_i) \approx \text{sig}(s) = \frac{1}{2}[1 + \tanh(a \cdot s + b)] \quad (3.5)$$

Die Dichtefunktion ergibt sich dann durch Ableiten der kumulativen Verteilung und besitzt die selben Parameter a und b :

$$P(s|D, H_i) \approx \text{sig}'(s) = \frac{a}{2}[1 - \tanh^2(a \cdot s + b)] \quad (3.6)$$

Die beiden Parameter können leicht mit Hilfe einer Kleinste-Quadrate-Regression (Niemeier, 2008) bestimmt werden. Um dies zu tun, müssen für beide Ereignisse ausreichend viele Detektionswerte gesammelt werden. Für Falschalarme kann dies geschehen, indem der Detektor auf Bilder ohne die gesuchten Objekte angewandt wird. Da kein Schwellwert genutzt wird, kommt auf diese Weise schnell eine sehr große Anzahl an Detektionswerten zusammen. Für korrekte Detektionen durchsucht man zuerst Trainingsbilder mit dem Detektor und sammelt anschließend die Detektionswerte in nächster Nähe zu manuell markierten Objektpositionen. Die für die Regression benötigten Beobachtungen (y_i, x_i) ergeben sich dann wie folgt:

$$y_i = P(S \leq s_i) = \frac{n(s \leq s_i)}{N}, \quad x_i = s_i \quad (3.7)$$

Die Gesamtzahl der gesammelten Beispielwerte eines Ereignisses ist N und der Anteil davon mit einem Wert kleiner gleich s_i ist n . Hier zeigt sich der Vorteil des Umweges über die kumulative Verteilung. Da die Detektion keine diskreten sondern kontinuierliche Werte zurückgibt, ist es auf diese Weise deutlich leichter, die für die Regression benötigten Beobachtungen zu bestimmen. Für die in Abb. 3.6 dargestellten Ergebnisse wurden Detektionswerte von 165.000 Falschalarmen und 2.500 korrekten Detektionen genutzt. Das Bestimmtheitsmaß der Regression (Niemeier, 2008) erreicht für beide Verteilungen einen Wert von 0,999 und bestätigt die Modellierung mittels Sigmoidfunktion.

4. Objektverfolgung

Im letzten Kapitel wurde detailliert die Objekterkennung im Einzelbild auf Basis eines impliziten Modells erläutert. Die gewonnenen Detektionen bilden die Grundlage für das nun folgende Verfahren zur Objektverfolgung. Wie in Abschnitt 2.2.1 beschrieben, bringt die Aufgabe, Personen in Luftbildsequenzen zu verfolgen, zahlreiche Schwierigkeiten mit sich. Aus diesem Grund wird in dieser Arbeit das MHT-Verfahren genutzt, welches alle Voraussetzungen erfüllt, um mit den gegebenen Herausforderungen umzugehen. In den nachfolgenden Abschnitten wird beschrieben, wie das Objektmodell für das Tracking erweitert wird, wie sich die Bewertungsfunktion für Zuordnungen zusammensetzt und wie mehrere globale Hypothesen in einem effektiven, sequentiellen Ansatz verfolgt werden können.

4.1. Veränderlichkeit des Objektmodells

Für die Detektion von Einzelpersonen in Luftbildern wurde ein Modell entwickelt, welches deren Aussehen und Position beschreibt. Nun ist es notwendig zu definieren, ob und wie sich die Komponenten des Objektmodells mit der Zeit verändern können.

4.1.1. Aussehen

Da sich Personen in Luftbildern visuell kaum unterscheiden, wird ihr Aussehen durch ein gemeinsames Modell beschrieben. Nun wäre es denkbar, auf dessen Grundlage für jede Person während des Trackings ein individuelles Modell abzuleiten (vgl. Jüngling (2011); Kalal u. a. (2012)). Sobald eine Person in einem weiteren Bild wiedererkannt wird, könnte diese Detektion als Beispiel genutzt werden, um das visuelle Modell durch ein erneutes Training zu individualisieren. Da Personen sich jedoch sehr stark ähneln, wird in dieser Arbeit davon abgesehen (vgl. Abb. 3.2).

Des Weiteren ließen sich Detektionen in neuen Bildern auch nutzen, um das allgemeine Personenmodell besser an die Bedingungen in der jeweiligen Sequenz anzupassen. Solche ein Ansatz wird z. B. in (Grabner u. a., 2008) zur Detektion von Fahrzeugen in Luftbildern beschrieben. Hierbei besteht die Schwierigkeit in der automatischen Auswahl von korrekten Beispielen für das sequentielle, unüberwachte Training des Klassifikators. Auf solch ein Vorgehen wird verzichtet und stattdessen mehr Gewicht auf ein umfangreiches initiales Training gelegt (s. Abs. 3.4.2).

Die Personenfarbe könnte in einigen Fällen zur Unterscheidung von Individuen dienen. Sie kann sich jedoch aufgrund der Kamera- und Personenbewegung abrupt ändern und ist deshalb kein zuverlässiges Merkmal. Zusammengefasst wird in dieser Arbeit ein einheitliches visuelles Modell genutzt, dass während des Trackings unverändert bleibt. Daher lassen sich Informationen über das Aussehen der Personen nicht für die Lösung des Zuordnungsproblems verwenden.

4.1.2. Position und Bewegung

Detektierte Personen werden durch ihren Mittelpunkt repräsentiert. Da hier von georeferenzierten Luftbildern ausgegangen wird, ist dessen Position nicht nur in Bild- sondern auch

4. Objektverfolgung

in Weltkoordinaten bekannt. Die Veränderung der Objektposition mit die Zeit sowie deren Prädiktion und Korrektur werden im Rahmen eines rekursiven Bayes'schen Ansatzes zur Zustandsschätzung behandelt (Blackman und Popoli, 1999). Die Basis bildet eine diskrete Kalman-Filterung (Niemeier, 2008), welche für normalverteilte Zustände und Messungen optimale Schätzwerte liefert. Obwohl diese Bedingung meist nicht exakt erfüllt wird, liefert der Ansatz in der Regel ausreichend gute Ergebnisse.

Bei einer Aufnahme Frequenz von zwei oder mehr Bildern pro Sekunde ändert sich die Bewegung einer Person nur wenig zwischen zwei Detektionen. Aus diesem Grund wird ein einfaches lineares Bewegungsmodell genutzt (s. Gl. 4.1) und der Zustandsvektor \vec{x} um die Geschwindigkeit in X- und Y-Richtung (\dot{x}, \dot{y}) ergänzt. Bei diesem Ansatz wird zudem von einer stückweise linearen Geschwindigkeit ausgegangen (*near-constant motion model*). Alle Abweichungen von dieser Annahme z. B. durch Abbiegen, Beschleunigen oder Abbremsen, werden nur implizit in Form eines erhöhten Prozessrauschens \vec{r} behandelt.

$$\vec{x}_k = T_{k-1} \cdot \vec{x}_{k-1} + \vec{r} \quad (4.1)$$

$$\vec{x} = \begin{pmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{pmatrix} \quad (4.2)$$

$$T = \begin{pmatrix} 1 & 0 & \delta t & 0 \\ 0 & 1 & 0 & \delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (4.3)$$

$$\vec{r} \sim \mathcal{N}(0, \sigma^2) \quad (4.4)$$

Initialisiert wird der Zustandsvektor jeder neu erkannten Personen mit der Position der jeweiligen Detektion. Die Varianz der Zustandswerte ist ebenfalls gleich der Varianz der Detektion, welche anhand von Trainingsdaten bestimmt werden kann. Die Unsicherheit der Geschwindigkeit kann auf Basis einer oder zwei erfolgter Detektionen ermittelt werden (Mallick und La Scala, 2008). In dieser Arbeit werden beide Verfahren kombiniert. Zuerst wird die Geschwindigkeit mit 0m/s und deren Varianz mit einem Drittel der quadrierten, maximalen Objektgeschwindigkeit angegeben (*single point initialization*). Wird die Person dann ein zweites Mal erkannt, werden Geschwindigkeit und Varianz aus der Differenz der Positionen neu berechnet (*two-point differencing*). Dieses Vorgehen ermöglicht bereits nach der ersten Detektion eine Prädiktion mit Hilfe des Kalman-Filters ohne dass die anfänglich falsche Geschwindigkeit anschließend einen zu großen Einfluss ausübt.

Das Prozessrauschen wird als normalverteilt und zeitlich unkorreliert angenommen (s. Gl. 4.4). Als Varianz der Position wird die Ungenauigkeit der Georeferenzierung der Bilddaten genutzt, welche sich aus Trainingssequenzen mit Hilfe von Passpunkten ermittelt lässt. Das Rauschen der Geschwindigkeit lässt sich ebenfalls im Vorfeld bestimmen, indem die Abweichung von Referenztrajektorien von einer linearen Bewegung analysiert wird.

Die wahren Zustände der Personen können nicht direkt, sondern nur mit Hilfe von Messungen bestimmt werden. Diesen Vorgang und die dabei auftretenden Fehler werden durch ein Beobachtungsmodell beschrieben. Dieses fällt hier sehr einfach aus, da außer der Position der Detektion keine weiteren Messwerte anfallen. Das Messrauschen kann mit Hilfe von Detektionen in Trainingsbildern ermittelt werden. Würde man den Optischen Fluss zwischen aufeinanderfolgenden Bildern berechnen, könnte man diesen als Messung für die Objektgeschwindigkeit heranziehen. Dieser Ansatz wird jedoch nicht verfolgt, da zum einen die Berechnung des Optischen Flusses zwischen zwei Luftbildern sehr lang dauert und zum anderen die Geschwindigkeit von sich schnell bewegenden Personen auf diese Weise nicht zuverlässig bestimmt werden kann.

Für Prädiktion und Korrektur der Objektzustände werden die bekannten Formeln des Kalman-Filters genutzt (Niemeier, 2008). Die Vorhersage erfolgt mit Hilfe des linearen Modells und unabhängig für jede Person. Es wird davon ausgegangen, dass diese Vereinfachung der Realität die Ergebnisse nur wenig beeinflusst. Erfolgt nach einer Prädiktion keine Korrektur z. B. weil eine Person kurzzeitig verdeckt ist, werden die geschätzten Zustände inklusive ihrer Genauigkeiten direkt übernommen und eventuell erneut prädiziert (vgl. Abs. 4.2.2). Sollte eine Person mehrfach nicht erkannt worden sein, nimmt die Unsicherheit ihrer Zustände auf diese Weise stetig zu.

4.2. Stochastische Bewertungsfunktion

Jedes Tracking-Verfahren benötigt eine Methode zur Bewertung potentieller Zuordnungen zwischen Detektionen und bekannten Objekten. Nur so kann entschieden werden, ob überhaupt eine Zuordnung erfolgen darf und falls ja, welche von mehreren Varianten die erfolgversprechendste ist. Beim MHT-Verfahren wird die wahrscheinlichste Erklärung für den Ursprung aller Detektionen und die Zustände aller Objekte für eine gewisse Zeitspanne gesucht. Die zugehörige Bewertungsfunktion, mit welcher sich die Wahrscheinlichkeit jeder Zuordnung berechnen lässt, wird in diesem Abschnitt behandelt.

4.2.1. Hypothesenwahrscheinlichkeit

Im MHT-Kontext hat eine Hypothese eine andere Bedeutung als in anderen Tracking-Verfahren. Hier ist eine einzelne Hypothese Θ_i^t eine globale, konsistente Erklärung für alle Objekte und alle Detektionen vom Beginn der Beobachtung bis zum Zeitpunkt t . Sie besteht aus einem Satz von Zuordnungsereignissen θ_k^t (*association events*), welche jeweils für ein bestimmten Zeitpunkt gelten:

$$\Theta_i^t = \{\theta_k^{1:t}\} \quad (4.5)$$

Die verwendeten Indizes symbolisieren hier und in den nachfolgenden Gleichungen, dass jeweils ein bestimmtes Element aus der Menge aller zum jeweiligen Zeitpunkt vorhandenen Elemente ausgewählt wurde. Jedes Zuordnungsereignis enthält eine mögliche Lösung des zugehörigen Korrespondenzproblems. Es besteht daher aus einer Liste von Zuordnungsoptionen (*assignment options*), die den Ursprung aller Detektionen und die Zustände aller Objekte im Bild zum Zeitpunkt t erklären. Folgende vier Optionen stehen zur Verfügung:

1. Detektion ist ein Falschalarm (*false alarm*, FA),
2. Detektion stammt von einem neuen Objekt (*new target*, NT),
3. Detektion kommt von einem bereits bekannten Objekt (*detected target*, DT),
4. Bekanntes Objekt wurde nicht erkannt (*lost target*, LT).

Die Aufgabe beim MHT-Verfahren besteht nun darin, aus allen möglichen Hypothesen Θ_i^t , diejenige auszuwählen, welche zum Zeitpunkt t unter Berücksichtigung alle bisherigen Beobachtungen $Z^{1:t}$ die größte a posteriori Wahrscheinlichkeit (MAP) besitzt:

$$\Theta_{MAP}^t = \operatorname{argmax}_{\Theta_i^t} P(\Theta_i^t | Z^{1:t}) \quad (4.6)$$

In jeder Hypothese sind die Zuordnungsereignisse bedingt abhängig von ihren Vorgängern. Die Wahrscheinlichkeit einer einzelnen Hypothese kann daher in der folgenden sequentiellen, modularen Weise dargestellt werden (Reid, 1979):

$$P(\Theta_i^t | Z^{1:t}) = \frac{1}{c_0} \cdot P(Z^t | \theta_h^t, \Theta_g^{t-1}, Z^{1:t-1}) \cdot P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) \cdot P(\Theta_g^{t-1} | Z^{1:t-1}) \quad (4.7)$$

4. Objektverfolgung

Der konstante Faktor c_0 dient allein zur Normalisierung und hat keinen Einfluss auf die MAP-Entscheidung. Davon abgesehen setzt sich die Gleichung aus drei Teilen zusammen:

1. der Wahrscheinlichkeit die Beobachtungen $Z^t = \{z_j^t\}$ zum Zeitpunkt t zu erhalten, gegeben aller bisherigen Beobachtungen und Zuordnungsereignisse inklusive des aktuellen,
2. der Wahrscheinlichkeit des aktuellen Zuordnungsereignisses, gegeben aller bisherigen Ereignisse und Beobachtungen sowie
3. der Wahrscheinlichkeit der Elternhypothese zum vorhergehenden Zeitpunkt.

Die sequentielle Schreibweise ermöglicht es, dass zu jedem Zeitpunkt nur die ersten beiden Teile ausgewertet werden müssen. Diese können wiederum in eine Form gebracht werden, welche die Beiträge der unterschiedlichen Optionen im aktuellen Zuordnungsereignis deutlicher herausstellt (Herleitung im Anhang A.1):

$$P(\Theta_i^t | Z^{1:t}) = \frac{1}{c_1} \cdot \prod_j^{N_{NT}} f_{NT}(z_j^t) \cdot \prod_j^{N_{FA}} f_{FA}(z_j^t) \cdot \prod_j^{N_{DT}} [f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P_D] \cdot \prod_j^{N_{LT}} (1 - P_D) \cdot P(\Theta_g^{t-1} | Z^{1:t-1}) \quad (4.8)$$

Die Anzahl Detektionen, welche neuen Objekten, Falschalarmen und bekannten Objekten zugeordnet werden und die Anzahl verschwundener Objekte sind mit N_{NT} , N_{FA} , N_{DT} und N_{LT} angegeben. Die Wahrscheinlichkeitsdichtefunktionen von neuen Objekten, Falschalarmen und wiederentdeckten Objekten werden mit f_{NT} , f_{FA} und f_{DT} bezeichnet und die Wahrscheinlichkeit ein vorhandenes Objekt zu erkennen mit P_D . Alle hypothesenunabhängige Terme sind in der Konstante c_1 zusammengefasst.

4.2.2. Hypothesenwert und Track Management

Gleichungen 4.7 und 4.8 geben die Wahrscheinlichkeit einer globalen Hypothese im MHT-Kontext an. Sie könnten daher genutzt werden, um alternative Hypothesen zu bewerten und die wahrscheinlichste von ihnen zu ermitteln. Es ist jedoch vorteilhaft, hierfür stattdessen den Hypothesenwert (*hypothesis score*) zu nutzen (Sittler, 1964; Bar-Shalom u. a., 2007) wie es z. B. in der trajektorien-orientierten Variante des MHT-Verfahrens getan wird (Blackman und Popoli, 1999). Der Wert einer Hypothese $L_{\Theta_i^t}$ ergibt sich aus dem Logarithmus ihrer Wahrscheinlichkeit, erweitert mit der logarithmierten Wahrscheinlichkeit der Alternativhypothese Θ_0^t , welche alle Beobachtungen zu Falschalarmen erklärt:

$$L_{\Theta_i^t} = \ln P(\Theta_i^t | Z^{1:t}) = \ln \frac{P(\Theta_i^t | Z^{1:t})}{P(\Theta_0^t | Z^{1:t})} + \ln P(\Theta_0^t | Z^{1:t}) = \ln \frac{P(\Theta_i^t | Z^{1:t})}{P(\Theta_0^t | Z^{1:t})} + c \quad (4.9)$$

Ersetzt man nun die Hypothesenwahrscheinlichkeiten durch Gleichung 4.8 und fasst alle konstanten Terme in c_2 zusammen, so erhält man:

$$L_{\Theta_i^t} = c_2 + \sum_j^{N_{FA}} \ln \frac{f_{FA}(z_j^t)}{f_{FA}(z_j^t)} + \sum_j^{N_{NT}} \ln \frac{f_{NT}(z_j^t)}{f_{FA}(z_j^t)} + \sum_j^{N_{DT}} \ln \frac{f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P_D}{f_{FA}(z_j^t)} + \sum_j^{N_{LT}} \ln(1 - P_D) + L_{\Theta_g^{t-1}} \quad (4.10)$$

Die einzelnen Beiträge ΔL der vier möglichen Zuordnungsoptionen zum Hypothesenwert sind somit:

$$\Delta L_{FA} = 0 \quad (4.11)$$

$$\Delta L_{NT} = \ln \frac{f_{NT}(z_j^t)}{f_{FA}(z_j^t)} \quad (4.12)$$

$$\Delta L_{DT} = \ln \frac{f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P_D}{f_{FA}(z_j^t)} \quad (4.13)$$

$$\Delta L_{LT} = \ln(1 - P_D) \quad (4.14)$$

Bei der Berechnung der Beiträge neuer und wiedererkannter Objekte ΔL_{NT} und ΔL_{DT} können numerische Probleme bei der Verhältnisbildung mit Wahrscheinlichkeiten nahe Null auftreten. Eine übliche Lösung dieses Problems besteht in der Begrenzung des Wertebereiches z. B. auf $1e-4 < x/y < 1e4$.

Der Wert einer Hypothese, wie er in Gleichung 4.10 definiert ist, besitzt einige entscheidende Vorteile. Da er eine vollständig dimensionslose Größe ist, erlaubt er einen direkten Vergleich von Hypothesen mit unterschiedlicher Anzahl enthaltener Objekte. Dies ist nicht der Fall bei der in Gleichung 4.8 präsentierten Hypothesenwahrscheinlichkeit (Bar-Shalom u. a., 2007). Hier müssten die enthaltenen Dichtefunktionen noch über den Detektionsbereich der jeweiligen Beobachtung integriert werden, um dimensionslose Wahrscheinlichkeiten zu erhalten.

Des Weiteren enthält der ursprüngliche MHT-Ansatz keine explizite Möglichkeit ermittelte Trajektorien zu bewerten (Reid, 1979) wie es für ein effektives *Track Management* wünschenswert wäre. Diese Einschränkung entfällt jedoch, sobald man auf den Hypothesenwert übergeht. Hier kann auf Basis der Beiträge der einzelnen Zuordnungsoptionen (s. Gleichungen 4.12–4.14) für jede Trajektorie ein individueller Wert (*track score*) bestimmt werden. Dieser ermöglicht es, die Trajektorie im Rahmen eines sequentiellen Quotiententests (Wald, 1945) als *bestätigt*, *vorläufig* oder *falsch* zu klassifizieren (Sittler, 1964). Hierfür wird der Trajektorienwert mit den folgenden zwei Schwellwerten verglichen:

$$T_{\text{bestätigt}} = \ln \frac{1 - \beta}{\alpha}, \quad T_{\text{falsch}} = \ln \frac{\beta}{1 - \alpha} \quad (4.15)$$

Die Wahrscheinlichkeit α , eine falsche Trajektorie zu bestätigen, und die Wahrscheinlichkeit β , eine korrekte Trajektorie zu löschen, sind Parameter, welche auf Grundlage von Anwendungsanforderungen manuell gesetzt werden müssen (z. B. $\alpha = 0,2, \beta = 0,1$; s. a. Blackman und Popoli (1999)).

Der Wert einer Trajektorie verringert sich, sobald keine neue Detektion zugeordnet werden kann oder diese nur schlecht zu den bisherigen Beobachtungen passt. Dieser Umstand kann ausgenutzt werden, um zu entscheiden, ob ein Objekt weiter verfolgt wird oder nicht (Blackman, 1986). Hierfür wird zuerst die Wahrscheinlichkeit P_{FD} definiert, eine als korrekt klassifizierte Trajektorie fälschlicherweise nicht weiter zu verfolgen. Dieser Wert ist meist deutlich geringer als β und kann z. B. auf Basis einer maximal zugelassenen Anzahl n von Zeitpunkten ohne Detektion definiert werden: $P_{FD} = (1 - P_D)^n$. Ein Objekt wird nun nicht mehr weiter verfolgt, wenn der Wert seiner Trajektorie L_i^t verglichen mit dem Maximum zum Zeitpunkt t^* zu stark gesunken ist:

$$\Delta L_i^t = L_i^t - L_i^{t^*} < \ln(P_{FD}), \quad t^* = \operatorname{argmax}_t L_i(t) \quad (4.16)$$

4.2.3. Bestimmung der Wahrscheinlichkeitsdichten

Zur Bewertung von Hypothesen und Trajektorien müssen die Beiträge der enthaltenen Zuordnungsoptionen berechnet werden. Für die Optionen ΔL_{NT} und ΔL_{DT} , gegeben in Glei-

4. Objektverfolgung



Abbildung 4.1.: Wahrscheinlichkeitsdichten für neu erkannte Objekte (NT) und Falschalarme (FA).
Dunkle Bereiche kennzeichnen Orte mit hoher Auftrittswahrscheinlichkeit.

chungen 4.12 und 4.13, werden die Wahrscheinlichkeitsdichten für neue Objekte f_{NT} , wiedererkannte Objekte f_{DT} und Falschalarme f_{FA} benötigt. Diese beschreiben die Auftrittswahrscheinlichkeit von Detektionen der jeweiligen Sorte im Raum der messbaren Zustände. Beschränken sich die Beobachtungen auf Objektpositionen, fällt dieser Raum mit dem Beobachtungsbereich bzw. dem Bildausschnitt zusammen.

Die Wahrscheinlichkeitsdichte der prädierten Zustände eines bekannten Objektes wird aufgrund der Kalman-Filterung (s. Abs. 4.1.2) durch eine mehrdimensionale Normalverteilung beschrieben. Der Wert von $f_{DT}(z_j^t)$ ist somit gleichbedeutend der Mahalanobis-Distanz zwischen den prädierten Zuständen und der jeweiligen Beobachtung (Reid, 1979).

Die Dichten von neuen Objekten und Falschalarmen werden meist mit einer Gleichverteilung beschrieben. Dies ist allerdings eine starke Vereinfachung der Wirklichkeit und kann zu ungenügenden Ergebnissen führen. In dieser Arbeit wird daher die Dichteverteilung ortsabhängig modelliert (s. Abb. 4.1) wie es u. a. in (Sittler, 1964) und (Blackman und Popoli, 1999) angeregt wird. Hierfür wurde eine Methode entwickelt, welche die Verteilung während des Trackings automatisch lernt und sich variierenden Gegebenheiten anpassen kann. Die nachfolgenden Erklärungen erfolgen für neu erkannte Objekte, sie gelten jedoch analog für Falschalarme.

Die Dichtefunktion f_{NT} beschreibt die Verteilung der Auftrittswahrscheinlichkeit neuer Objekte insgesamt. Sie lässt sich aufteilen in die zu erwartende Anzahl neuer Objekte \hat{N}_{NT} zum Zeitpunkt t und die normalisierte Dichteverteilung eines einzelnen neuen Objektes p_{NT} :

$$f_{NT}(z^t) = \hat{N}_{NT}(t) \cdot p_{NT}(z^t) \quad (4.17)$$

Die Zahl neuer Objekte schwankt üblicherweise mit der Zeit. Man erhält eine gute Schätzung für das aktuelle Bild, wenn man über die Anzahl neuer Objekte in vorangegangenen Aufnahmen mittelt. Der hierfür betrachtete Zeitraum kann anwendungsabhängig festgelegt werden. Da im MHT-Verfahren mehrere Hypothesen gleichzeitig verfolgt werden, gibt es in der Regel unterschiedliche Aussagen über die Zahl neu erkannter Objekte für einen bestimmten Zeitpunkt. Dieses Problem könnte behoben werden, indem zuerst in den entsprechenden Zuordnungsereignissen aller Hypothesen die NT-Optionen gezählt und diese Werte anschließend mit der jeweiligen Hypothesenwahrscheinlichkeit gewichtet aufsummiert werden würden. Dieser Ansatz würde jedoch die Zahl neuer Objekte über- und die Zahl an Falschalarmen

unterschätzen. Die Ursache hierfür liegt in der Tatsache, dass viele Detektionen zuerst als neues Objekt behandelt, später jedoch als falsche Trajektorien aussortiert werden (s. Abs. 4.2.2). Aus diesem Grund ist es notwendig, nach jedem Zeitschritt die Zahl neuer Objekte für alle vorangegangenen Zeitpunkte auf Basis der aktuellen Hypothesen neu zu bestimmen.

Hierfür sollten zuerst die Trajektorien aller Hypothesen akkumuliert und ihre jeweilige Wahrscheinlichkeit berechnet werden. Diese ergibt sich aus der Summe der Wahrscheinlichkeiten aller Hypothesen, in denen die Trajektorie enthalten ist. Anschließend wird jede einzelne Trajektorie überprüft, ob sie aus einem einzelnen Falschalarm besteht oder als falsch klassifiziert worden ist. Ist dies nicht der Fall, wird die Anzahl neuer Objekte zum Startzeitpunkt der Trajektorie um deren Wahrscheinlichkeit erhöht. Für dieses Verfahren wird die Wahrscheinlichkeit einer einzelnen Hypothese $P(\Theta_i^t)$ benötigt. Diese ergibt sich direkt aus dem Hypothesenwert $L_{\Theta_i^t}$, wenn dieser wie folgt mit der Summe der Werte aller N_{Θ^t} betrachteten Hypothesen normalisiert wird:

$$P(\Theta_i^t) = \frac{\exp(L_{\Theta_i^t})}{\sum_j^{N_{\Theta^t}} \exp(L_{\Theta_j^t})} \quad (4.18)$$

Die Gleichung unterscheidet sich leicht von der in (Demos u. a., 1990) präsentierten. Das Berücksichtigen einer implizit vorhandenen Falschalarm-Hypothese ist in dem hier vorgestellten MHT-Ansatz nicht notwendig und kann sogar zu schlechteren Ergebnissen führen, wenn nur sehr wenige Hypothesen verfolgt werden. Der Wert einer Hypothese steigt in den meisten Fällen mit zunehmender Länge des betrachteten Zeitraumes. Wird er dann in die Exponentialfunktion in Gleichung 4.18 eingesetzt, kann es zu numerischen Problemen kommen. Diese lassen sich vermeiden, wenn die Werte aller Hypothesen zuvor um ihren Mittelwert reduziert werden.

Die Anzahl neuer Objekte im ersten Bild kann vor Beginn des Trackings abgeschätzt werden, indem für jede Detektion die Objektwahrscheinlichkeit auf Basis des Detektionswertes berechnet wird. Da zum ersten Zeitpunkt neue Objekte im gesamten Beobachtungsbereich und nicht nur am Bildrand auftreten können, ist dieser Wert meist relativ hoch und nicht repräsentativ für die restliche Sequenz. Er sollte daher anschließend nicht weiter berücksichtigt werden.

Für Gleichung 4.17 wird neben der zu erwartenden Anzahl neuer Objekte auch die Dichtefunktion p_{NT} benötigt. Diese beschreibt die Verteilung aller messbaren Zustände eines neu erkannten Objektes im Zustandsraum. Da in dieser Arbeit nur die Position von Objekten gemessen wird, beschränkt sich der nachfolgende Absatz auf die Bestimmung der räumlichen Dichtefunktion. Die vorgestellte Methode lässt sich jedoch leicht auf weitere Zustände übertragen.

Die räumliche Wahrscheinlichkeitsdichte neuer Objekte wird hauptsächlich durch das Aussehen der betrachteten Szene beeinflusst und wird daher als statisch angenommen. Um sie zu ermitteln, wird statt eines gleitenden zeitlichen Fensters ein räumlicher Akkumulator genutzt (vgl. Perko und Leonardis (2010)). Mit der gleichen Methode wie zur Bestimmung der Anzahl neuer Objekte, werden auch hier die Wahrscheinlichkeiten aller Trajektorien ermittelt und jeweils an der Stelle der ersten Detektion in den Akkumulator eingetragen.

Die kontinuierliche, räumliche Dichte folgt dann, nachdem der Akkumulator mit einem Gauß-Kernel fester Bandbreite gefiltert und anschließend normalisiert worden ist (s. Abb. 4.1). Der Akkumulator kann eine geringere Auflösung als das Originalbild besitzen, um das Verfahren zu beschleunigen. Liegt kein Vorwissen über die räumliche Dichte vor, sollte er mit einem kleinen Wert ϵ initialisiert werden, was für das erste Bild in einer Gleichverteilung resultiert.

4. Objektverfolgung

4.2.4. Integration des Detektionswertes

Die vorgestellte Bewertungsfunktion bestimmte die Zuordnungswahrscheinlichkeit bisher allein auf Basis der messbaren Attribute der Detektionen. In dieser Arbeit ist dies der Ort einer möglichen Person. Je nachdem wo eine Detektion erfolgt, ob in der Nähe einer zuvor erkannten Person oder am Rand des Bildes, unterscheiden sich die Wahrscheinlichkeiten der unterschiedlichen Zuordnungsmöglichkeiten. Die Qualität der Detektion blieb in der Bewertungsfunktion bisher noch unberücksichtigt, obwohl sie für das Tracking eine wichtige Information darstellt. Üblicherweise werden in einem Tracking-by-Detection-Verfahren nur diejenigen Detektionen genutzt, welche einen bestimmten Schwellwert überschreiten. Dieses Vorgehen reduziert die Komplexität des anschließend zu lösenden Korrespondenzproblems, es bringt jedoch auch einige Nachteile mit sich.

Durch die strikte, unumkehrbare Einteilung von Detektionen in Falschalarme und mögliche Objektpositionen, werden die beiden Prozesse Erkennen und Verfolgen deutlich voneinander getrennt. Informationen über Detektionen, die unterhalb des Schwellwertes liegen, stehen im Tracking nicht mehr zur Verfügung. Die Schwierigkeit besteht nun darin, den Schwellwert so festzulegen, dass möglichst alle Objektbeobachtungen erhalten bleiben und gleichzeitig sämtliche Falschalarme aussortieren werden. In Anwendungen, in denen sich korrekte Detektionen deutlich von falschen unterscheiden, ist dies leicht möglich. Völlig anders stellt sich die Situation dar, wenn, wie in dieser Arbeit, die Trennbarkeit von Objekt und Hintergrund nicht eindeutig möglich ist. Hier hängen die Detektions- und Tracking-Ergebnisse stark von der Wahl des Schwellwertes ab.

Ein weiterer Nachteil der Standardmethode ist der, dass alle Detektionen oberhalb des Schwellwertes gleichgewichtet in das Tracking eingehen. Der Abstand zum Schwellwert ist jedoch ein wichtiges Indiz für die Qualität und Zuverlässigkeit der Detektion und sollte daher auch im Tracking genutzt werden. Besonders Anwendungen mit geringem Signal-Rausch-Verhältnis können davon profitieren. Hier lässt sich der Detektionsschwellwert so weit absenken, dass auch schwach erkennbare Objekte verfolgt werden können. Falsche Detektionen, die nun in größerer Zahl auftreten, lassen sich jedoch leicht identifizieren und schnell aussortieren.

Um die angesprochenen Nachteile der Standardmethode zu eliminieren, wird in dieser Arbeit der Detektionswert in die stochastische Bewertungsfunktion integriert. Die Trennung zwischen Detektion und Tracking wird damit aufgehoben und die zwei mächtigen Ansätze der aussehensbasierten Objekterkennung mit implizitem Modell und des Multi-Hypothesen-Trackings direkt miteinander verbunden. Die Grundlage hierfür bilden die stochastische Beschreibung des Detektionswertes in Abschnitt 3.6 und ein Ansatz, welcher in Radaranwendungen genutzt wird, um das Signal-Rausch-Verhältnis in das MHT-Verfahren zu integrieren (Blackman und Popoli, 1999).

Unter der Annahme, dass die Zustände einer Detektion z_j unabhängig von ihrem Wert s_j gemessen werden, lässt sich Gleichung 3.4 für die a posteriori Wahrscheinlichkeit eines Detektionsereignisses H_i folgendermaßen umschreiben:

$$P(H_i|s_j, z_j, D) \propto P(s_j|D, H_i) \cdot P(z_j|D, H_i) \cdot P(D|H_i) \cdot P(H_i), \quad i \in \{0, 1\} \quad (4.19)$$

Berücksichtigt man nun noch, dass nur Beobachtungen im Tracking verwendet werden, deren Detektionswert über einem Schwellwert T liegen, ergibt sich:

$$P(H_i|s_j, z_j, T, D) \propto P(s_j|T, D, H_i) \cdot P(z_j|T, D, H_i) \cdot P(S > T|D, H_i) \cdot P(D|H_i) \cdot P(H_i) \quad (4.20)$$

Die Verbindung zur den entsprechenden Komponenten der MHT-Grundgleichung 4.8 sieht folgendermaßen aus:

$$f_{NT}(z_j) = P(z_j | T, D, H_1) \cdot P(S > T | D, H_1) \cdot P(D | H_1) \cdot P(H_1) \quad (4.21)$$

$$f_{FA}(z_j) = P(z_j | T, D, H_0) \cdot P(S > T | D, H_0) \cdot P(D | H_0) \cdot P(H_0) \quad (4.22)$$

$$f_{DT}(z_j) \cdot P_D = P(z_j | T, D, H_1) \cdot P(S > T | D, H_1) \cdot P(D | H_1) \quad (4.23)$$

Integriert man noch den Detektionswert s_j wie in Gleichung 4.20 beschrieben, ergibt sich bspw. für die Wahrscheinlichkeit eines neuen Objektes folgendes:

$$\begin{aligned} f_{NT}(s_j, z_j) &= P(s_j | T, D, H_1) \cdot f_{NT}(z_j) \\ &= \frac{P(s_j | D, H_1)}{P(S > T | D, H_1)} \cdot f_{NT}(z_j) = \frac{P(s_j | D, H_1)}{1 - P(S \leq T | D, H_1)} \cdot f_{NT}(z_j) \end{aligned} \quad (4.24)$$

Die obige Gleichung zeigt, wie die stochastische Modellierung des Detektionswertes in Abschnitt 3.6 nun dabei hilft, die Ergebnisse der Objekterkennung in die Bewertungsfunktion des Tracking-Verfahrens zu integrieren. Überträgt man dieses Ergebnis abschließend auf die einzelnen Zuordnungsoptionen, ergibt sich folgendes Bild:

$$\Delta L_{FA} = 0 \quad (4.25)$$

$$\Delta L_{NT} = \ln \frac{P(s_j^t | T, D, H_1) \cdot f_{NT}(z_j^t)}{P(s_j^t | T, D, H_0) \cdot f_{FA}(z_j^t)} \quad (4.26)$$

$$\Delta L_{DT} = \ln \frac{P(s_j^t | T, D, H_1) \cdot f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P(S > T | D, H_1) P_D}{P(s_j^t | T, D, H_0) \cdot f_{FA}(z_j^t)} \quad (4.27)$$

$$\Delta L_{LT} = \ln[1 - P(S > T | D, H_1) \cdot P_D] \quad (4.28)$$

Die Integration des Detektionswertes bewirkt für ΔL_{NT} und ΔL_{DT} jeweils einen zusätzlichen Term im Nenner und Zähler. Der Einfluss des Schwellwertes auf die Detektionswahrscheinlichkeit wird mit $P(S > T | D, H_1)$ explizit ausgewiesen. Dadurch ist diese nun nur noch von der Qualität des Detektionsverfahrens und der Erkennbarkeit der Objekte in der jeweiligen Situation abhängig. Zudem muss der Schwellwert nun nicht mehr manuell gesetzt werden, sondern kann anhand von Gleichung 3.5 und der Vorgabe einer entsprechenden Wahrscheinlichkeit für $P(S > T | D, H_1) = 1 - P(S \leq s | D, H_1)$ direkt berechnet werden.

4.3. Multi-Hypothesen-Tracking

Bisher wurde die im MHT-Verfahren genutzte Bewertungsfunktion vorgestellt und erweitert. Mit ihr ist es möglich, alternative Hypothesen zu evaluieren und entsprechend ihren Wahrscheinlichkeiten zu ordnen. In diesem Abschnitt soll es nun um die zweite Hauptkomponente des MHT-Ansatzes gehen, dem Verfahren, mit welchem mehrere globale Hypothesen gleichzeitig verfolgt werden können. Im Zentrum steht dabei die Frage, wie sich neue Hypothesen möglichst effektiv aus den bisherigen erzeugt lassen sobald ein neues Bild eintrifft und sich der zu berücksichtigende Zeitraum vergrößert (vgl. Gl. 4.7). Hierfür werden Verfahren benötigt, welche die Komplexität des Tracking-Problems so weit reduzieren, dass eine Auswertung in Echtzeit möglich ist, aber gleichzeitig das Ergebnis ausreichend nah an der global optimalen Lösung liegt.

4.3.1. Datenstruktur

Bevor die einzelnen Schritte des MHT-Verfahrens detailliert erklärt werden, wird die zugrunde liegende Datenstruktur beschrieben. Wie für Methoden der Objektverfolgung üblich, ba-

4. Objektverfolgung

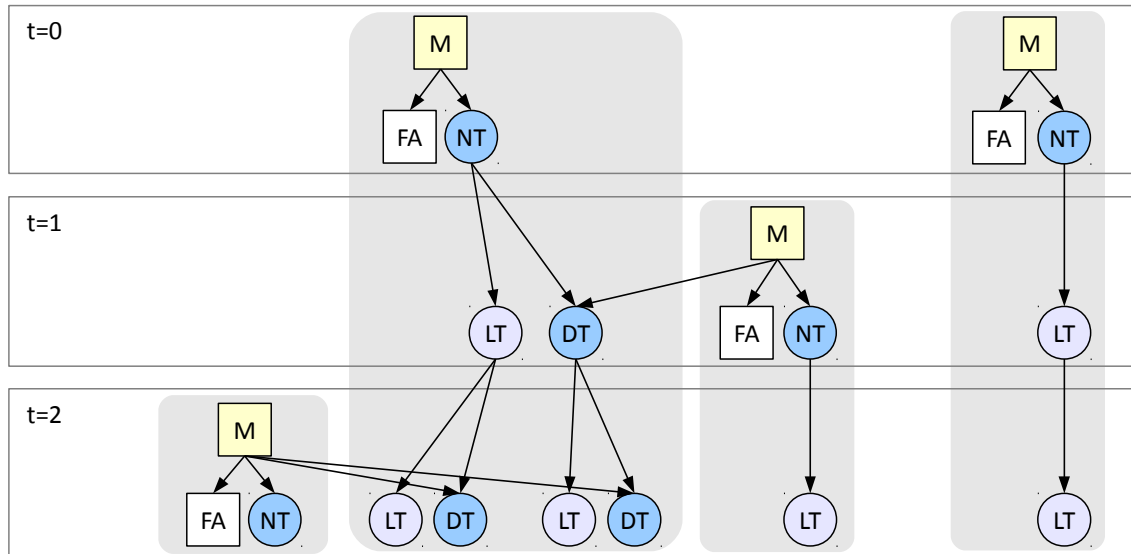


Abbildung 4.2.: Beispiel eines möglichen Tracking-Graphs mit vier Detektionen [M] in den ersten drei Bildern. Diese bilden Wurzelknoten während Hypothesen immer auf Blattknoten zeigen. Die Kanten beschreiben Abhängigkeiten und Interaktionen, was für diesen Graph bedeutet, dass er aus zwei unabhängigen Clustern besteht.

siert sie auf einem gerichteten Graphen und wird nachfolgend als *Tracking-Graph*¹ bezeichnet. Dieser speichert alle Hypothesen, Beobachtungen und Trajektorien sowie deren wechselseitige Abhängigkeiten in einer einzigen Struktur (s. Abb. 4.2). Seine Knoten repräsentieren Beobachtungen, Trajektorien und Zuordnungsoptionen. Die Kanten enthalten den Beitrag der Optionen zum Hypothesenwert (s. Gl. 4.25 bis 4.28) und verbinden die zeitlich benachbarten Zuordnungsereignisse jeder Hypothese miteinander. Zusätzlich verkörpern sie Abhängigkeiten zwischen Hypothesen, was eine wichtige Information ist, um das Gesamtproblem später in kleinere, einfacherer Teilprobleme gliedern zu können.

Diese werden als Cluster bezeichnet und umfassen jeweils eine Liste interagierender Hypothesen. Eine einzelne Hypothese besteht wiederum aus zwei Listen mit Zeigern zu Blattknoten des Tracking-Graphs. Die erste ist für aktive Trajektorien, welche auch zum nächsten Zeitpunkt berücksichtigt werden müssen, die zweite Liste enthält inaktive Trajektorien von verschwundenen Objekten oder Falschalarmen. Geht man von diesen Blattknoten aus in die Vergangenheit, also entgegen der Kantenrichtung, zeigen sich sämtliche Zuordnungsereignisse der jeweiligen Hypothese und man erhält eine konsistente Erklärung für alle bisherigen Beobachtungen innerhalb des jeweiligen Clusters. Des Weiteren ist es vorteilhaft, während des Trackings eine zusätzliche Liste mit Beobachtungen zu führen, die von der jeweiligen Hypothese im aktuellen Zuordnungsereignis berücksichtigt werden müssen.

4.3.2. Ablauf

In diesem Abschnitt werden alle Schritte des MHT-Verfahrens kurz erklärt, welche notwendig sind, um die ursprüngliche, hypothesen-orientierte Variante von Reid (1979) in Echtzeit ablaufen zu lassen (s. Abb. 4.3). Die besonders wichtigen Methoden zur Hypothesengenerierung und zum Clustern werden in nachfolgenden Abschnitten gesondert behandelt.

Detektionen empfangen und Objekte präzisieren. Da MHT ein sequentieller Ansatz ist, müssen die nachfolgenden Verfahrensschritte für jedes Bild bzw. jeden Zeitpunkt wiederholt

¹Der Tracking-Graph wurde auf Basis der freien Graphenbibliothek *LEMON* implementiert.

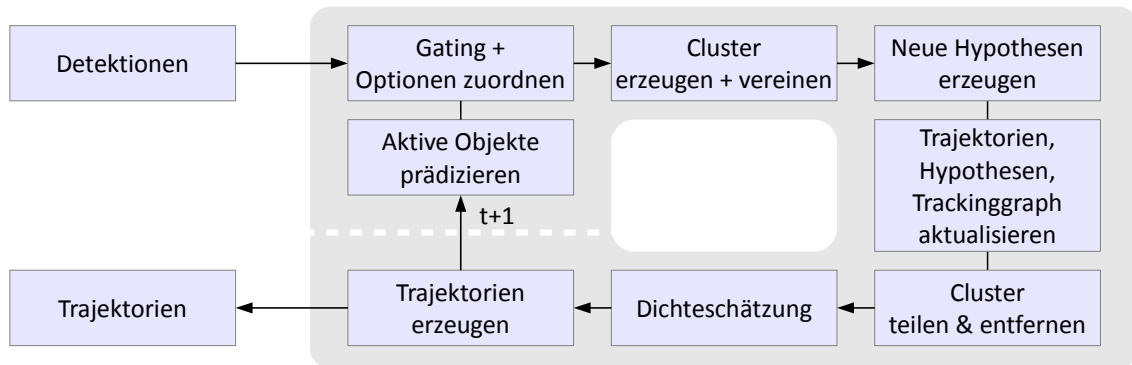


Abbildung 4.3.: Ablaufdiagramm des MHT-Verfahrens, so wie es in dieser Arbeit dargestellt wird.

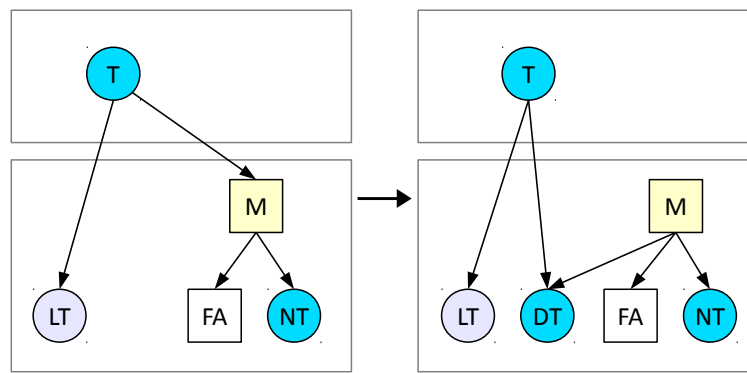


Abbildung 4.4.: Die linke Seite zeigt den Tracking-Graph, nachdem durch Einfügen von neuen Knoten und Kanten alle möglichen Zuordnungsoptionen erzeugt wurden. Die rechte Seite stellt dar, wie der Graph verändert werden muss, sollte die DT-Option in einem Zuordnungsereignis enthalten sein. Der T-Knoten repräsentiert eine aktive Trajektorie (NT, LT oder DT).

werden. Zuerst wird für jede Detektion im aktuellen Bild ein neuer Knoten in den Tracking-Graph eingefügt und die Zustände aller aktiven Objekte mittels Kalman-Filter präzidiert.

Gating und Clustern. Anschließend werden diese Objekte mit den aktuellen Beobachtungen verbunden (s. Abb. 4.4). Die dabei erzeugten Kanten erhalten den Wert der DT-Option ΔL_{DT} . Um die Komplexität des später zu lösenden Zuordnungsproblem zu reduzieren, werden sehr unwahrscheinliche Paarungen von vornherein ausgeschlossen. Dieser als *Gating* bezeichnete Vorgang kann bereits sehr viel Rechenzeit beanspruchen, da prinzipiell jedes Objekt mit jeder Beobachtung verglichen werden muss. Daher wird die effektive Gating-Strategie von Collins und Uhlmann (1992) genutzt. Hier werden die Positionen aller Beobachtungen in zweidimensionale Bäume eingetragen, was die Suche nach Detektionen in der Nähe eines präzidierten Objektes enorm beschleunigt. Anschließend folgt eine Kaskade immer aufwendiger Zustandsvergleiche, um viele Kandidaten bereits zu Beginn aussortieren zu können. In einem abschließenden Test wird sichergestellt, dass die Zuordnung den Wert einer Trajektorie nicht unter den Schwellwert T_{falsch} absenkt (vgl. Abs. 4.2.2). Nachdem alle zulässigen Paarungen erzeugt wurden, muss gewährleistet werden, dass alle Hypothesen eines Clusters die gleichen Detektionen in ihren Listen führen, auch wenn diese nicht direkt mit den eigenen Objekten verknüpft wurden. Danach kann die Liste der Cluster aktualisiert werden, indem interagierende Cluster vereint und neue erzeugt werden (s. Abs. 4.3.4).

4. Objektverfolgung

Hypothesen erzeugen. Da die DT-Option bereits während des Gatings erstellt wurde, müssen die aktuellen Zuordnungsprobleme durch Einfügen der noch fehlenden Optionen in den Graph vervollständigt werden. Daher werden nun alle aktiven Objekte mit der LT-Option verbunden und alle Detektionen mit der NT- und FA-Option (s. Abb. 4.4). Die dabei entstehenden Kanten erhalten den Wert der jeweiligen Zuordnungsoption. Nun können unabhängig für jeden Cluster neue Hypothesen erzeugt werden. Dies geschieht auf Basis der bisherigen Hypothesen und der Lösung des aktuellen Zuordnungsproblems (s. Abs. 4.3.3). Alle folgenden Prozessierungsschritte dienen der Nachbereitung.

Trajektorien und Hypothesen aktualisieren. Zuerst werden alle Optionen betrachtet, die in den aktuellen Zuordnungsereignissen der neuen Hypothesen enthalten sind. Für jede DT-Option werden die prädizierten Objektzustände mit der zugeordneten Beobachtungen korrigiert. Dann wird ein neuer DT-Knoten in den Graphen eingefügt und die bisherige Kante, wie in Abbildung 4.4 dargestellt, geändert. Objekte mit der LT-Option werden weiterverfolgt oder deaktiviert, je nach Status und Wert ihrer Trajektorie (s. Abs. 4.2.2). Für jede NT-Option wird ein neuer Kalman-Filter mit der zugehörigen Detektion initialisiert. Nachdem alle gewählten Optionen abgearbeitet worden sind, können die Listen der aktiven und inaktiven Trajektorien aller Hypothesen entsprechend aktualisiert werden.

Tracking-Graph bereinigen und clustern. Der Tracking-Graph kann nun von allen ungenutzten Knoten und Kanten bereinigt werden. So wird sichergestellt, dass er ausschließlich die aktuellen Hypothesen und ihre jeweiligen Abhängigkeiten repräsentiert. Zuerst werden alle Kanten gelöscht, die während des Gatings erstellt wurden. Danach werden die Listen der aktiven und inaktiven Trajektorien aller Hypothesen analysiert. Angefangen von den Blattknoten werden alle vorhergehenden Knoten im Tracking-Graph markiert bis die Wurzelknoten erreicht sind. Wird hierbei ein Knoten erreicht, der bereits markiert worden ist, kann Zeit gespart und mit der nächsten Trajektorie fortgefahren werden. Knoten, welche am Ende dieser Prozedur nicht markiert worden sind, werden von keiner Hypothese genutzt und können daher gelöscht werden. Im Anschluss an diese Bereinigung wird die Struktur aller Cluster analysiert, um diese bei Bedarf zu teilen oder zu deaktivieren (s. Abs. 4.3.4).

Dichteschätzung. Einer der letzten Schritte beinhaltet die Schätzung der Dichteverteilung von neuen Objekten und Fehlalarmen entsprechend Abschnitt 4.2.3. Hierfür werden zuerst die inaktiven Cluster prozessiert. Da diese sich auch in Zukunft nicht mehr ändern, werden deren Ergebnisse in permanenten Listen und Akkumulatoren gespeichert. Um die aktuell gültige Dichteverteilung zu erhalten, werden davon temporäre Kopien erstellt und die Ergebnisse der aktiven Cluster ergänzt. Dieses Vorgehen spart Zeit, da nur ein Teil aller Cluster in jedem Zeitschritt neu analysiert werden muss.

Bildung von Trajektorien. Abschließend ergeben sich die Tracking-Ergebnisse aus der Hypothese mit dem größten Wert innerhalb eines inaktiven Clusters. Alle restlichen Hypothesen sind weniger wahrscheinlich und bleiben unberücksichtigt. Jedes bestätigte Objekt in der wahrscheinlichsten Hypothese erzeugt eine eigene Trajektorie. Diese wird zurückgeschnitten bis zu dem Zeitpunkt, an welchem der Trajektorienwert das letzte Mal angestiegen ist und eine Aktualisierung mit einer zuverlässigen Messung erfolgte. Dieses Vorgehen zum Erzeugen von Trajektorien ist besonders geeignet, wenn Bildsequenzen mit einem kurzen Zeitversatz von z. B. zehn Aufnahmen analysiert werden können. Für besonders zeitkritische Anwendungen, in welchen für jedes Bild sofort ein Ergebnis benötigt wird, müssen andere Methoden eingesetzt werden (vgl. Blackman und Popoli (1999)).

4.3.3. Hypothesengenerierung

Das Erzeugen neuer Hypothesen gehört zu den wichtigsten Schritten im MHT-Verfahren. Jedes Mal wenn ein neues Bild eintrifft, vergrößert sich der Zeitraum, für welchen die global wahrscheinlichste Hypothese gefunden werden muss. Wie man an der MHT-Grundgleichung 4.7 gut erkennen kann, muss hierfür nicht die gesamte Sequenz ausgewertet werden. Die neuen Hypothesen lassen sich vielmehr in sequentieller Weise auf Basis der bisherigen ermitteln (Reid, 1979). Hierbei ergibt sich jedoch das Problem, dass deren Anzahl exponentiell mit der Zeit zunimmt. Aus einer einzigen Hypothese lassen sich für jeden Zeitschritt so viele neue Hypothesen erzeugen, wie das aktuelle Zuordnungsproblem Lösungen besitzt. Um den MHT-Ansatz praktikabel zu machen, ist es daher notwendig, die Anzahl der Hypothesen zu beschränken.

Zur Bewältigung dieser Aufgabe wurden im Laufe der Zeit zahlreiche Vorschläge veröffentlicht, welche in *screening methods* und *pruning methods* eingeteilt werden können (Kurien, 1990). Zur ersten Gruppe gehören Methoden wie z. B. Gating, welche im Vorfeld die mögliche Anzahl neuer Hypothesen einschränken. Verfahren der zweiten Kategorie werden dagegen eingesetzt, um nach der Erzeugung neuer Hypothesen deren Anzahl zu reduzieren. So kann z. B. pro Objekt nur eine maximale Zahl unterschiedlicher Varianten zugelassen oder von jeder Hypothese eine Mindestwahrscheinlichkeit gefordert werden. All diese Methoden lösen das Problem der steigenden Komplexität jedoch nicht grundsätzlich. Deutlich vielversprechender ist das in (Cox und Hingorani, 1996) veröffentlichte Verfahren, welches auf einer optimierten Variante von Murty's Algorithmus (Murty, 1968) beruht. Es lässt sich auf den hypothesen-orientierten MHT-Ansatz anwenden und bestimmt für jeden Zeitpunkt die n wahrscheinlichsten Hypothesen. Die Komplexität, Qualität und Laufzeit des MHT-Ansatzes lassen sich so mit einem einzigen Parameter präzise steuern. Es gibt kein anderes Verfahren, welches dies auf eine ähnlich effektive und elegante Weise vollbringt.

Die Methode funktioniert im Detail wie folgt (Cox und Hingorani, 1996; Blackman und Popoli, 1999): Sobald ein neues Bild eintrifft, wird für jede bisherige Hypothese die optimale Lösung des aktuellen Zuordnungsproblems bestimmt. Die sich so ergebenden neuen Hypothesen werden in eine temporäre Liste möglicher Lösungen aufgenommen und entsprechend ihres Wertes geordnet. Dann wird die beste Hypothese von der Spitze dieser Liste entfernt und als erste Lösung übernommen. Anschließend wird mit Hilfe von Murty's Algorithmus die zweitbeste Lösung des Zuordnungsproblems dieser Hypothese bestimmt und so eine weitere mögliche Hypothese erzeugt. Diese wird nun ebenfalls entsprechend ihres Wertes in die temporäre Liste eingeordnet. Diese Prozedur wird so oft wiederholt, bis die Liste der endgültigen Lösungen den gewünschten Umfang erreicht hat oder die temporäre Liste keine weiteren Hypothesen mehr enthält. In (Miller u. a., 1997) werden drei Optimierungen zur Beschleunigung dieses Verfahrens vorgeschlagen. Sie zielen darauf ab, sowohl die Anzahl der Zuordnungsprobleme zu reduzieren, für welche die optimale Lösung gefunden werden muss, als auch die Zeit, welche jeweils dafür benötigt wird.

Bei der Generierung neuer Hypothesen muss für das jeweils betrachtete Zuordnungsproblem immer die optimale Lösung gefunden werden. Da in dieser Arbeit nur Einzelpersonen erkannt werden und es aufgrund der Draufsicht zu keinen gegenseitigen Verdeckungen kommt, sind nur 1:1-Zuordnungen erlaubt. Sobald sich die Aufgabe als bipartiter Graph bzw. lineares Zuordnungsproblem darstellen lässt, kann eine Lösung leicht mit Standardverfahren wie der Ungarischen Methode (Kuhn, 1955) berechnet werden (vgl. a. Pentico (2007)). Hierfür ist es notwendig das Problem so zu erweitern, dass es eine quadratische Form besitzt (Miller u. a., 1997). Dies kann erreicht werden, indem temporäre Knoten und Kanten, wie in Abbildung 4.5 gezeigt, hinzugefügt werden.

4. Objektverfolgung

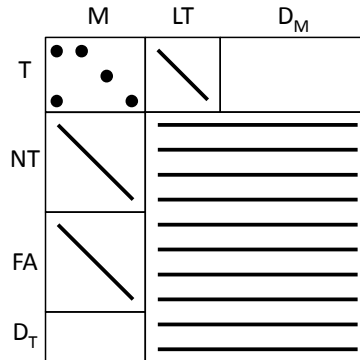


Abbildung 4.5.: Schematische Matrixdarstellung eines Zuordnungsproblems im MHT-Verfahren. Temporäre Knoten (D_M , D_T) und Kanten (rechte untere Submatrix) wurden hinzugefügt, um eine quadratische Form für den Zuordnungsalgorithmus zu gewährleisten. In echten Problemen werden entweder D_M - oder D_T -Knoten benötigt.

4.3.4. Clustern

Im letzten Abschnitt wurde eine effiziente Methode präsentiert, neue Hypothesen zu erzeugen. Obwohl diese einen Durchbruch für die schnelle Ausführung des hypothesen-orientierten MHT-Ansatzes darstellt, wird dessen volles Potential nur dann ausgeschöpft, wenn auch ein effizientes Cluster-Verfahren eingesetzt wird. Ein einzelnes Cluster ist definiert als eine Gruppe interagierender Hypothesen, welche keine Abhängigkeiten mit Hypothesen anderer Cluster besitzen. Diese einfache Regel erlaubt eine Partitionierung des Tracking-Problems sowie eine parallele Prozessierung. Der entscheidende Vorteil ist jedoch, dass die Anzahl der berücksichtigten Hypothesen stark erhöht werden kann. Bestimmt man z. B. die 10 besten Hypothesen in 5 Clustern, so entspricht dies ca. $5^{10} \approx 10$ Mio. Hypothesen für das Gesamtproblem (vgl. Blackman (1986); Werthmann (1992)). Die Möglichkeit, mit geringem zusätzlichem Aufwand deutlich mehr Hypothesen gleichzeitig zu verfolgen, erlaubt es, der global optimalen Lösung deutlich näher zu kommen.

Cluster-Verfahren für den hypothesen-orientierten MHT-Ansatz wurden bereits von (Reid, 1979) und anderen (Roy u. a., 1997; Antunes u. a., 2011) vorgestellt. Diese beruhen jedoch immer auf einer Analyse des sog. *hypotheses trees*, welcher in der hier vorgestellten MHT-Umsetzung jedoch nicht mehr benötigt wird. Nachfolgend werden die vier Grundaufgaben des Clusters vorgestellt und erläutert, wie sie durch die Nutzung des Tracking-Graphs (vgl. Abb. 4.2) stark vereinfacht werden.

Cluster erzeugen und vereinigen

Wenn neue Detektionen eintreffen, werden sie während des Gatings mit existierenden Objekten verbunden. Jede Beobachtung, die nicht von einem Objekt beansprucht wird, bildet einen neuen Cluster.

Sobald Hypothesen unterschiedlicher Cluster Anspruch auf die selbe Detektion erheben, müssen deren Cluster zu einem sog. *Supercluster* vereint werden. Solch interagierende Cluster können leicht identifiziert werden, indem nach Zusammenhangskomponenten im Tracking-Graph gesucht wird (s. Abb. 4.2). Anschließend müssen alle Cluster, die Knoten der selben Komponente enthalten, vereint werden. Diese Prozedur kann für mehrere Cluster beschleunigt werden, indem diese zuerst in aufsteigender Reihenfolge entsprechend der Anzahl an enthaltenen Hypothesen geordnet werden. Anschließend werden iterativ benachbarte Cluster zusammengefasst bis nur noch ein Supercluster übrig ist.

Zwei Cluster werden vereint, indem man ihre Hypothesen zusammenfasst. Für zwei Hypothesen bedeutet dies, dass ihre Hypothesenwerte addiert und die Listen der aktiven und inaktiven Trajektorien zusammengefasst werden. Die Anzahl der Hypothesen in einem Supercluster ist das Produkt der Hypothesenanzahl der beiden ursprünglichen Cluster, da jede Hypothese im ersten mit jeder Hypothese im zweiten Cluster kombiniert werden muss. Ein Supercluster würde daher im Allgemeinen deutlich mehr als die vorgegebene Anzahl an Hypothesen enthalten. Werden jedoch die Hypothesen der beteiligten zwei Cluster zuvor entsprechend ihres Wertes geordnet, so kann die Anzahl leicht auf die n besten beschränkt werden, ohne sämtliche Kombinationen durchgehen zu müssen.

Aufgrund dieser Vorgehensweise fallen einige Hypothesen bereits während der Vereinigung von Clustern weg. Im Anschluss muss daher immer überprüft werden, ob jedes aktive Objekt noch von mindestens einer Hypothese referenziert wird. Ist dies nicht der Fall, so kann der zugehörige Blattknoten gelöscht werden. Dies hat wiederum zur Folge, dass Kanten gelöscht werden, welche eventuell während des Gating erstellt wurden. Daher sollten neue Cluster immer erst im Anschluss an die Vereinigung erzeugt werden.

Cluster zerteilen und löschen

Ein Cluster muss aufgeteilt werden, wenn er mehrere unabhängige Teile enthält. Dies kann vorkommen, wenn Hypothesen gelöscht werden und zuvor bestehende Abhängigkeiten bei der Aktualisierung des Tracking-Graphs verschwinden. Die Prozedur zum Aufteilen war bisher äußerst aufwändig, da komplizierte Interaktionen zwischen unterschiedlichen Hypothesen beachtet werden müssen. Durch den Tracking-Graph vereinfacht sich das Verfahren jedoch wieder auf eine simple Suche nach Zusammenhangskomponenten.

Enthält ein Cluster Hypothesen, die auf unterschiedliche Komponenten verweisen, muss er entsprechend aufgeteilt werden. Hierfür werden die enthaltenen Hypothesen einzeln analysiert und deren Knoten entsprechend ihrer Komponentenummer in neuen Hypothesen zusammengefasst. Der Wert der neuen Hypothese ergibt sich aus der Summe der Trajektorienwerte. Das Aufteilen von Hypothesen eines Clusters erzeugt üblicherweise viele identische Teilhypothesen, wovon jedoch nur die erste einem neuen Cluster hinzugefügt werden darf.

Ein Cluster kann deaktiviert werden, wenn keine seiner Hypothesen mehr ein aktives Objekt enthält. Vor dem endgültigen Löschen müssen jedoch noch Objekttrajektorien auf Basis der wahrscheinlichsten Hypothese erzeugt werden.

5. Evaluierung

In diesem Kapitel wird evaluiert, wie gut das bisher beschriebene System die Zielsetzung dieser Arbeit, die automatische Gewinnung möglichst vollständiger Trajektorien von Einzelpersonen aus Luftbildsequenzen, erfüllt. Darüber hinaus werden auch die zahlreichen anwendungsunabhängigen Verbesserungsvorschläge einzeln untersucht und bewertet. Zur besseren Übersicht sind die Experimente je nach Schwerpunkt der Objekterkennung oder der Objektverfolgung zugeordnet. Beide Bereiche werden am Ende des jeweiligen Abschnittes als Ganzes bewertet und diskutiert.

5.1. Grundlagen

Bevor die Experimente zu Detektion und Tracking erläutert werden, folgen in diesem Abschnitt einige grundlegende Anmerkungen zu den verwendeten Bildsequenzen, den Referenzdaten und Evaluationsmaßen.

5.1.1. Bilddaten

Sämtliche in dieser Arbeit genutzte Luftbildsequenzen stammen vom 3K-Kamerasystem des Deutschen Zentrums für Luft- und Raumfahrt (DLR), welches bei Großveranstaltungen, zur Verkehrsüberwachung und im Katastrophenfall eingesetzt wird (Thomas u. a., 2008). Es besteht aus drei Spiegelreflexkameras, welche so ausgerichtet sind, dass ein möglichst großer Bereich gleichzeitig aufgenommen werden kann. Während eine Kamera in Nadirrichtung zeigt, sind die anderen beiden um 35° verschwenkt.

Alle Luftbilder durchlaufen nach der Aufnahme eine Reihe von Vorverarbeitungsschritten, um die nachfolgende Auswertung zu erleichtern. Hierzu zählen u. a. die direkte Georeferenzierung mittels GPS/IMU-Navigationssystem sowie die anschließende Orthorektifizierung auf Basis eines digitalen Geländemodells. Je nach Flughöhe, welche üblicherweise zwischen 1000 m und 1800 m liegt, Sensorgröße und Kameraausrichtung besitzen die Orthophotos eine Bodenauflösung von 12 cm bis 20 cm und decken jeweils einen Bereich von etwa 700 m x 900 m ab. Für die Methodenentwicklung und -evaluierung wurden bestimmte Gebiete kleiner 100 m x 100 m ausgewählt, um die Auswertung zu beschleunigen und auf interessante Bereiche zu fokussieren.

Da das Kamerasystem an einem Flugzeug befestigt ist, kann ein bestimmtes Gebiet während eines Überfluges nur einige wenige Male beobachtet werden. Die genaue Anzahl der Bilder hängt von der Aufnahmefrequenz und der Ausrichtung des Kamerasystems ab und liegt zwischen 4 und 30. Bei einer Frequenz von ca. 2 Hz ergibt dies einen Beobachtungszeitraum von 2 s bis 15 s. Die innere Lagegenauigkeit der Einzelbilder in einer Sequenz beträgt wenige Pixel und wird von vielen Faktoren beeinflusst, wie der Blickrichtung, der Flughöhe oder dem verfügbaren Geländemodell (Kurz u. a., 2007). Werden Bilder von allen drei Kameras in einer Sequenz vereint, kann es beim Übergang von einer Kamera auf die nächste auch zu größeren Abweichungen kommen.

In dieser Arbeit wurden aus den verfügbaren Luftbilddaten eine Vielzahl von Sequenzen erstellt, welche sowohl unterschiedliche Szenarien als auch variierende Schwierigkeiten abdecken. Dies soll eine möglichst allgemeine Bewertung der Leistungsfähigkeit der entwickelten

5. Evaluierung

Verfahren erlauben. Zusätzlich wurden die Bilddaten in Trainings- und Testsequenzen unterteilt. Während erstere u. a. für die Methodenentwicklung, das Klassifikatortraining und die Parameterbestimmung genutzt wurden, kamen letztere ausschließlich bei abschließenden Evaluierungen zum Einsatz.

5.1.2. Referenzdaten

In dieser Arbeit dienen manuell erfasste Positionen und Trajektorien von Personen als Referenzdaten und werden für unterschiedlichste Zwecke eingesetzt. Für das Training des Klassifikators werden bspw. Bildpositionen benötigt, um Beispieldaten erzeugen zu können. Da nicht alle Personen in einem Bild erkennbar sind, z. B. weil sie Teil einer Menschengruppe sind oder hinter einem anderen Objekt verschwinden, müssen diese Positionen sorgfältig ausgewählt werden. Erfasst wird hierbei immer die Bildposition des Mittelpunktes einer Person. Um die Objektklasse mit ihrer Varianz möglichst vollständig zu repräsentieren, wurden insgesamt 2.500 Personen manuell markiert.

Als Referenz für Objekterkennung und -verfolgung werden die Positionen bzw. Trajektorien aller Personen in einem Bild bzw. einer Sequenz benötigt. Diese müssen ebenfalls manuell erfasst werden, was im Rahmen dieser Arbeit für sieben Trainings- und sechs Testsequenzen geschehen ist. Insgesamt sind hierbei etwa 33.000 Positionen von über 2.400 Personen erfasst worden. Die Referenzdaten wurden im CVML-Format (List und Fisher, 2004) gespeichert und zusammen mit den zugehörigen Bildsequenzen öffentlich zugänglich gemacht¹, um auch anderen Wissenschaftlern zur Verfügung zu stehen.

Die Trajektorie jeder Person wurde meist ohne Unterbrechung erfasst. Dies bedeutet, dass ihre jeweilige Position auch dann markiert wurde, wenn die Person nicht mehr erkennbar, ihre Lage jedoch einigermaßen sicher abgeschätzt werden konnte. Dies ist der Grund, warum für das Training des Klassifikators separate Daten erhoben werden mussten. Die vollständige Erfassung der Positionen und Trajektorien, auch von visuell nicht erkennbaren Einzelpersonen, soll dazu dienen, die Grenzen des in dieser Arbeit entwickelten Systems aufzuzeigen.

Da die Referenzdaten manuell in den Sequenzen erfasst wurden, sind sie abhängig von den Bilddaten und stellen keine vollkommen unabhängige *Ground Truth* dar. Eine Evaluierung auf Basis dieser Daten ist somit immer nur ein Vergleich zwischen manueller und automatischer Erfassung. Ein identisches Ergebnis ist nicht möglich, da auch die Referenzdaten Fehler enthalten. Es wird jedoch für die automatischen Verfahren eine ähnlich hohe Qualität angestrebt, wie sie die manuelle Erfassung im Allgemeinen liefert.

5.1.3. Evaluierungsmaße

Wie gut die erzeugten Ergebnisse mit den Referenzdaten übereinstimmen, wird mit Hilfe von Evaluierungsmaßen beschrieben. Eine umfangreiche Übersicht über verschiedene Maße zur Bewertung von Detektions- und Tracking-Verfahren findet man in (Baumann u. a., 2008). Sie alle basieren auf einer deterministischen Zuordnung von Detektionen zu Referenzobjekten anhand eines vorgegebenen, maximal zulässigen Abstandswertes. Aufgrund der kompakten Form der Personen, wird in dieser Arbeit die euklidische Distanz der Mittelpunkte als Entscheidungskriterium gewählt. Für eine korrekte Zuordnung bzw. Detektion darf diese nicht größer als 50 cm sein. Da Merge- und Split-Situationen im betrachteten Anwendungsszenario nicht möglich sind, werden nur 1:1-Zuordnungen zugelassen. Für jedes Bild wird deshalb immer eine global optimale GNN-Zuordnung auf Basis des Abstandes durchgeführt. Abbildung 5.1 veranschaulicht die in dieser Arbeit verwendeten Bewertungsmaße anhand eines einfachen Beispiels.

¹http://www.ipf.kit.edu/downloads_People_Tracking.php

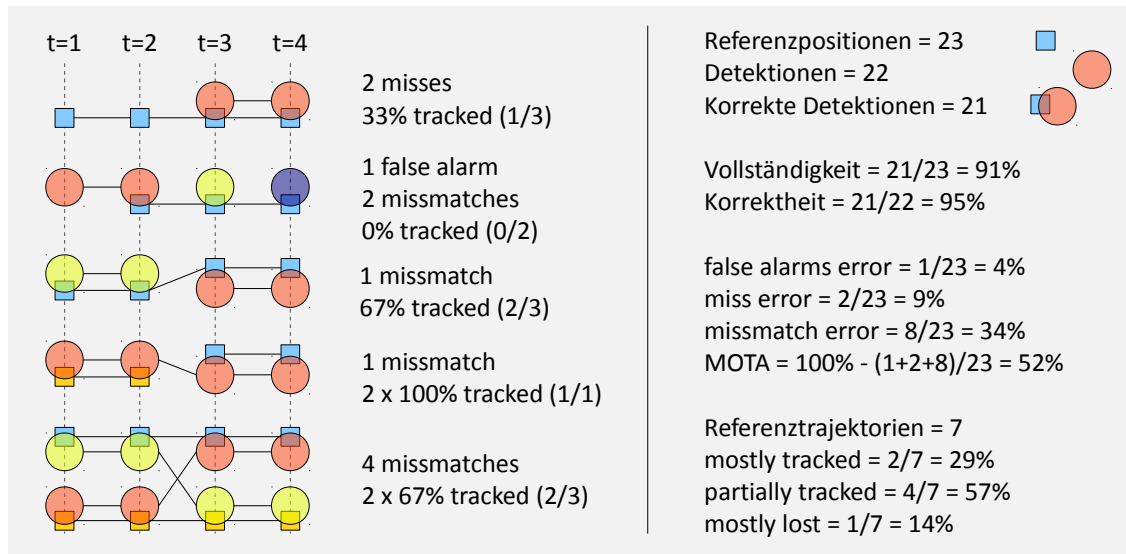


Abbildung 5.1.: Verwendeten Bewertungsmaße für Detektion und Tracking berechnet für eine Beispielsequenz mit vier Bildern ($t=1:4$), sieben Referenztrajektorien (Kästchen) und neun automatisch erzeugte Trajektorien bzw. Positionen (Kreise).

Die Qualität der Detektion wird mit Vollständigkeit (*recall*) und Korrektheit (*precision*) beschrieben und immer global für die gesamte Bildsequenz berechnet:

$$\text{Vollständigkeit} = \frac{\text{Anzahl der korrekten Detektionen}}{\text{Anzahl aller Objektpositionen}} \quad (5.1)$$

$$\text{Korrektheit} = \frac{\text{Anzahl der korrekten Detektionen}}{\text{Anzahl aller Detektionen}} \quad (5.2)$$

In den nachfolgenden Experimenten werden die beiden Maßzahlen meist gemeinsam in Form eines Precision-Recall-Diagrammes dargestellt, um eine Vergleichbarkeit unabhängig vom Detektionsschwellwert zu ermöglichen. Das häufig zur Evaluierung genutzte ROC-Diagramm ist für die Bewertung der Detektion ungeeignet, da sich die hierfür benötigte wahre Anzahl der *Nicht-Objekte* nicht bestimmen lässt.

Zur Bewertung der Tracking-Ergebnisse reichen die Vollständigkeit und Korrektheit der Positionen nicht mehr aus. Hier kommt es vor allem auf möglichst korrekte Trajektorien an. Aus den vielen möglichen Bewertungsmaßen wurden u. a. die CLEAR-MOT-Maße (Bernardin und Stiefelhagen, 2008) ausgewählt, da sie besonders häufig genutzt werden. Die *Multiple Object Tracking Accuracy* (MOTA) ist ein kumulatives Fehlermaß, welches bei einem perfekten Ergebnis 100% erreicht. Treten Falschalarme (*false alarms*), Schlupf (*misses*) oder Fehlzuzuweisungen (*mismatches*) auf, reduziert sich dieser Wert. Für all diese Fehler werden auch eigene Maße berechnet, welche sich aus dem Verhältnis der Anzahl des jeweiligen Fehlers mit der Anzahl an Objektpositionen ergeben. Anders als in (Bernardin und Stiefelhagen, 2008) definiert, werden in dieser Arbeit als Fehlzuzuweisungen sowohl *fragments* als auch *ID switches* gezählt (vgl. Li u. a. (2009)). Das Fehlermaß ist damit etwas strenger und leichter nachzuvollziehen. Die *Multiple Object Tracking Precision* (MOTP) beschreibt die Lagegenauigkeit aller Trajektorien indem der mittlere Abstand zu Referenzpositionen angegeben wird.

Neben den CLEAR-MOT-Maßen werden noch weitere genutzt, die stärker trajektorien- bzw. objektbezogen sind. Hierbei wird untersucht, wie gut die Referenztrajektorien durch eine oder mehrere automatisch generierte Trajektorien abgebildet werden (Li u. a., 2009). Anschließend kann der Anteil an Trajektorien angegeben werden, die zu 80% und mehr (*mostly tracked*), zu 80% bis 20% (*partially tracked*) und zu 20% und weniger (*mostly lost*) erkannt worden sind.

5. Evaluierung

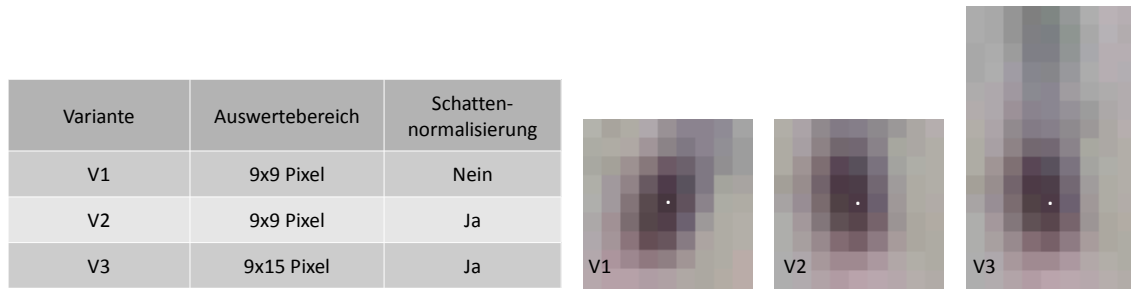


Abbildung 5.2.: Vorgaben für die drei unterschiedlichen Varianten im Experiment zum Personenschatten und ihre Auswirkungen auf den Detektor.

5.2. Objekterkennung

Nachfolgend werden die verschiedenen, in Kapitel 3 präsentierten Methoden evaluiert. Abschließend erfolgt eine vom Tracking unabhängige Bewertung der Objekterkennung.

Die optimale Einstellung der Detektionsparameter kann aufgrund ihrer Anzahl und des relativ zeitaufwändigen iterativen Klassifikatortrainings nicht für alle gemeinsam bestimmt werden. Unter der Annahme einer geringen Korrelation wird ihre endgültige Konfiguration daher nacheinander ermittelt. Innerhalb eines bestimmten Experiment wird jedoch für jede untersuchte Variante ein eigener Klassifikator trainiert, um voneinander unabhängige Ergebnisse zu erhalten und vergleichen zu können. Beim Training wird die Gesamtheit der verfügbaren Beispiele in jeder Iteration zufällig aufgeteilt. 80 % werden für das Training selbst genutzt und 20 % dienen zur Berechnung des Testfehlers.

5.2.1. Personenschatten

In den Abschnitten 3.2.2 und 3.3.2 wurde beschrieben, wie im Detektionsprozess mit dem Personenschatten umgegangen wird. In einem Vorverarbeitungsschritt wird das Bild so gedreht, dass der Schatten immer in die gleiche Richtung zeigt. Dies soll die Varianz reduzieren und die Detektionsleistung erhöhen. Darüber hinaus wird die Form des Detektors so ausgelegt, dass auch ein Teil des Personenschattens enthalten ist. Dies soll die Modellierung des Schattens als Teil des impliziten visuellen Objektmodells ermöglichen und ebenfalls die Detektionsleistung steigern.

Um diese Annahmen zu überprüfen, wurden die Detektion in drei unterschiedlichen Varianten durchgeführt (s. Abb. 5.2). Ihnen allen gemein ist, dass nur Haar-Merkmale auf dem Intensitätskanal des i1i2i3-Farbraumes genutzt wurden und 50 Schwellwerte als Basisklassifikatoren. Ansonsten verwenden Variante 1 und 2 einen 9 x 9 großen Auswertebereich, der auf den Körper der Person beschränkt ist, wohingegen Variante 3 mit einem 9 x 15 großen Bereich auch den Schatten mit erfassen kann. Variante 2 und 3 führen im Gegensatz zu Variante 1 auch eine Drehung des Bildes zur Normalisierung der Schattenrichtung durch.

Abbildung 5.3 zeigt die Detektionsleistung der unterschiedlichen Varianten für drei Bildsequenzen. In Sequenz 1 besitzen die Personen einen sehr langen Schatten und sind aufgrund des niedrigen Kontrastes nur schwer zu erkennen. Hier steigt die Detektionsleistung deutlich von Variante 1 zu 2 und auch von 2 zu 3. In Sequenz 2 werfen die Personen nur einen sehr kurzen Schatten und sind sehr gut zu erkennen. Hier steigt die Detektionsleistung ebenfalls deutlich von Variante 1 zu 2, jedoch kaum noch zu Variante 3. In Sequenz 3 scheint keine Sonne, so dass die Personen keinen Schatten besitzen und der Kontrast nur durchschnittlich ist. Hier verbessern sich die Ergebnisse nur leicht zwischen Variante 1 und 3.

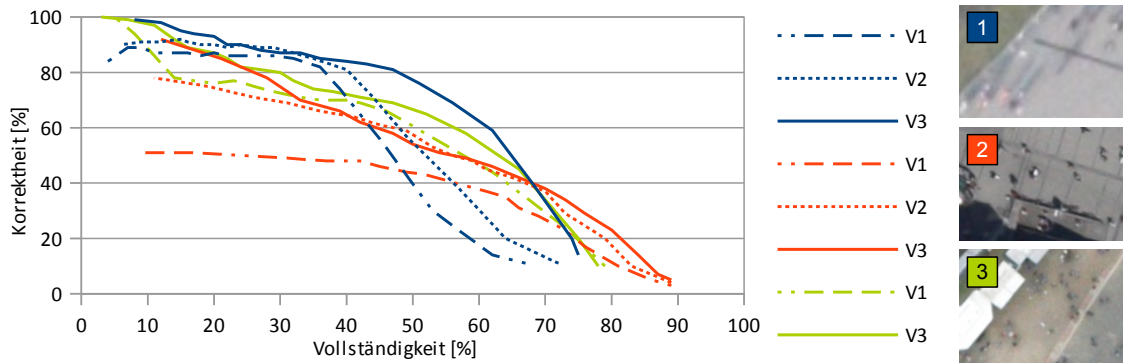


Abbildung 5.3.: Detektionsleistung der drei Varianten des Schattenexperimentes für Bildsequenzen mit unterschiedlichen Eigenschaften.

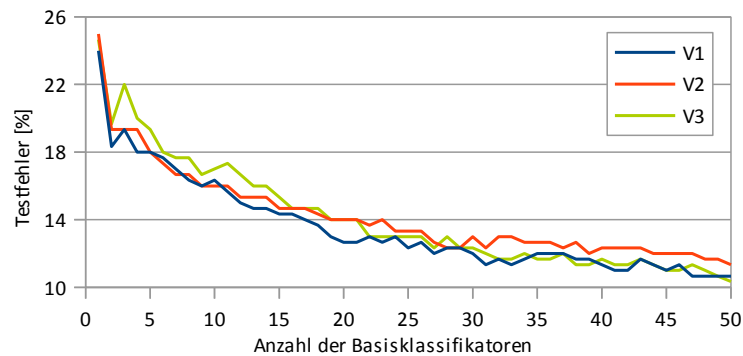


Abbildung 5.4.: Testfehler des AdaBoost-Klassifikators in Abhängigkeit der Anzahl an Basisklassifikatoren für alle drei Varianten des Experiments zum Personenschatten.

Die Ergebnisse bestätigen die anfangs aufgeführten Annahmen und zeigen deutlich, wie das beschriebene Vorgehen die Detektionsleistung erhöht. Durch die Normalisierung der Schattenrichtung und die Vergrößerung des Auswertebereiches lässt sich der Personenschatten effektiv in das implizite, visuelle Objektmodell integrieren. Der Umfang der Qualitätssteigerung hängt davon ab, wie deutlich der Schatten ausgeprägt ist und wie wichtig er zum Erkennen der Personen unter den jeweiligen Verhältnissen ist. Die leichte Verbesserung von Variante 1 zu 3 in den Ergebnissen von Sequenz 3 könnte darauf hindeuten, dass allein ein größeres Detektionsfenster Vorteile für die Objekterkennung bringt. Vergleicht man die Ergebnisse von Sequenz 3 mit denen der anderen beiden, so wird deutlich, dass der Detektor auch ohne das Vorhandensein von Personenschatten leistungsfähig ist.

Abbildung 5.4 zeigt einen weiteren Aspekt, der allgemein beim Entwerfen eines Detektors berücksichtigt werden muss. Obwohl sich die Detektionsergebnisse der drei untersuchten Varianten teils deutlich unterscheiden, liegt der Testfehler bei allen Klassifikatoren auf ähnlichem Niveau. Unterschiedliche Parametereinstellungen sollten daher immer anhand ihrer Auswirkung auf die Detektionsergebnisse verglichen werden und nicht aufgrund des Testfehlers beim Klassifikatortraining. Die Ursache hierfür liegt wahrscheinlich in der Trainingsweise begründet, in welcher für jeden Detektor die individuell geeignetsten Hintergrundbeispiele gesammelt werden.

5.2.2. Detektor

In den folgenden Unterabschnitten werden die zur Detektion genutzten Bildmerkmale evaluiert.

5. Evaluierung

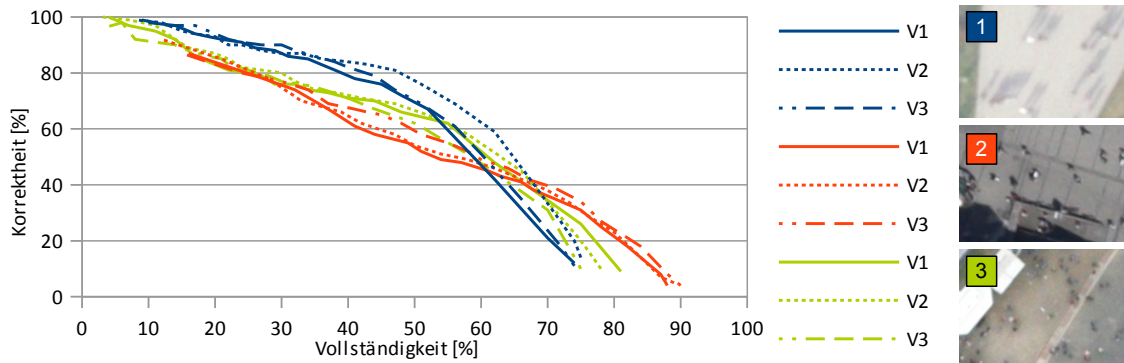


Abbildung 5.5.: Detektionsergebnisse der drei Varianten des Experiments zu Haar-Merkmalen.

Variante	Standard Merkmale	Verallgem. Merkmale	Standard + Betrag	Verallgem. + Betrag
V1	45	-	-	-
V2	30	13	-	-
V3	10	2	26	11

Tabelle 5.1.: Art und Anzahl der von den 50 Basisklassifikatoren ausgewählten Merkmale in den drei Varianten des Experiments zu Haar-Merkmalen.

Haar-Merkmale

In dieser Arbeit wurden die Haar-Merkmale verallgemeinert, um zusätzliche Varianten für Ecken und Linienenden zu erzeugen (s. Abs. 3.3.1). Diese Merkmale sollen dazu dienen, die Form der Personen noch besser zu beschreiben und eine höhere Detektionsqualität zu erzielen. Um dies zu überprüfen, wurde ein Experiment in drei Varianten durchgeführt. Die Einstellungen sind ähnlich wie im letzten Experiment in Abs. 5.2.1, nur dass diesmal grundsätzlich die Schattenrichtung normalisiert und mit einem 9×15 Pixel großen Detektor gearbeitet wird. Die Unterschiede liegen in der Art der genutzten Haar-Merkmale. Variante 1 nutzt nur den erweiterten Standard-Satz aus (Lienhart und Maydt, 2002), Variante 2 dagegen 11 zusätzliche, auf die Form der Personen angepasste Merkmale und Variante 3 alle Merkmale der Variante 2, wobei zusätzlich noch der Betrag der Merkmalswerte verwendet wird.

Abbildung 5.5 stellt die Detektionsergebnisse der drei Varianten auf ähnlichen Bildsequenzen wie im vorhergehenden Experiment dar. Für Sequenz 1 liegen die Varianten 1 und 3 auf ähnlichem Niveau während Variante 2 leicht besser abschneidet. In Sequenz 2 erreichen Variante 1 und 2 ähnliche Ergebnisse, Variante 3 ist nur etwas besser. In Sequenz 3 gibt es kaum Unterschiede zwischen allen drei Varianten. In Tabelle 5.1 ist aufgeführt, welche Merkmale in den verschiedenen Varianten automatisch ausgewählt wurden.

Insgesamt kann festgehalten werden, dass der Einfluss der verschiedenen Varianten der Haar-Merkmale auf die Ergebnisse der Objekterkennung gering ist. Dies liegt wahrscheinlich daran, dass der Satz von (Lienhart und Maydt, 2002) bereits ausreichend umfangreich ist, um die relativ einfache Form der Personen beschreiben zu können. Darauf deuten auch die Ergebnisse in Tabelle 5.1 hin. In Variante 2 und 3 stammen nur etwa ein Viertel aller Merkmale aus dem zusätzlichen Satz. Variante 3, in welcher zusätzlich zum direkten Merkmalswert auch dessen Betrag genutzt wird, um eine Invarianz gegenüber hellen und dunklen Personen zu erzielen, führt zu keiner eindeutigen Steigerung der Detektionsleistung, obwohl etliche dieser Merkmale ausgewählt wurden. Dies liegt wahrscheinlich daran, dass in den Sequenzen nur ein Bruchteil der Personen hell erscheinen und der AdaBoost-Klassifikator

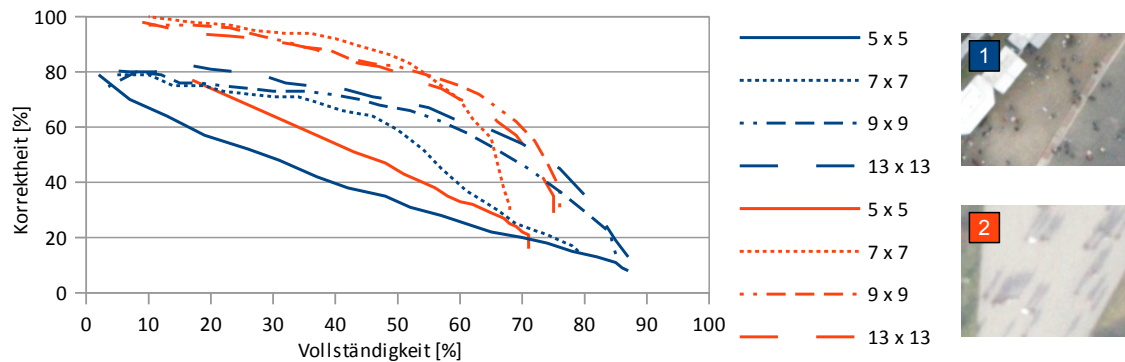


Abbildung 5.6.: Detektionsleistung der Rechteckmerkmale für zwei Sequenzen bei variierender Detektorgröße.

diese Variation bereits durch Kombination der 50 Basisklassifikatoren abbilden kann. Es ist zu vermuten, dass die Vorteile der zusätzlichen Haar-Merkmale bei komplexeren Objektformen oder bei Verwendung von weniger Basisklassifikatoren deutlicher hervortreten.

In den weiteren Experimenten werden die in Variante 2 verwendeten Haar-Merkmalen genutzt, da diese in einigen Fällen eine leichte Leistungssteigerung hervorrufen. Auf die zusätzlichen Merkmale in Variante 3 wird dagegen verzichtet, da hier die verdoppelte Anzahl an Merkmalen die Zeit für das Klassifikatortraining deutlich erhöht, ohne zu einer Verbesserung der Ergebnisse zu führen.

Rechteckmerkmale

In dieser Arbeit wurden die sog. Rechteckmerkmale entwickelt, um komplementäre Informationen zu den auf die Objektform fokussierten Haar-Merkmalen für die Detektion nutzen zu können. In diesem Experiment werden die Rechteckmerkmale jedoch vorerst separat untersucht. Hierfür wurde ein Detektor allein mit Rechteckmerkmalen, einem 9x9 Pixel großen Auswertebereich und unter Verwendung aller drei Kanäle des i1i2i3-Farbraumes trainiert. Anschließend wurde die Detektionsleistung bei variierender Detektorgröße getestet.

In Abbildung 5.6 sind repräsentative Ergebnisse für zwei Bildsequenzen dargestellt. Die Detektionsleistung ist für einen 5x5 Pixel großen Detektor am geringsten, sie steigt bei 7x7 Pixel deutlich an und konvergiert schließlich ab einer Größe von 9x9 Pixel. Ein zu klein gewählter Auswertebereich verschlechtert also die Ergebnisse der Objekterkennung mit Rechteckmerkmalen deutlich. Ein 9x9 Pixel großer Bereich liefert gute Ergebnisse ohne zu viele Merkmale zu ermöglichen. Im Gegensatz zu einem 13x13 großen Bereich mit 49.000 möglichen Merkmalen, lassen sich nur 12.000 extrahieren, was den Trainingsprozess deutlich beschleunigt.

In Tabelle 5.2 sind diejenigen Rechteckmerkmale dargestellt, welche bei einem 9x9 Pixel großen Auswertebereich von den 50 Basisklassifikatoren ausgewählt wurden. Die meisten nutzen den Intensitätskanal i1. Dies zeigt, dass die Farbkanäle i2 und i3 deutlich weniger für die Objekterkennung nützliche Informationen enthalten. Insgesamt wurden deutlich mehr Varianz- als Mittelwertmerkmale ausgewählt. Dies lässt den Schluss zu, dass Informationen über die Objektform, wenn auch verschlüsselt als Texturparameter, für das Erkennen von Personen in Luftbildsequenzen wesentlich nützlicher sind als reine Farbwerte.

5. Evaluierung

Farbkanal	Rechteck-Merkmale Mittelwert	Rechteck-Merkmale Varianz
i1	9	20
i2	3	2
i3	2	1

Tabelle 5.2.: Art, Anzahl und Farbkanal der, von 50 Basisklassifikatoren ausgewählten Rechteckmerkmale.

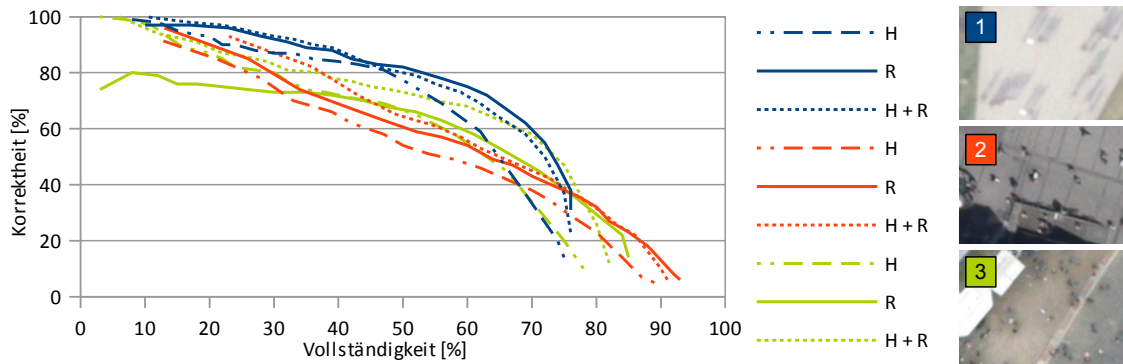


Abbildung 5.7.: Detektionsleistung für verschiedene Bildsequenzen unter Nutzung folgender Bildmerkmale: Haar-Merkmale allein (H), Rechteckmerkmale allein (R) und Haar- und Rechteckmerkmale kombiniert (H+R).

Kombination der Bildmerkmale

Die gemeinsame Nutzung von Haar- und Rechteckmerkmalen soll die Detektionsleistung steigern, da komplementäre Informationen in den Prozess zur Objekterkennung eingebracht werden. Abbildung 5.7 zeigt die Ergebnisse eines Experiments, in welchem Haar- und Rechteckmerkmale einzeln und zusammen genutzt wurden.

Für die erste Sequenz liefern die Haar-Merkmale allein das schlechteste Ergebnis. Die anderen beiden Varianten liegen auf einem höheren, ähnlichen Niveau. In Sequenz 2 erzielen Haar-Merkmale ebenfalls das niedrigste Ergebnis. Die restlichen Varianten liegen jedoch nur leicht höher. In Sequenz 3 liefert die Kombination der Merkmale die besten Ergebnisse, die anderen beiden Varianten sind etwas schlechter.

Es ist interessant, dass die Ergebnisse der Haar-Merkmale meist niedriger als die der Rechteckmerkmale ausfallen. Dies könnte daher rühren, dass die Form der Personen doch so stark variiert, dass sie sich durch ein Varianz-Rechteckmerkmal besser beschreiben lässt. Dieses ist nicht so anfällig gegenüber Formvariationen wie ein bestimmtes Haar-Merkmal. Insgesamt betrachtet liefert die Kombination von Haar- und Rechteckmerkmalen jedoch in allen Sequenzen mit die besten Ergebnisse, auch wenn diese teilweise nur leicht besser sind als die Rechteckmerkmale allein.

Ein großer Unterschied zwischen den einzelnen Merkmalsarten und einer Kombination zeigt sich im Training. Konnten im ersten Fall zu den 2.500 Objektbeispielen die gleiche Anzahl Hintergrundbeispiele automatisch gesammelt werden, so war dies im zweiten Fall nicht möglich. Hier ließen sich aus den verwendeten Bildsequenzen nur 1.200 Hintergrundbeispiele extrahieren. Die Kombination komplementärer Bildmerkmale in einem Detektor reduziert somit den Aufwand für das iterative Training.

Darüber hinaus ist ein weiterer Aspekt bemerkenswert. Die in Abbildung 5.7 dargestellten Sequenzen wurden ebenfalls zur automatischen Gewinnung von Hintergrundbeispielen für

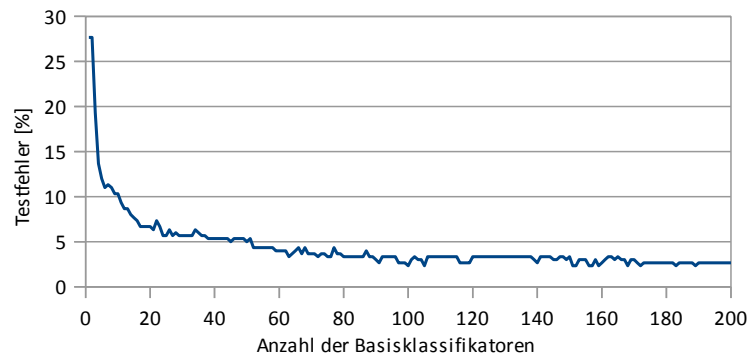


Abbildung 5.8.: Testfehler des AdaBoost-Klassifikators in Abhängigkeit von der Anzahl verwendeter Basisklassifikatoren (Schwellwerte).

das Klassifikatortraining genutzt. Obwohl sie im aktuellen Experiment etliche Fehldetektionen aufweisen, wurden diese Stellen im Vorfeld nicht automatisch als Hintergrundbeispiele ausgewählt. Die Ursache hierfür liegt darin, dass diese Orte eine sehr hohe Ähnlichkeit mit der Objektklasse aufweisen und daher nicht verwendet wurden. Dies impliziert, dass sich die Korrektheit mit dem in Kapitel 3 beschriebenen Verfahren zur Objekterkennung kaum noch steigern lässt, da alle übrigen Falschalarme sich anhand ihres äußeren Erscheinungsbildes nicht von echten Personen unterscheiden lassen. Eine Verbesserung wird daher nur möglich sein, wenn zusätzliche, komplementäre Informationen in den Detektionsprozess eingebracht werden können. In dieser Arbeit geschieht dies, indem auch die Konsistenz der Detektionen über die Zeit analysiert wird (s. Kap. 4).

5.2.3. Klassifikator

Anzahl Basisklassifikatoren

Bei allen bisherigen Experimenten zu Bildmerkmalen und Personenschatten bestand der AdaBoost-Klassifikator aus 50 Basisklassifikatoren. Nun soll der Einfluss der Anzahl näher untersucht und der optimale Wert bestimmt werden. Die Art der Basisklassifikatoren spielt dabei nur eine untergeordnete Rolle. Bisher wurden Schwellwerte genutzt. Üblich sind auch Entscheidungsbäume geringer Tiefe. Diese benötigen jedoch bei gleicher Anzahl insgesamt deutlich mehr Merkmale als einfache Schwellwerte. Für das AdaBoost-Verfahren (Freund und Schapire, 1997) ist es ausreichend, wenn die Basisklassifikatoren besser als der Zufall funktionieren.

Für das Experiment wurde ein Training mit 200 Schwellwerten durchgeführt, wobei sowohl Haar- als auch Rechteckmerkmale mit ihrem optimalen Auswertebereich genutzt wurden. Anschließend wurde die Anzahl der Basisklassifikatoren reduziert und verschiedene Evaluationen durchgeführt. In Abbildung 5.8 ist der Testfehler des Klassifikators in Abhängigkeit von der Anzahl an Basisklassifikatoren dargestellt. Man erkennt, dass der Fehler sehr schnell sinkt und ab etwa 80 Schwellwerten bei 2% bis 3% konvergiert. Dies zeigt, dass der AdaBoost-Klassifikator die Trainingsbeispiele anhand der definierten Merkmale sehr gut unterscheiden kann und auch viele Basisklassifikatoren nicht zu einer Überanpassung führen.

In Grafik 5.9 sind die Detektionsergebnisse für zwei Sequenzen dargestellt, wobei die Anzahl der Basisklassifikatoren zwischen 200 und 25 variiert. Man erkennt, dass sich mit steigender Anzahl die Ergebnisse verbessern und gegen ein Maximum konvergieren. In der linken Abbildung geschieht dies bereits ab 50 Klassifikatoren, in der rechten erst ab 100. Diese Ergebnisse decken sich im wesentlichen mit dem Verlauf des Testfehlers. Zukünftig wird daher immer mit 100 Schwellwerten als Basisklassifikatoren gearbeitet. Insgesamt sind so deut-

5. Evaluierung

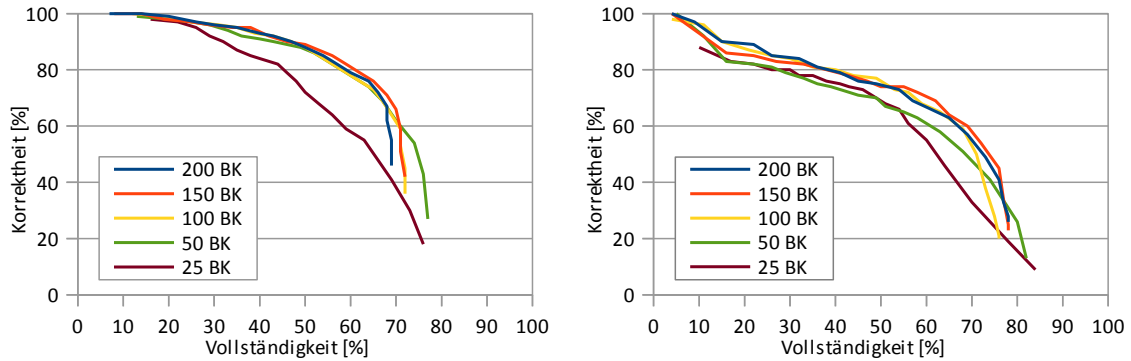


Abbildung 5.9.: Detektionsleistung für zwei Sequenzen in Abhängigkeit der Anzahl an Basisklassifikatoren (BK).

Variante	Beispiele in Objektnähe	Auswahlkriterium
V1	ja	Wert des Beispiels
V2	nein	Wert des Beispiels
V3	ja	Zufall

Tabelle 5.3.: Konfiguration der drei Varianten des Experiments zur Untersuchung der Vorschläge zur Verbesserung der üblichen Trainingsmethode.

lich bessere Ergebnisse möglich als in einer früheren Arbeit dargestellt (Schmidt und Hinz, 2011), in welcher nur 16 Schwellwerte genutzt wurden. Zusätzlich reduziert sich die Anzahl der Bildmerkmale enorm, von 21.000 im Training auf 100 für die Objekterkennung.

Trainingsmethoden

Im nachfolgenden Experiment soll ermittelt werden, welche Auswirkungen die in Abschnitt 3.4.2 vorgeschlagenen Anpassungen der üblichen Trainingsprozedur auf die Detektionsergebnisse haben. Der Versuch umfasst drei Varianten, deren jeweilige Konfiguration in Tabelle 5.3 zusammengefasst ist. Sie unterscheiden sich dahingehend, ob Negativbeispiele auch in Objektnähe gesammelt werden und ob deren Auswahl zufällig oder aufgrund des jeweiligen Wertes erfolgt. Die Ergebnisse des Experiments sind in Abbildung 5.10 dargestellt.

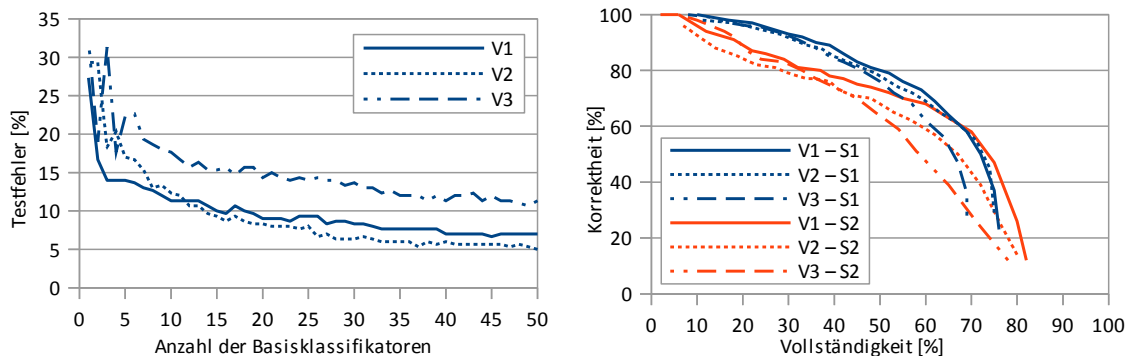


Abbildung 5.10.: Ergebnisse des Trainings (links) und der Objekterkennung für zwei Sequenzen S1 und S2 (rechts) der drei Varianten des Experiments zum Sammeln von Hintergrundbeispielen im Training (s. Abs. 5.2.3).

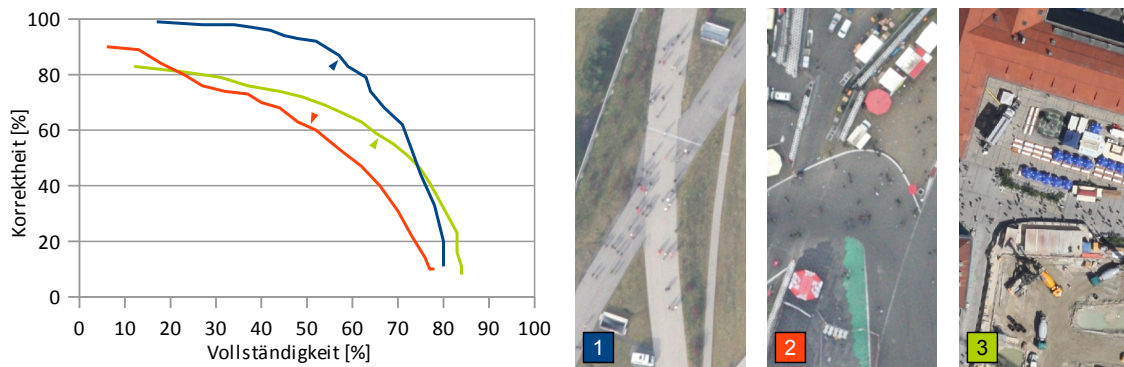


Abbildung 5.11.: Detektionsergebnisse für drei Testsequenzen. Die Werte, welche mit einem üblichen Detektionsschwellwert von z. B. 0,5 Konfidenz erreicht werden, sind besonders markiert.

Betrachtet man den Testfehler beim Klassifikatortraining, so ist Variante 3 meist 5 % schlechter als die anderen beiden. Dies deutet darauf hin, dass bei der zufälligen Auswahl von Hintergrundbeispielen auch solche genommen wurden, die Personen sehr ähnlich sind. Damit erschwert sich das Trennungsproblem und der Testfehler steigt an. Variante 1 hat einen leicht höheren Testfehler als Variante 2. Die Ursache könnte sein, dass die zusätzlichen Hintergrundbeispiele in Objektnähe, welche häufig aus dem Personenschatten stammen, ebenfalls das Trennungsproblem erschweren.

In Abbildung 5.10 werden auch die Auswirkungen auf die Detektionsergebnisse dargestellt. Hier ist zu beobachten, dass Variante 3 deutlich schlechter abschneidet als die anderen beiden. Variante 1 führt zu den besten Ergebnissen, wobei in Sequenz 1 Variante 2 auf ähnlichem Niveau liegt. Generell lässt sich folgern, dass die automatische Auswahl der Hintergrundbeispiele immer anhand ihres Wertes und nicht zufällig erfolgen sollte. Zudem kann es sich positiv auswirken, wenn auch Beispiele in Objektnähe gesammelt werden, statt solche Bilder oder Bereiche im Training überhaupt nicht zu berücksichtigen.

5.2.4. Bewertung der Detektion

Nachdem bisher die unterschiedlichen Komponenten der Detektion einzeln untersucht wurden, folgt nun eine abschließende Evaluierung des Gesamtprozesses. Folgende Einstellungen werden hierfür genutzt: Schattennormalisierung, Detektorkonfiguration wie in Tab. 3.1, 100 Schwellwerte als Basisklassifikatoren und die in Abs. 3.4.2 beschriebene Trainingsmethode. Abbildung 5.11 zeigt die Detektionsergebnisse für drei verschiedene Testsequenzen. Wird ein sehr hoher Detektionsschwellwert genutzt, steigt die Korrektheit auf maximal 80 % bis 100 % bei einer Vollständigkeit von unter 20 %. Senkt man den Schwellwert dagegen stark ab, erreicht die Vollständigkeit 80 % bei einer Korrektheit von 10 %.

Die Höhe der Korrektheit wird u. a. dadurch beeinflusst, wie viele personenähnliche Objekte in den Bilddaten enthalten sind. Dies ist z. B. in Sequenz 2 und 3 deutlich stärker der Fall als in Sequenz 1, was eine etwa 20 % geringere Korrektheit zur Folge hat. Wie in Abschnitt 5.2.2 bereits angesprochen, stößt die Detektion mit implizitem visuellem Modell hier an ihre Grenzen. Allein auf Basis des äußeren Erscheinungsbildes wird sich keine deutliche Steigerung der Korrektheit mehr erreichen lassen. Die spektrale Signatur von Personen in Luftbildern mit einer Bodenauflösung von unter 10 cm ist schlicht nicht charakteristisch genug, um eine eindeutige Detektion zu erlauben. Es werden zusätzliche, komplementäre Informationen benötigt, um die Anzahl der Falschalarme weiter zu reduzieren.

Die Vollständigkeit erreicht in den Testsequenzen maximal 80 %. Hier zeigt sich ebenfalls der Einfluss der niedrigen Auflösung und des geringen Signal-Rausch-Verhältnisses. Dies trifft

5. Evaluierung

speziell für Sequenz 2 zu. Da hier im Gegensatz zu den anderen Sequenzen keine Sonne scheint, besitzen die Personen keinen Schatten und heben sich nur schwach vom dunklen Untergrund ab. Die Vollständigkeit ist daher meist 10 % niedriger als in den anderen Sequenzen. Die Erkennbarkeit von Einzelpersonen und damit die Höhe der Vollständigkeit wird auch durch die Personendichte stark beeinflusst. Kommt es zu Gruppenbildung, was in allen drei Sequenzen zu beobachten ist, sind einzelne Personen visuell nicht mehr erkennbar. Da die Referenzdaten trotzdem für jede Person eine Position enthalten, sei sie auch nur durch den Erfasser geschätzt bzw. interpoliert, erreichen die Detektionsergebnisse nie eine Vollständigkeit von 100 %. Der Wert ließe sich u. a. durch Auswertung der Bildsequenz steigern. Sollte die Einzelperson nur für einen kurzen Moment nicht sichtbar sein, könnte ihre Position mit Hilfe eines Bewegungsmodells geschätzt werden. Es wäre auch denkbar einen zusätzlichen Detektionsprozess für Personengruppen oder gar Menschenmengen zu entwickeln. So ließen sich all die Situationen behandeln, in denen Einzelpersonen zwar noch physisch im Beobachtungsgebiet vorhanden, jedoch visuell Teil eines größeren Objektes geworden sind. Ihre jeweilige Position ließe sich dann jedoch nur noch grob abschätzen.

In Abbildung 5.11 ist die Stelle markiert, welche mit einem häufig verwendeten Detektionsschwellwert von 0,5 Konfidenz erreicht wird. Die Korrektheit liegt zwischen 60 % und 90 % bei einer Vollständigkeit von 50 % bis 70 %. Sollen diese Detektionen als Grundlage für ein Tracking-by-Detection-Verfahren genutzt werden, sind diese Werte insgesamt zu niedrig, um gute Ergebnisse erzielen zu können. Der in dieser Arbeit vorgeschlagenen Tracking-Ansatz soll trotz dieser Probleme möglichst vollständige Trajektorien liefern. Er erhöht die Vollständigkeit korrekter Detektionen, indem der Schwellwert abgesenkt und die Detektionswerte stochastisch beschrieben werden. Gleichzeitig soll der steigende Anzahl an Falschalarmen und Mehrdeutigkeiten durch das Berücksichtigen mehrerer alternativen Hypothesen entgegengewirkt werden.

5.3. Objektverfolgung

In diesem Abschnitt werden sowohl das Gesamtsystem zur Objektverfolgung als auch einige seiner Komponenten evaluiert. Ähnlich wie bei der Objekterkennung müssen auch hier etliche Parameter gesetzt werden. Die meisten sind nur schwach korreliert und können nacheinander mit Hilfe von Trainingssequenzen bestimmt werden. Obwohl einige Parameter in Abhängigkeit von den Eigenschaften der jeweiligen Bildsequenz gesetzt werden müssten, um optimale Ergebnisse zu erzielen, wird in den nachfolgenden Experimenten aus Gründen einer besseren Vergleichbarkeit für alle Sequenzen immer der gleiche Parametersatz verwendet.

5.3.1. Hypothesenanzahl

Die Besonderheit des MHT-Verfahrens ist die Möglichkeit mehrere Tracking-Hypothesen gleichzeitig zu verfolgen und so der global optimalen Lösung für die Sequenz näher zu kommen als andere Ansätze. Die Abhängigkeit zwischen Hypothesenanzahl und Qualität der erzeugten Trajektorien wird nun untersucht. Im durchgeführten Experiment wurde die Auftrittswahrscheinlichkeit von neuen Objekten und Falschalarmen nicht automatisch bestimmt sondern vorgegeben und war somit für alle Durchläufe einer Sequenz gleich. In Abbildung 5.12 sind die Ergebnisse des Versuches dargestellt.

Für alle drei Sequenzen sind bei steigender Hypothesenanzahl die gleichen Effekte zu beobachten. Im Vergleich zu den Ergebnissen mit einer einzigen Hypothese (*Single Hypothesis Tracking*, SHT), nimmt bei mehr Hypothesen die Anzahl der Trajektorien, welche weniger als 20 % erkannt wurden um maximal 6 % bis 15 % ab und der Anteil von Trajektorien, welche zu über 80 % erkannt wurden steigt um 4 % bis 10 % an. Der Anteil an Falschalarmen in den Ergebnissen nimmt während dessen um 3 % bis 9 % zu. Weiterhin ist zu beobachten, dass

5.3. Objektverfolgung

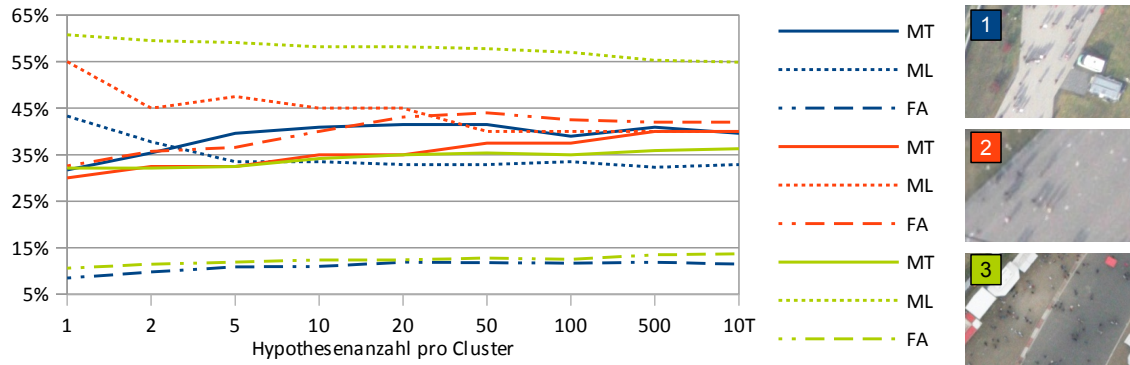


Abbildung 5.12.: Tracking-Kennzahlen (*mostly tracked*, *mostly lost*, *false alarms*) in Abhängigkeit der Anzahl verfolgter Hypothesen pro Cluster für drei Sequenzen.

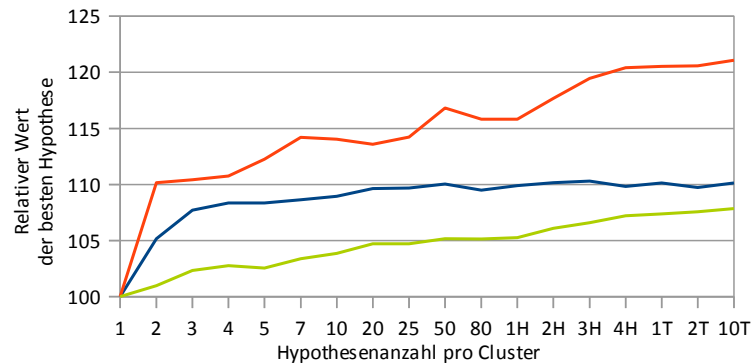


Abbildung 5.13.: Relativer Wert der besten Hypothese in Abhängigkeit der Anzahl verfolgter Hypothesen pro Cluster für drei Sequenzen.

die größten Änderungen bereits bei einer leichten Erhöhung der Hypothesenanzahl eintreten. Um diesen Effekt näher zu untersuchen, ist in Abbildung 5.13 der Verlauf des Wertes der jeweils besten Hypothese dargestellt. Hier erkennt man, dass bereits 20 Hypothesen eine Steigerung von zwischen 5 % bis 15 % gegenüber SHT hervorrufen können. Anschließend müssen für weitere, teils nur geringe Zunahmen des Hypothesenwertes deutlich mehr Hypothesen pro Cluster verfolgt werden. Vergleicht man Abb. 5.12 und 5.13, so kann man feststellen, dass der Hypothesenwert durch Hinzunahme weiterer Hypothesen zum Teil deutlich ansteigt, ohne dass dies jedoch zu einer zusätzlichen Steigerung der Tracking-Ergebnisse führt.

Die Resultate belegen den Vorteil von MHT gegenüber SHT. Die Vollständigkeit der Trajektorien nimmt im Allgemeinen deutlich zu. Dieser Umstand führt jedoch gleichzeitig zu einem leichten Anstieg der Falschalarme, da auch diese, sollten sie in konsistenter Weise mehrmals auftreten, besser verfolgt werden. Dies ist besonders dann der Fall, wenn in einer Szene viele personenähnliche Objekte vorhanden sind und demzufolge auch viele Fehldetektionen hervorrufen. Um diesem Effekt entgegenzuwirken, sind weitere Analysemethoden notwendig, welche die erzeugten Trajektorien anhand zusätzlicher, komplementärer Informationen klassifizieren. Des Weiteren verdeutlichen die Ergebnisse auch den positiven Effekt des Cluster-Verfahrens. Dadurch, dass bereits wenige Hypothesen pro Cluster die globale Anzahl an Hypothesen enorm steigern, reichen zum Tracking meist relativ wenige Hypothesen aus, um deutlich näher an das globale Optimum zu kommen.

5. Evaluierung

Methode	Sequenz 1		Sequenz 2		Sequenz 3	
	Auto	Fix	Auto	Fix	Auto	Fix
Falschalarme	30%	29%	4%	7%	23%	29%
Schlupf	40%	36%	48%	43%	52%	49%

Tabelle 5.4.: Tracking-Ergebnisse für drei Sequenzen, einmal mit der automatischen Methode zur Bestimmung der Verteilung der Auftrittswahrscheinlichkeit (Auto) und einmal mit fest vorgegebenen Werten und einer Gleichverteilung (Fix).

5.3.2. Auftrittswahrscheinlichkeit

Damit das MHT-Verfahren zu besseren Tracking-Ergebnissen führt, muss die stochastische Bewertungsfunktion die Zusammenhänge in der Szene möglichst realistisch wiedergeben. Aus diesem Grund wurde eine automatische Methode zur adaptiven Bestimmung der Auftrittswahrscheinlichkeit von neuen Objekten und Falschalarmen entwickelt (s. Abs. 4.2.3). Ihr Einfluss auf die Tracking-Ergebnisse wird nun dargestellt. In dem Versuch wurden 20 Hypothesen pro Cluster sowie ein gleitendes Zeitfenster von 2,5 s genutzt. Tabelle 5.4 stellt die Ergebnisse des Experimentes dar, in welchem einmal die vorgeschlagene Methode und einmal fest vorgegebene Werte mit einer Gleichverteilung verwendet wurden. Die Werte für neue Objekte und Falschalarme wurden vorab aus den Referenzdaten bzw. den Detektionsergebnissen ermittelt.

Man sieht, dass die Menge der Falschalarme mit der automatischen Methode in Sequenz 1 um 1 % höher liegt und in Sequenzen 2 und 3 um 3 % bzw. 6 % niedriger. Der Schlupf erhöht sich mit der automatischen Methode in allen Sequenzen um 3 % bis 5 %. Die Werte zeigen, dass die vorgeschlagene Methode nicht generell zu besseren Ergebnissen bei Schlupf und Falschalarmen führt, verglichen mit einer Gleichverteilung mit festen Werten. Hierbei ist jedoch zu berücksichtigen, dass in einer echten Anwendung die vorgegebenen Werte deutlich schlechter geschätzt worden wären, als dies in diesem Experiment durch Analyse der Referenzdaten möglich war.

Eine genauere Betrachtung der Verteilung neu und falsch erkannter Objekte zeigt jedoch eine Schwäche der eigenen Methode. Da sie sich auf die eigenen Ergebnisse aus vorhergehenden Zeitpunkten stützt, kann es zur Bestätigung einer fehlerhaften Verteilung kommen. Obwohl in deren Bestimmung alternative Hypothesen mit einfließen, kann diese Situation nicht gänzlich ausgeschlossen werden. Problematisch sind vor allem Bereiche mit hoher Personendichte (vgl. Abb. 4.1). Wenn beispielsweise Personen nur kurz erkannt werden und dann in einer Gruppe verschwinden, wird ihre gesamte Trajektorie meist als Falschalarm deklariert und erhöht damit die Auftrittswahrscheinlichkeit für Fehldetektionen speziell in den Bereichen, wo sich sehr viele Personen befinden. Gleichzeitig steigt auch die Wahrscheinlichkeit neuer Personen in Bereichen, wo Einzelpersonen sich aus Gruppen lösen.

Insgesamt stellt das Verfahren jedoch eine gute Möglichkeit im Falle nicht vorhandener Vorinformationen, die Auftrittswahrscheinlichkeit im Rahmen des MHT-Verfahrens automatisch und adaptiv zu ermitteln. Eine Verbesserung ließe sich erreichen, wenn sowohl die Detektionsleistung als auch die Klassifizierung der Trajektorien optimiert werden würden. Des Weiteren wäre es auch denkbar, die Verteilungen nicht ausschließlich adaptiv zu ermitteln, sondern Vorwissen mit einzubringen. Dies würde sich vor allem bei sehr kurzen Sequenzen, wie sie zur Durchführung dieses Experiments genutzt wurden, positiv auswirken.

			Sequenz 1		Sequenz 2		Sequenz 3	
$P(s > T D, H_1)$	T	$P(s > T D, H_0)$	MOTA D	MOTA	MOTA D	MOTA	MOTA D	MOTA
100%	-1,00	100%	39%	8%	40%	24%	-8%	-
99,9%	-0,36	77%	41%	2%	42%	25%	0%	-
99,5%	-0,21	21%	44%	17%	48%	27%	12%	-63%
99%	-0,15	8%	44%	30%	49%	37%	14%	-48%
98%	-0,08	3%	45%	37%	44%	42%	17%	-18%
95%	0,00	0,6%	41%	40%	45%	45%	21%	-21%
90%	0,07	0,2%	38%	32%	39%	38%	20%	5%
80%	0,15	0,1%	28%	28%	31%	30%	15%	18%

Tabelle 5.5.: Tracking-Ergebnisse mit und ohne Detektionswert (D) für verschiedene Schwellwerte (T).

5.3.3. Integration des Detektionswertes

Die Integration der Objekterkennung in das MHT-Verfahren mittels stochastischer Modellierung des Detektionswertes (vgl. Abb. 2.3) ist einer der zentralen Beiträge dieser Arbeit. In dem folgenden Experiment wurden die Auswirkungen dieses Vorgehens auf die Tracking-Ergebnisse untersucht. Hierfür wurden drei Sequenzen bei variierendem Detektionsschwellwert prozessiert, einmal mit und einmal ohne den Detektionswert zu nutzen. Die Ergebnisse sind in Tabelle 5.5 dargestellt.

Die ersten drei Spalten zeigen den Zusammenhang zwischen Schwellwert und der Verteilung des Detektionswertes für korrekte und falsche Detektionen (vgl. Abb. 3.6). Statt den Schwellwert direkt vorzugeben, ergibt sich dieser aus der Wahrscheinlichkeit, dass eine korrekte Detektion erkannt wird (s. erste Spalte). Vergleicht man die Ergebnisse mit und ohne Detektionswert, so treten die Vorteile der Integration deutlich hervor. Erstens ist die MOTA-Kennzahl für alle drei Sequenzen mit Detektionswert deutlich höher als ohne, was vor allem auf einen niedrigeren Anteil an Schlupf und Falschalarmen zurückzuführen ist. Zweitens bleiben die Ergebnisse auch bei einem niedrigeren Schwellwert auf hohem Niveau, obwohl deutlich mehr Fehldetektionen in den Tracking-Prozess einfließen. Insgesamt lassen sich gute Ergebnisse über einen größeren Schwellwertbereich erzielen, was beweist, dass die Integration des Detektionswertes die Abhängigkeit des Trackingverfahrens vom Schwellwert reduziert.

Völlig ohne Detektionsschwellwert zu arbeiten, ist jedoch nicht zu empfehlen. Obwohl die zusätzlich ins Tracking eingebrachten Detektionen nur ein sehr geringes Gewicht besitzen, können sie doch die Ergebnisse unter Umständen verschlechtern, wie es bei Sequenz 2 und 3 zu beobachten ist. Dies liegt u. a. daran, dass die Auftrittswahrscheinlichkeit von Falschalarmen stark zunimmt und die Generierung korrekter Trajektorien erschwert. Zudem treten deutlich mehr Mehrdeutigkeiten auf, was die Prozessierungszeit erhöht und zu mehr Zuordnungsfehlern führt. Eine hohe Wahrscheinlichkeit für korrekte Detektionen von 98 % liefert in den meisten Fällen ausreichend gute Ergebnisse.

5.3.4. Laufzeit

Das MHT-Verfahren ist deutlich rechenaufwändiger als andere Tracking-Verfahren, was vor allem daran liegt, dass mehrere Hypothesen parallel verfolgt werden. Wie sich dieser Umstand auf die Laufzeit auswirkt, wurde in dem folgenden Experiment näher untersucht. Obwohl die absolute Laufzeit von vielen, vom Algorithmus unabhängigen Faktoren beeinflusst wird, so soll das Experiment dennoch die Größenordnung aufzeigen, welche man mit dem

5. Evaluierung

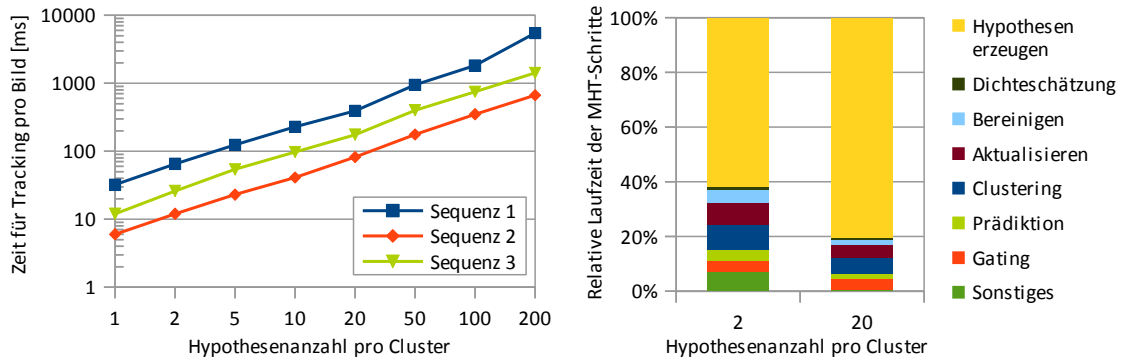



Abbildung 5.14.: Absolute MHT-Laufzeit pro Bild für drei verschiedene Sequenzen ohne Detektion (links). Relative Laufzeit der einzelnen MHT-Schritte (rechts).

MHT-Verfahren erreichen kann. Durchgeführt wurde der Versuch auf einem Standard-PC² unter Verwendung eines einzelnen Prozessorkerns. Die Implementierung der Methoden erfolgte vollständig in C++. Abbildung 5.14 stellt die Ergebnisse der Laufzeitmessung dar.

Man erkennt, dass die absolute Laufzeit, wie erwartet, mit steigender Anzahl an Hypothesen pro Cluster zunimmt. Lässt man die Zeit für die Detektion außer acht und nimmt eine Aufnahme­frequenz von 500 ms pro Bild an, so ist das Tracking-Verfahren je nach Sequenz bei 20 bis 100 Hypothesen pro Cluster echtzeitfähig. Unterschiede zwischen den Sequenzen sind auf die jeweiligen Gegebenheiten zurückzuführen. So sind etwa 400 Detektionen und 200 Personen pro Bild in Sequenz 1 enthalten, 100 Detektionen und 80 Personen in Sequenz 2, und 200 Detektionen und 160 Personen in Sequenz 3. In Abbildung 5.14 ist auch die relative Laufzeit der diversen MHT-Prozessierungsschritte gemittelt über mehrere Sequenzen dargestellt. Der zeitaufwändigste Schritt ist die Generierung neuer Hypothesen. Dieser benötigt in den zwei untersuchten Varianten mit 2 und 20 Hypothesen pro Cluster 60 % bzw. 80 % der gesamten Laufzeit. Die restlichen Schritte brauchen jeweils nur wenige Prozent. Das Verfahren zur Bestimmung der n besten Hypothesen wurde in dieser Arbeit mit Hilfe der Vorschläge aus (Miller u. a., 1997) beschleunigt. Es konnten jedoch nur zwei von drei Verbesserungen nachimplementiert werden. Ließe sich auch die letzte Verbesserung umsetzen, kann eine weitere Absenkung der Laufzeit und eine nahezu lineare Abhängigkeit von der Hypothesenanzahl erwartet werden. Letzteres ist aktuell nicht der Fall, wie sich eindeutig in beiden Grafiken in Abb. 5.14 unter Beachtung der logarithmische Zeitskala erkennen lässt.

Obwohl noch weitere Laufzeitverbesserungen möglich sind, u. a. auch durch die Prozessierung auf mehreren Kernen, zeigen die Resultate, dass das MHT-Verfahren auch bei sehr vielen Objekten und Detektionen in Echtzeit ausgeführt werden kann. Besonders die Möglichkeit die Laufzeit über die Anzahl der gewünschten Hypothesen beeinflussen zu können, erlaubt es, den Algorithmus an wechselnden Gegebenheiten und Anforderungen adaptiv und online anzupassen. Diese Möglichkeit bringt für viele Anwendungen große Vorteile. Für die vollständige Prozessierung einer Luftbildsequenz benötigt man für die Detektion jedoch wesentlich mehr Zeit als für das Tracking. Unter Verwendung der in Abs. 5.2.4 aufgeführten Parameter und vier Prozessoren benötigt die Detektion für 100 x 100 Pixel etwa 110 ms bis 180 ms. Soll also das Gesamtsystem in Echtzeit funktionieren, muss der Beobachtungsbereich entsprechend eingeschränkt oder die Detektion insgesamt beschleunigt werden (vgl. Abs. 3.5).

²Intel Quad Core 2 mit jeweils 2,66 GHz, 8 GB RAM, 300GB HDD, Windows 7



Sequenz	FA	Miss	MM	MOTP	MT	PT	ML
1	4%	47%	2%	13cm	42%	17%	42%
2	15%	64%	1%	9cm	29%	7%	63%
3	31%	38%	5%	19cm	29%	50%	21%
4	23%	51%	2%	15cm	27%	26%	47%

Tabelle 5.6.: Tracking-Kennzahlen (*false alarm error* (FA), *miss error* (Miss), *missmatch error* (MM), *mostly tracked* (MT), *partially tracked* (PT), *mostly lost* (ML)) für vier Testsequenzen.

5.3.5. Bewertung des Trackings

In diesem Abschnitt wird das Verfahren zur Objektverfolgung als Ganzes untersucht und bewertet. Hierfür wurden vier unterschiedliche Testsequenzen mit dem selben Parametersatz prozessiert. Die Qualität der Tracking-Ergebnisse ist in Tab. 5.6 dargestellt.

Der Anteil an Falschalarmen liegt insgesamt zwischen 4 % und 31 %, Schlupf bei 38 % bis 64 % und Fehlzweisungen bei 1 % bis 5 %. Der stark variierende Anteil an Falschalarmen hat diverse Ursachen. Er hängt vor allem von der Größe des betrachteten Bereichs und dessen Komplexität ab. So enthält Sequenz 3 im Vergleich zu Sequenz 1 besonders viele personenähnliche Objekte, welche Fehldetektionen hervorrufen. Auch die Vollständigkeit schwankt stark zwischen den Sequenzen. Sie hängt vor allem von der jeweiligen Erkennbarkeit der vorhandenen Personen ab. Der Anteil an Fehlzweisungen ist allgemein sehr niedrig. Nur in Sequenz 3, in welcher Menschen aus zwei unterschiedlichen Richtungen in einer Engstelle aneinander vorbei laufen, nimmt er deutlich zu. Die mittlere Lagegenauigkeit der Detektionen liegt unter 1,5 Pixel. In Sequenz 2, in welcher der fehlende Personenschatten die exakte Lokalisierung erleichtert, ist die Abweichung besonders gering. Der Anteil an Trajektorien, die zu 80 % und mehr erfasst wurden, liegt zwischen 27 % und 42 %, solche, die zu maximal 20 % erkannt wurden bei 21 % bis 63 %. Die besonders schwachen Ergebnisse in Sequenz 2 haben ihre Ursache im niedrigen Kontrast und der Tatsache, dass diese Sequenz nur 4 Bilder umfasst. Da bei schlechter Sichtbarkeit mehr konsistente Detektionen notwendig sind, damit eine Trajektorie als korrekt klassifiziert wird, wirkt sich die Kürze der Sequenz deutlich negativ auf die Ergebnisse aus.

In Abbildung 5.15 sind die Werte für Korrektheit und Vollständigkeit der erkannten Objektpositionen nach der Detektion und nach dem Tracking gemeinsam dargestellt. Hier zeigt sich, dass in einigen Sequenzen das Tracking zu einer Steigerung der Korrektheit und Vollständigkeit führt. Die Sequenzanalyse kann somit zusätzliche Informationen in den Detektionsprozess mit einbringen, die bei einer rein bildweisen Betrachtung nicht verfügbar sind. Die Grafik verdeutlicht zudem noch einen weiteren Aspekt. Obwohl der Detektionsschwellwert stückweise so weit abgesenkt wurde, bis sämtliche Beobachtungen zugelassen waren, so fallen die PR-Kurven der Tracking-Ergebnisse nicht so weit ab wie die der Detektion, sondern verbleiben auf hohem Niveau. Hier macht sich vor allem der Vorteil durch die stochastische Modellierung des Detektionswertes bemerkbar, wodurch schlechte Beobachtungen einen geringeren Einfluss erhalten (vgl. Abs. 5.3.3). Die Gegenüberstellung zeigt auch, dass die Detektion insgesamt den größten Einfluss auf die Ergebnisse hat. Das Tracking-Verfahren kann den Anteil an Falschalarmen und Schlupf nur in einem begrenzten Maße verringern. Für deutlich

5. Evaluierung

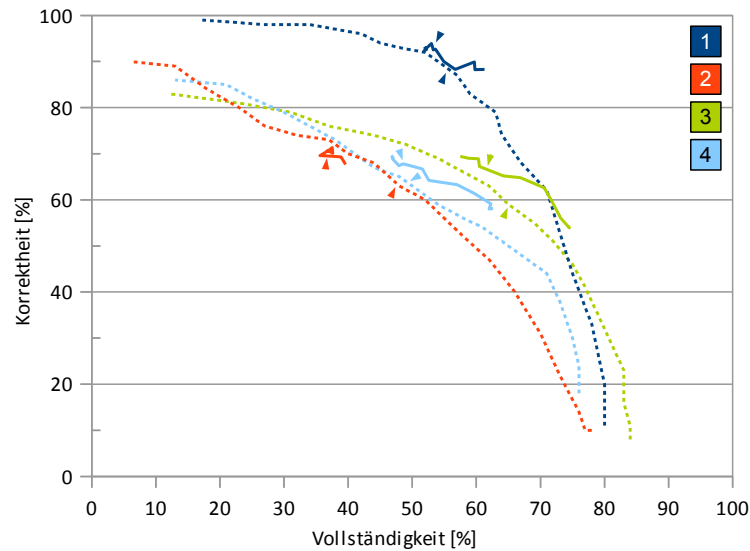


Abbildung 5.15.: Vergleich der Detektionsergebnisse nach der Detektion (gestrichelt) mit denen nach Abschluss des Trackings (durchgezogen) für die vier Sequenzen aus Tab. 5.6. Markiert sind ähnlich wie in Abb. 5.11 die Werte, welche mit einem moderaten Schwellwert erreicht werden.

bessere Ergebnisse muss vor allem die Leistungsfähigkeit der Objekterkennung gesteigert werden.

Eine genaue Betrachtung der gewonnenen Trajektorien offenbart einige Schwachstellen des Tracking-Verfahrens, welche in Abbildung 5.16 exemplarisch aufgeführt sind. Im linken Bild erkennt man, dass die Trajektorien einiger Personen quer zur allgemeinen Bewegungsrichtung verlaufen bzw. starke Sprünge aufweisen. Diese Phänomene treten vor allem am Beginn einer Sequenz und in Bereichen mit höher Objektdichte auf. Ihre Ursache liegt wahrscheinlich in der relativ einfachen Modellierung der Personenbewegung. Hier ließe sich eine Verbesserung erreichen, wenn die Bewegung nicht mit Null, sondern z. B. in Abhängigkeit der benachbarten Personen initialisiert werden würde. Zudem könnten weitere Regularisierungsmethoden, wie die Integration von mehr Vorwissen über das generell Bewegungsverhalten von Personen oder die Berücksichtigung der Laufwege benachbarter Personen, dabei helfen, die Prädiktion der Objektposition und die Zuordnung besonders in Bereichen höher Objektdichte zu verbessern. Im mittleren Bild in Abbildung 5.16 sind viele helle Flecken auf dem Weg zu erkennen. Obwohl die meisten unberücksichtigt bleiben, werden einige dennoch als Personen erkannt und durch die Sequenz verfolgt. Diese konsistenten Falschalarme treten häufig auf. Sie lassen sich nicht allein auf Basis der äußeren Erscheinung eliminieren. Notwendig ist die Integration von zusätzlichem Wissens über korrekte Detektionen und Trajektorien. Das rechte Bild zeigt ebenfalls ein häufig auftretendes Problem. Sobald die Objektdichte lokal zunimmt, sind Einzelpersonen nicht mehr sichtbar und bleiben unerkannt. Hält dieser Zustand für mehrere Bilder in Folge an, so kann auch das Tracking-Verfahren diesen Schlupf nicht mehr ausgleichen. Hier ist eine Erweiterung des Detektionsprozesses notwendig, um auch Personengruppen erkennen zu können.

Obwohl es gelungen ist ein System zu entwickeln, welches aus Luftbildsequenzen die Trajektorien von etwa einem Drittel aller Personen unter schwierigen Bedingungen rekonstruieren kann (s. Tab. 5.6), so verdeutlicht dieses Ergebnis auch, dass eine mikroskopische Verhaltensanalyse aktuell nur eingeschränkt durchgeführt werden kann. Es ist zu erwarten, dass die Umsetzung der aufgeführten Verbesserungsvorschläge zu einer Steigerung der Tracking-Ergebnisse führen wird. Es stellt sich jedoch die generelle Frage, ob sich bei der geringen Auflösung der Luftbilder und der schlechten Erkennbarkeit von Einzelpersonen, jemals Tra-

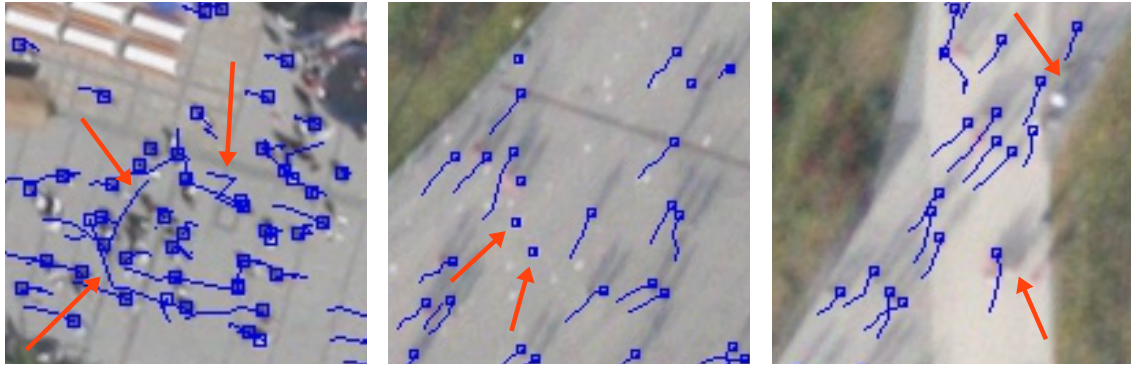


Abbildung 5.16.: Beispiele für beim Tracking auftretende Probleme. Dargestellt sind die aktuelle sowie die letzten vier Positionen jeder automatisch gewonnenen Trajektorie.

jektorien automatisch in ausreichender Qualität ermitteln lassen, um anschließend eine detaillierte, mikroskopische Verhaltensanalyse durchführen zu können. Andere Anwendungen, welche nicht auf die vollständige Extraktion der Trajektorien aller Einzelpersonen angewiesen sind, können jedoch bereits von den gewonnenen Daten profitieren. Anhand der Positionen lassen sich bspw. die Anzahl der Personen und deren räumliche Verteilung relativ genau schätzen (Butenuth u. a., 2011). Auch können die Trajektorien genutzt werden, um für bestimmte Bereiche Geschwindigkeitshistogramme oder Bewegungsfelder abzuleiten und so das lokale Bewegungsverhalten allgemein zu charakterisieren.

6. Zusammenfassung und Ausblick

In diesem abschließenden Kapitel wird zuerst die gesamte Arbeit mit ihren wesentlichen Beiträgen und Erkenntnissen zusammenfassend dargestellt. Anschließend werden Lösungsansätze für aktuelle Probleme benannt und interessante Bereiche für weiterführende Forschungsaktivitäten skizziert.

6.1. Zusammenfassung

Die vorliegende Dissertation hat sich erstmals umfassend mit der Fragestellung beschäftigt, in wie weit es möglich ist, einzelne Personen in Luftbildsequenzen automatisch zu erkennen und über die Zeit zu verfolgen. Die so gewonnenen Daten können die Grundlage bilden für weitergehende, groß- und kleinräumige Analysen zur Verteilung und zum Bewegungsverhalten von Personen. Diese Informationen lassen sich u. a. für eine bessere Koordination von Großveranstaltungen, eine Evaluation von Infrastrukturanlagen oder zum tieferen Verständnis des menschlichen Bewegungsverhaltens einsetzen.

Luftbildsequenzen werden bereits erfolgreich zur Verkehrsüberwachung eingesetzt. Bei der Detektion und Verfolgung von Einzelpersonen ergeben sich jedoch besondere Herausforderungen. Die Flughöhe von über 1000 m und die Bodenauflösung von 10 cm bis 20 cm pro Pixel führen z. B. dazu, dass Personen häufig kaum zu erkennen sind. Ein freistehender, von oben aufgenommener Mensch erscheint in den Bilddaten lediglich als kompakter, meist dunkler Fleck von 5 bis 10 Pixeln Größe und besitzt je nach Wetterlage einen Schatten. Bei der automatischen Analyse der Luftbilder können daher leicht Verwechslungen auftreten, zum einen mit menschenähnlichen Objekten, die besonders häufig in urbanen Gebieten vorkommen, und zum anderen zwischen verschiedenen, nah beieinander laufenden Personen.

Auf Basis einer detaillierten Literaturrecherche und eines Methodenvergleiches wurde eine Auswertestrategie konzipiert, welche besonders geeignet ist, mit den schwierigen Rahmenbedingungen und Herausforderungen umzugehen. Ihr Ziel ist es, möglichst vollständige Trajektorien aller sich in einem bestimmten Gebiet befindlichen Personen ohne manuelle Eingriffe bestimmen zu können. Um dies zu erreichen, wurden erstmals die beiden mächtigen Methoden der aussehensbasierten Detektion mit implizitem Objektmodell und des Multi-Hypothesen-Trackings (MHT) in einem Verfahren kombiniert. Hierfür war es notwendig, die Ergebnisse der Objekterkennung stochastisch zu beschreiben und den MHT-Formalismus entsprechend zu erweitern. Als Folge konnte der Einfluss des Detektionsschwellwertes stark abgeschwächt und die deutliche Trennung von Detektion und Tracking aufgehoben werden. Die Objekterkennung wurde so in den Bereich des Trackings verschoben. Hier stehen neben Informationen aus der Sequenzanalyse nun auch die Wahrscheinlichkeiten aller Detektionen zur Verfügung, was insgesamt zu deutlich besseren Ergebnissen führt.

Für die Suche nach potentiellen Objektpositionen im Luftbild wurde ein aussehensbasierter Ansatz mit implizitem Modell gewählt und an die gegebenen Anforderungen adaptiert. Um die äußere Erscheinung der Personen besonders gut beschreiben zu können, wurden geeignete Bildmerkmale entwickelt. Zum einen wurde die Bildungsvorschrift für Haar-Merkmale verallgemeinert, was die Definition neuer und individuell an das Objekt angepasster Formmerkmale erlaubt. Des Weiteren wurden Rechteckmerkmale für Farbwert und Varianz entworfen, um zur Form komplementäre Informationen verfügbar zu haben und so die Detektionsergebnisse

6. Zusammenfassung und Ausblick

zu verbessern. Eine weitere Steigerung konnte erreicht werden, indem der Personenschatten mit in das visuelle Objektmodell integriert wurde. Hierfür muss das gesamte Bild im Vorfeld so gedreht werden, dass der Schatten immer in die gleiche Richtung zeigt. Zusätzlich ist auch der Auswertebereich des Detektors so zu erweitern, dass er neben der Person auch einen Teil ihres Schattens mit umfasst. Da das implizite Objektmodell einem Lernverfahren anhand von vielen Beispielen beigebracht werden muss, spielen diese für eine erfolgreiche Detektion eine entscheidende Rolle. In dieser Arbeit wurde das übliche, unüberwachte Verfahren zum Sammeln von Hintergrundbeispielen auf zwei Arten verbessert. Zum einen wurden Trainingsbilder mit ausmaskierten Personen verwendet, damit auch Negativbeispiele aus dem nahen Umfeld gesammelt werden können und dort später weniger Fehldetektionen auftreten. Zum anderen diente der Konfidenzwert jedes potentiellen Beispiels als Auswahlkriterium, um zu verhindern, dass personenähnliche Objekte der Hintergrundklasse zugeordnet werden.

Personen werden in dieser Arbeit mittels eines Tracking-by-Detection-Ansatzes verfolgt. Hierfür wurde das besonders leistungsfähige MHT-Verfahren ausgewählt. Dieses verfolgt gleichzeitig mehrere alternative Erklärungen für den Ursprung sämtlicher Detektionen und möglicher Objektbewegungen, bis sich im Laufe der Zeit die wahrscheinlichste von ihnen durchsetzt. Da sich auf diese Weise schwierige Entscheidungen in mehrdeutigen Situationen bis zu ihrer Klärung aufschieben lassen, besitzen die gewonnenen Trajektorien eine höhere Qualität als wenn nur eine einzige Hypothese verfolgt werden würde. Das MHT-Verfahren wurde in dieser Arbeit zudem in mehrfacher Weise verbessert. Zum einen wurden die Vorteile der hypothesen- und trajektorienorientierten MHT-Variante vereint. Nun lassen sich sowohl die Anzahl der verfolgten Hypothesen präzise steuern, als auch die Trajektorien mit Hilfe des Quotiententests bewerten. Des Weiteren wurde ein Verfahren entwickelt, welches die Auftrittswahrscheinlichkeit von Falschalarmen und neuen Objekten adaptiv und automatisch bestimmen kann und dabei alle aktuellen Hypothesen berücksichtigt. Dadurch entfällt die Notwendigkeit, diese Wahrscheinlichkeitsverteilungen manuell festzulegen. Darüber hinaus konnte auch die Umsetzung des besonders wichtigen Clusterverfahrens zur Erzeugung von leichter zu lösenden Teilproblemen durch eine verbesserte Datenstruktur enorm vereinfacht werden.

Obwohl sämtliche Entwicklungen in dieser Arbeit mit der Absicht erfolgt sind, das Erkennen und Verfolgen von Personen in Luftbildsequenzen zu optimieren, sind sie in der Regel unabhängig von diesem Anwendungsfall. Von den verbesserten Methoden und Abläufen können daher leicht auch andere Bereiche profitieren, in denen ähnliche Aufgabenstellungen zu bewältigen sind.

Die Zielsetzung der Arbeit, mittels Luftbildsequenzen die Trajektorien aller abgebildeten Personen möglichst vollständig zu ermitteln, ließ sich nur im begrenzten Maße erfüllen. In anspruchsvollen, realistischen Szenarien konnten zwischen 25% und 40% aller Personen zum größten Teil korrekt verfolgt werden, eine deutliche Mehrheit jedoch nur teilweise oder kaum. Besonders gute Ergebnisse ließen sich in Situationen mit niedriger Personendichte erzielen, da hier die einzelnen Personen deutlich erkennbar sind. Es kommt jedoch relativ häufig vor, dass Personen sich in Gruppen oder in großer Nähe zu weiteren Personen aufhalten. In diesen Fällen sind Individuen visuell kaum mehr zu unterscheiden und der aussehensbasierte Ansatz zur Objekterkennung stößt an seine Grenzen. Die geringe Vollständigkeit der Detektion ist daher auch die Hauptursache dafür, dass sich nur ein Teil aller Trajektorien weitestgehend vollständig bestimmen lässt. Als Folge können die gewonnenen Daten aktuell nur begrenzt für detaillierte, personenbezogene Verhaltensanalysen eingesetzt werden. Bereits möglich sind jedoch makroskopische Auswertungen zum allgemeinen Bewegungsverhalten und der Verteilung von Personen in einem bestimmten Gebiet.

6.2. Ausblick

In diesem Abschnitt werden Ansatzpunkte für zukünftige Forschungsaktivitäten aufgezeigt, in welchen die Grenzen der verwendeten Verfahren überwunden oder weitergehende Fragestellungen ergründet werden können.

Eine wichtige Erkenntnis aus dieser Arbeit ist das Wissen über die Grenzen der mikroskopischen, objektbezogenen Bildauswertung. Obwohl es im Detail noch Möglichkeiten zur Verbesserung gibt, wird sich allein auf diese Weise nie das Bewegungsverhalten sämtlicher Personen in einem Luftbild erfassen lassen. Notwendig ist daher eine umfassendere Betrachtung des Problems, ähnlich wie es in (Hinz u. a., 2008) für die Verkehrsbeobachtung mittels Luftbildern geschehen ist. Das aktuelle System müsste um weitere Module zur meso- und makroskopischen Bildauswertung ergänzt werden, damit auch Phänomene wie Personengruppen und Menschenmassen behandelt werden können, die ebenfalls im Luftbild beobachtet werden können. Die Verknüpfung der Module untereinander könnte zudem durch gegenseitige Kontrolle und gleichzeitige Nutzung komplementärer Ansätze zu einer allgemeinen Steigerung der Ergebnisqualität führen. Zu beiden Bereichen gibt es bereits etliche Veröffentlichungen, deren Ergebnisse auf den Luftbildfall übertragen werden müssten. In Arbeiten zur makroskopischen Bildauswertung liegt der Schwerpunkt auf der Beobachtung von Menschenmassen (*crowd analysis*). Meist wird die Anzahl, die Dichte oder die generelle Bewegungsrichtung der Personen in einem bestimmten Gebiet ermittelt, ohne einzelne Personen explizit zu erkennen (Zhan u. a., 2008; Hinz, 2009; Jacques Junior u. a., 2010). Besonders interessant sind Verfahren, in denen solche Analysen als Vorwissen zum Erkennen und Verfolgen von Einzelpersonen genutzt werden (Rodriguez u. a., 2009, 2011). Auch zu Personengruppen gibt es schon einige Veröffentlichungen (Gennari und Hager, 2004; Henriques u. a., 2011). Hier stehen besonders die korrekte Modellierung von Split- und Merge-Situationen im Vordergrund sowie die entsprechende Erweiterung bestehender Ansätze wie dem MHT-Verfahren (Mucientes und Burgard, 2006; Joo und Chellappa, 2007; Lau u. a., 2010).

Neben der Erweiterung des Systems um neue Module für Personengruppen und Menschenmassen, können auch die bereits vorhandenen Methoden weiter optimiert werden. Vor allem eine Steigerung der Detektionsleistung ließe deutlich verbesserte Ergebnisse erwarten (vgl. Daum und Fitzgerald (1994)). Neben neuartigen Ansätzen für besonders niedrig aufgelöste Objekte (Jiang u. a., 2012), scheint vor allem die Hinzunahme von Kontextinformationen (Heitz und Koller, 2008; Perko und Leonardis, 2010) eine vielversprechende Möglichkeit, um den Einschränkungen durch die besonders schwache Signatur von Personen in Luftbildern entgegenzuwirken. Des Weiteren wäre es auch interessant zu untersuchen, ob Veränderungen des Klassifikatortrainings, z. B. durch Online-Learning-Methoden (Grabner u. a., 2008), eine Verbesserung bringen würden. In diesem Zusammenhang ließe sich eventuell auch die robuste Gewinnung von Hintergrundbeispielen weiterentwickeln. So wäre es denkbar, statt des Konfidenzwertes eine adaptiv berechnete Wahrscheinlichkeit als Auswahlkriterium zu nutzen. Das iterative Training könnte dann automatisch abgebrochen werden, wenn nur noch Beispiele über einer bestimmten Objektwahrscheinlichkeit gefunden werden würden.

Zur Steigerung der Tracking-Ergebnisse scheinen zwei Ansätze am erfolgversprechendsten. Zum einen könnte auch hier versucht werden, durch Hinzunahme von Kontextinformationen, z. B. über benachbarte Personen oder markante Objekte, das Korrespondenzproblem weiter zu vereinfachen (Reilly u. a., 2010a; Xiao u. a., 2010; Dinh u. a., 2011). Hierbei wäre zu beachten, dass sich diese Informationen möglichst ebenfalls direkt aus den Bilddaten in einfacher und robuster Weise ermitteln lassen sollten. Zum anderen könnte das aktuell verwendete Bewegungsmodell durch ein komplexeres ersetzt werden, welches das Verhalten der Personen realistischer und mit Bezug zu ihren jeweiligen Nachbarn (Pellegrini u. a., 2009; Li u. a., 2008; Robin u. a., 2009) oder relativ zu festen Bildpunkten beschreibt. Hiervon würde auch die stochastische Bewertungsfunktion profitieren, auf welcher das MHT-Verfahren basiert. Die Wahrscheinlichkeitsverteilung wiedererkannter Objekte ließe sich durch ein fortschrittli-

6. Zusammenfassung und Ausblick

chere Bewegungsmodell besser als bisher beschreiben. Eine andere Schwachstelle der Bewertungsfunktion stellt die konstante Detektionswahrscheinlichkeit dar. Realistischer ist es, dass sich diese aufgrund vielfältiger Einflussfaktoren ständig verändert. Hier wäre zu untersuchen, welche Faktoren dies genau sind und ob sie sich robust ermitteln und stochastisch beschreiben lassen.

Literaturverzeichnis

- Anton Andriyenko und Konrad Schindler. Globally optimal multi-target tracking on a hexagonal lattice. In *Computer Vision – ECCV 2010*, Bd. 6311 aus *LNCS*, S. 466–479, Springer, 2010. DOI 10.1007/978-3-642-15549-9_34.
- G. Antonini, S. V. Martinez, M. Bierlaire, und J. P. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69(2):159–180, 2006. DOI 10.1007/s11263-005-4797-0.
- David M. Antunes, David Martins de Matos, und José Gaspar. A library for implementing the multiple hypothesis tracking algorithm. *ArXiv e-prints*, S. 1–13, 2011.
- Y. Bar-Shalom, S.S. Blackman, und R.J. Fitzgerald. Dimensionless score function for multiple hypothesis tracking. *Aerospace and Electronic Systems, IEEE Trans. on*, 43(1):392–400, 2007. DOI 10.1109/TAES.2007.357141.
- Y. Bar-Shalom, F. Daum, und J. Huang. The probabilistic data association filter. *Control Systems Magazine, IEEE*, 29(6):82–100, 2009. DOI 10.1109/MCS.2009.934469.
- Yaakov Bar-Shalom und William Dale Blair, editors. *Multitarget-Multisensor Tracking: Applications and Advances*, Bd. 3 aus *Artech House radar library*. Artech House, 2000. ISBN 978-1580530910.
- Yaakov Bar-Shalom und Thomas E. Fortmann. *Tracking and Data Association*. Mathematics in Science and Engineering. Academic Press, 1988. ISBN 978-0-12-079760-7.
- Axel Baumann, Marco Boltz, Julia Ebling, Matthias Koenig, Hartmut S. Loos, Marcel Merkel, Wolfgang Niem, Jan Karl Warzelhan, und Jie Yu. A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, 2008:1–30, 2008. DOI 10.1155/2008/824726.
- C. Benedek, T. Sziranyi, Z. Kato, und J. Zerubia. Detection of object motion regions in aerial image pairs with a multilayer markovian model. *Image Processing, IEEE Trans. on*, 18(10):2303–2315, 2009. DOI 10.1109/TIP.2009.2025808.
- J. Berclaz, F. Fleuret, E. Turetken, und P. Fua. Multiple object tracking using k-shortest paths optimization. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 33(9):1806–1819, 2011. DOI 10.1109/TPAMI.2011.21.
- Keni Bernardin und Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. DOI 10.1155/2008/246309.
- Irving Biederman, Robert J. Mezzanotte, und Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, 1982. DOI 10.1016/0010-0285(82)90007-X.
- Samuel Blackman und Robert Popoli. *Design and analysis of modern tracking systems*. Artech House, 1999. ISBN 978-1-58053-006-4.
- Samuel S. Blackman. *Multiple-target tracking with radar applications*. Artech House, 1986. ISBN 978-0890061794.
- Samuel S. Blackman. Multiple hypothesis tracking for multiple target tracking. *Aerospace and Electronic Systems Magazine, IEEE*, 19(1):5–18, 2004. DOI 10.1109/MAES.2004.1263228.
- H. A. P. Blom. An efficient filter for abruptly changing systems. In *Decision and Control, IEEE Conf. on*, Bd. 23, S. 656–658, 1984. DOI 10.1109/CDC.1984.272089.

Literaturverzeichnis

- Avrim L. Blum und Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997. DOI 10.1016/S0004–3702(97)00063–5.
- M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, und L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 33(9):1820–1833, 2011. DOI 10.1109/TPAMI.2010.232.
- Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. DOI 10.1023/A:1009715923555.
- Florian Burkert, Florian Schmidt, Matthias Butenuth, und Stefan Hinz. People tracking and trajectory interpretation in aerial image sequences. In *Photogrammetric Computer Vision and Image Analysis*, Bd. XXXVIII Part 3A aus *IAPRS*, S. 209–214, ISPRS Commission III, 2010.
- Matthias Butenuth, Florian Burkert, Angelika Kneidl, André Borrmann, Florian Schmidt, Stefan Hinz, Beril Sirmacek, und Dirk Hartmann. Integrating pedestrian simulation, tracking and event detection for crowd analysis. In *First IEEE ICCV Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, S. 150–157, 2011. DOI 10.1109/ICCVW.2011.6130237.
- H. Cheng und D. Butler. Segmentation of aerial surveillance video using a mixture of experts. In *Digital Image Computing: Techniques and Applications*, S. 454–461, 2005. DOI 10.1109/DICTA.2005.73.
- J.B. Collins und J.K. Uhlmann. Efficient gating in data association with multivariate gaussian distributed states. *Aerospace and Electronic Systems, IEEE Trans. on*, 28(3):909–916, 1992. DOI 10.1109/7.256316.
- D. Comaniciu und P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 24(5):603–619, 2002. DOI 10.1109/34.1000236.
- D. Comaniciu, V. Ramesh, und P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 25(5):564–577, 2003. DOI 10.1109/TPAMI.2003.1195991.
- I.J. Cox und S.L. Hingorani. An efficient implementation of Reid’s multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 18(2):138–150, 1996. DOI 10.1109/34.481539.
- Ingemar J. Cox. A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision*, 10(1):53–66, 1993. DOI 10.1007/BF01440847.
- Franklin C. Crow. Summed-area tables for texture mapping. In *Computer Graphics and Interactive Techniques, Conf. on*, S. 207–212, ACM, 1984. DOI 10.1145/800031.808600.
- N. Dalal und B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, Bd. 1, S. 886–893, 2005. DOI 10.1109/CVPR.2005.177.
- Navneet Dalal, Bill Triggs, und Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, Bd. 3952 aus *LNCS*, S. 428–441, Springer, 2006. DOI 10.1007/11744047_33.
- Frederick E. Daum und Robert J. Fitzgerald. Importance of resolution in multiple-target tracking. In *Signal and Data Processing of Small Targets, Conf. on*, Bd. 2235, S. 329–338, 1994. DOI 10.1117/12.179063.
- Tinne De Laet. *Rigorously Bayesian Multitarget Tracking and Localization*. Dissertation, Katholieke Universiteit Leuven, Belgium, 2010.
- G. C. Demos, R. A. Ribas, T. J. Broida, und S. S. Blackman. Applications of MHT to dim moving targets. In *Signal and Data Processing of Small Targets, Conf. on*, Bd. 1305, S. 297–309, SPIE, 1990. DOI 10.1117/12.21598.
- Thang Ba Dinh, Nam Vo, und G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 1177–1184, 2011. DOI 10.1109/CVPR.2011.5995733.

- Piotr Dollar, Christian Wojek, Bernt Schiele, und Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 34(4):743–761, 2012. DOI 10.1109/TPAMI.2011.155.
- Richard O. Duda, Peter E. Hart, und David G. Stork. *Pattern classification*. Wiley, 2001. ISBN 978-0-471-05669-0.
- Line Eikvil, Lars Aurdal, und Hans Koren. Classification-based vehicle detection in high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(1):65–72, 2009. DOI 10.1016/j.isprsjprs.2008.09.005.
- Yoav Freund und Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. DOI 10.1006/jcss.1997.1504.
- Jerome Friedman, Trevor Hastie, und Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000.
- Carolina Galleguillos und Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010. DOI 10.1016/j.cviu.2010.02.004.
- W. Ge und R. T. Collins. Marked point processes for crowd counting. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 2913–2920, 2009. DOI 10.1109/CVPRW.2009.5206621.
- G. Gennari und G.D. Hager. Probabilistic data association methods in visual tracking of groups. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, Bd. 2, S. 876–881, 2004. DOI 10.1109/CVPR.2004.1315257.
- David Gerónimo, Antonio M. López, Angel D. Sappa, und Thorsten Graf. Survey of pedestrian detection for advanced driver assistance systems. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 32(7):1239–1258, 2010. DOI 10.1109/TPAMI.2009.122.
- J.-M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, und H. Geerts. Color invariance. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 23(12):1338–1350, 2001. DOI 10.1109/34.977559.
- H. Grabner, T. T. Nguyen, B. Gruber, und H. Bischof. On-line boosting-based car detection from aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(3):382–396, 2008. DOI 10.1016/j.isprsjprs.2007.10.005.
- G. Heitz und D. Koller. Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision*, S. 30–43, 2008. DOI 10.1007/978-3-540-88682-2_4.
- Dirk Helbing und Péter Molnár. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51(5):4282–4286, 1995. DOI 10.1103/PhysRevE.51.4282.
- J.F. Henriques, R. Caseiro, und J. Batista. Globally optimal solution to multi-object tracking with merged measurements. In *Computer Vision, IEEE Int. Conf. on*, S. 2470–2477, 2011. DOI 10.1109/ICCV.2011.6126532.
- S. Hinz. Detection and counting of cars in aerial images. In *Image Processing, IEEE Int. Conf. on*, Bd. 3, S. 997–1000, 2003. DOI 10.1109/ICIP.2003.1247415.
- S. Hinz. Integrating local and global features for vehicle detection in high resolution aerial imagery. In *Photogrammetric Image Analysis*, Bd. XXXIV, Part 3/W8 aus *IAPRS*, S. 119–124, 2005.
- S. Hinz, D. Lenhart, und J. Leitloff. Detection and tracking of vehicles in low frame rate aerial image sequences. In *Proceedings of Workshop on High-Resolution Earth Imaging for Geospatial Information*, S. 1–6, 2007.
- Stefan Hinz. Density and motion estimation of people in crowded environments based on aerial image sequences. In *ISPRS Hannover Workshop 2009: High-Resolution Earth Imaging for Geospatial Information*, Bd. XXXVIII-1-4-7/W5 aus *IAPRS*, S. 1–6, ISPRS, 2009.

Literaturverzeichnis

- Stefan Hinz, Dominik Lenhart, und Jens Leitloff. Traffic extraction and characterisation from optical remote sensing data. *The Photogrammetric Record*, 23(124):424–440, 2008. DOI 10.1111/j.1477-9730.2008.00497.x.
- S.P. Hoogendoorn, H.J. van Zuylen, M. Schreuder, B. Gorte, und G. Vosselman. Microscopic traffic data collection by remote sensing. *Transportation Research Record*, 1855:121–128, 2003. DOI 10.3141/1855-15.
- Weiming Hu, Tieniu Tan, Liang Wang, und Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004. DOI 10.1109/TSMCC.2004.829274.
- Chang Huang, Bo Wu, und Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, Bd. 5303 aus LNCS, S. 788–801, Springer, 2008. DOI 10.1007/978-3-540-88688-4_58.
- J.C.S. Jacques Junior, S.R. Musse, und C.R. Jung. Crowd analysis using computer vision techniques. *Signal Processing Magazine, IEEE*, 27(5):66–77, 2010. DOI 10.1109/MSP.2010.937394.
- A. K. Jain, R. P. W. Duin, und J. Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 22(1):4–37, 2000. DOI 10.1109/34.824819.
- Khuloud Jaqaman, Dinah Loerke, Marcel Mettlen, Hirotaka Kuwata, Sergio Grinstein, Sandra L Schmid, und Gaudenz Danuser. Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, 5(8):695–702, 2008. DOI 10.1038/nmeth.1237.
- Bernd Jähne. *Digitale Bildverarbeitung*. Springer, Berlin, 2005. ISBN 978-3-540-24999-3.
- N. Jiang, H. Su, W. Liu, und Y. Wu. Discriminative metric preservation for tracking low-resolution targets. *Image Processing, IEEE Trans. on*, 21(3):1284–1297, 2012. DOI 10.1109/TIP.2011.2167345.
- Shan Jiang, Xiaobo Zhou, Tom Kirchhausen, und Stephen T. C. Wong. Detection of molecular particles in live cells via machine learning. *Cytometry Part A*, 71A(8):563–575, 2007. DOI 10.1002/cyto.a.20404.
- Kai Jüngling. *Ein generisches System zur automatischen Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten*. Dissertation, Institut für Photogrammetrie und Fernerkundung, Karlsruher Institut für Technologie (KIT), 2011. URN urn:nbn:de:swb:90-223579.
- P.-M. Jodoin, M. Mignotte, und C. Rosenberger. Segmentation framework based on label field fusion. *Image Processing, IEEE Trans. on*, 16(10):2535–2550, 2007. DOI 10.1109/TIP.2007.903841.
- Seong-Wook Joo und Rama Chellappa. A multiple-hypothesis approach for multiobject visual tracking. *Image Processing, IEEE Trans. on*, 16(11):2849–2854, 2007. DOI 10.1109/TIP.2007.906254.
- Zdenek Kalal, Krystian Mikolajczyk, und Jiri Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 34(7):1409–1422, 2012. DOI 10.1109/TPAMI.2011.239.
- R. Kaucic, A.G. Amitha Perera, G. Brooksby, J. Kaufhold, und A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, Bd. 1, S. 990–997, 2005. DOI 10.1109/CVPR.2005.53.
- H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. DOI 10.1002/nav.3800020109.
- R. Kumar, H. Sawhney, S. Samarasekera, S. Hsu, Hai Tao, Yanlin Guo, K. Hanna, A. Pope, R. Wildes, D. Hirvonen, M. Hansen, und P. Burt. Aerial video surveillance and exploitation. *Proceedings of the IEEE*, 89(10):1518–1539, 2001. DOI 10.1109/5.959344.
- Thomas Kurien. *Multitarget-Multisensor Tracking: Advanced Applications*, Kapitel: Issues in the design of practical multitarget tracking algorithms, S. 43–83. Artech House Radar Library. Artech House, 1990. ISBN 0-89006-377-X.
- Franz Kurz, Rupert Müller, Manfred Stephani, Peter Reinartz, und Manfred Schroeder. Calibration of a wide-angle digital camera system for near real time scenarios. In *High Resolution Earth Imaging for Geospatial Information, ISPRS Workshop*, S. 1–6, ISPRS, 2007.

- F. Lafarge, X. Descombes, J. Zerubia, und M. Pierrot-Deseilligny. Automatic building extraction from dems using an object approach and application to the 3d-city modeling. *Journal of Photogrammetry and Remote Sensing*, 63(3):365–381, 2008.
- Boris Lau, Kai Arras, und Wolfram Burgard. Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics*, 2:19–30, 2010. DOI 10.1007/s12369-009-0036-0.
- Bastian Leibe, Ale Leonardis, und Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008. DOI 10.1007/s11263-007-0095-3.
- Leica Geosystems. Leica RCD30 series datasheet, 2012. URL http://www.leica-geosystems.com/downloads123/zz/airborne/RCD30/brochures-datasheet/Leica_RCD30_DS_en.pdf.
- J. Leitloff, S. Hinz, und U. Stilla. Vehicle detection in very high resolution satellite images of city areas. *Geoscience and Remote Sensing, IEEE Trans. on*, 48(7):2795–2806, 2010. DOI 10.1109/TGRS.2010.2043109.
- K. Levi und Y. Weiss. Learning object detection from a small number of examples: the importance of good features. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, Bd. 2, S. 53–60, 2004. DOI 10.1109/CVPR.2004.1315144.
- Kang Li und Takeo Kanade. Nonnegativ mixed-norm preconditioning for microscopy image segmentation. In *Information Processing in Medical Imaging*, Bd. 5636 aus LNCS, S. 362–373. Springer, 2009. DOI 10.1007/978-3-642-02498-6_30.
- Kang Li, Eric D. Miller, Mei Chen, Takeo Kanade, Lee E. Weiss, und Phil G. Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical Image Analysis*, 12(5):546–566, 2008. DOI 10.1016/j.media.2008.06.001.
- Yuan Li, Chang Huang, und R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 2953–2960, 2009. DOI 10.1109/CVPR.2009.5206735.
- R. Lienhart und J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing, IEEE Int. Conf. on*, Bd. 1, S. 900–903, 2002. DOI 10.1109/ICIP.2002.1038171.
- Yuping Lin, Qian Yu, und Gérard Medioni. Efficient detection and tracking of moving objects in geo-coordinates. *Machine Vision and Applications*, 22:505–520, 2011. DOI 10.1007/s00138-010-0264-1.
- T. List und R.B. Fisher. Cvml - an xml-based computer vision markup language. In *Pattern Recognition, Int. Conf. on*, Bd. 1, S. 789–792, 2004. DOI 10.1109/ICPR.2004.1334335.
- Zhiming Liu und Chengjun Liu. Fusion of color, local spatial and global frequency information for face recognition. *Pattern Recognition*, 43(8):2882–2890, 2010. DOI 10.1016/j.patcog.2010.03.003.
- Mahendra Mallick und Barbara La Scala. Comparison of single-point and two-point difference track initiation algorithms using position measurements. *Acta Automatica Sinica*, 34(3):258–265, 2008. DOI 10.3724/SPJ.1004.2008.00258.
- Alexandre Matov, Marcus M. Edvall, Ge Yang, und Gaudenz Danuser. Optimal-flow minimum-cost correspondence assignment in particle flow tracking. *Computer Vision and Image Understanding*, 115(4):531–540, 2011. DOI 10.1016/j.cviu.2011.01.001.
- G. Medioni, I. Cohen, F. Bremond, S. Hongeng, und R. Nevatia. Event detection and analysis from video streams. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 23(8):873–889, 2001. DOI 10.1109/34.946990.
- Erik Meijering, Oleh Dzyubachyk, Ihor Smal, und Wiggert A. van Cappellen. Tracking in cell and developmental biology. *Seminars in Cell & Developmental Biology*, 20(8):894–902, 2009. DOI 10.1016/j.semcd.2009.07.004.

Literaturverzeichnis

- Microsoft. UltraCam-Xp technical specification, 2011. URL <http://download.microsoft.com/download/7/4/3/743EFD09-258B-4BFA-8D56-3148C60DD137/UCAMTechnicalDocuments/UltraCamXp-Specs.pdf>.
- Andrew Miller, Pavel Babenko, Min Hu, und Mubarak Shah. Person tracking in uav video. In *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007*, Bd. 4625 aus LNCS, S. 215–220, Springer, 2008. DOI 10.1007/978-3-540-68585-2_19.
- M.L. Miller, H.S. Stone, und I.J. Cox. Optimizing Murty's ranked assignment method. *Aerospace and Electronic Systems, IEEE Trans. on*, 33(3):851–862, 1997. DOI 10.1109/7.599256.
- Manuel Mucientes und Wolfram Burgard. Multiple hypothesis tracking of clusters of people. In *Intelligent Robots and Systems, IEEE Int. Conf. on*, S. 692–697, 2006. DOI 10.1109/IROS.2006.282614.
- Katta G. Murty. An algorithm for ranking all the assignments in order of increasing cost. *Operations Research*, 16(3):682–687, 1968.
- Fatemeh Karimi Nejadasl, Ben G.H. Gorte, und Serge P. Hoogendoorn. Optical flow based vehicle tracking strengthened by statistical decisions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 61(3-4):159–169, 2006. DOI 10.1016/j.isprsjprs.2006.09.007.
- Bernd Neumann. *Handbuch der künstlichen Intelligenz*, Kapitel: Bildverstehen - Ein Überblick, S. 815–841. Oldenbourg, 2003. ISBN 3-486-27212-8.
- Alexandru Niculescu-Mizil und Rich Caruana. Obtaining calibrated probabilities from boosting. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, S. 413–420, AUA Press, 2005.
- Wolfgang Niemeier. *Ausgleichsrechnung*. de Gruyter, 2008. ISBN 978-3110190557.
- Yu-Ichi Ohta, Takeo Kanade, und Toshiyuki Sakai. Color information for region segmentation. *Computer Graphics and Image Processing*, 13(3):222–241, 1980. DOI 10.1016/0146-664X(80)90047-7.
- T. Ojala, M. Pietikainen, und T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 24(7):971–987, 2002. DOI 10.1109/TPAMI.2002.1017623.
- O. Oreifej, R. Mehran, und M. Shah. Human identity recognition in aerial images. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 709–716, 2010. DOI 10.1109/CVPR.2010.5540147.
- M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, und T. Poggio. Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 193–199, 1997. DOI 10.1109/CVPR.1997.609319.
- Donovan H. Parks und Martin D. Levine. Is local colour normalization good enough for local appearance-based classification? *Machine Vision and Applications*, 21(5):789–796, 2010. DOI 10.1007/s00138-009-0186-y.
- Stefano Pellegrini, Andreas Ess, Konrad Schindler, und Luc van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, IEEE Int. Conf. on*, S. 261–268, 2009. DOI 10.1109/ICCV.2009.5459260.
- David W. Pentico. Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, 176(2):774–793, 2007. DOI 10.1016/j.ejor.2005.09.014.
- Roland Perko und Ales Leonardis. A framework for visual-context-aware object detection in still images. *Computer Vision and Image Understanding*, 114(6):700–711, 2010. DOI 10.1016/j.cviu.2010.03.005.
- John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, S. 61–74, MIT Press, 1999.
- R.L. Popp, K.R. Pattipati, und Y. Bar-Shalom. m-best s-d assignment algorithm with application to multitarget tracking. *Aerospace and Electronic Systems, IEEE Trans. on*, 37(1):22–39, 2001. DOI 10.1109/7.913665.

- G.W. Pulford. Taxonomy of multiple target tracking methods. In *Radar, Sonar and Navigation, IEE Proc.*, Bd. 152, S. 291–304, 2005. DOI 10.1049/ip-rsn:20045064.
- Ibrahim Reda und Afshin Andreashinz. Solar position algorithm for solar radiation applications. *Solar Energy*, 76(5):577–589, 2004. DOI 10.1016/j.solener.2003.12.003.
- D. Reid. An algorithm for tracking multiple targets. *Automatic Control, IEEE Trans. on*, 24(6):843–854, 1979. DOI 10.1109/TAC.1979.1102177.
- Vladimir Reilly, Haroon Idrees, und Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In *Computer Vision – ECCV 2010*, Bd. 6313 aus LNCS, S. 186–199, Springer, 2010a. DOI 10.1007/978-3-642-15558-1_14.
- Vladimir Reilly, Berkan Solmaz, und Mubarak Shah. Geometric constraints for human detection in aerial imagery. In *Computer Vision – ECCV 2010*, Bd. 6316 aus LNCS, S. 252–265, Springer, 2010b. DOI 10.1007/978-3-642-15567-3_19.
- Branko Ristic, Sanjeev Arulampalam, und Neil Gordon. *Beyond the Kalman filter : particle filters for tracking applications*. Artech House radar library. Artech House, 2004. ISBN 978-1-58053-631-8.
- Th. Robin, G. Antonini, M. Bierlaire, und J. Cruz. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1):36–56, 2009. DOI 10.1016/j.trb.2008.06.010.
- Mikel Rodriguez, Saad Ali, und Takeo Kanade. Tracking in unstructured crowded scenes. In *Computer Vision, IEEE Int. Conf. on*, S. 1389–1396, 2009. DOI 10.1109/ICCV.2009.5459301.
- Mikel Rodriguez, Ivan Laptev, Josef Sivic, und Jean-Yves Audibert. Density-aware person detection and tracking in crowds. In *Computer Vision, IEEE Int. Conf. on*, S. 2423–2430, IEEE, 2011. DOI 10.1109/ICCV.2011.6126526.
- Carsten Rother, Vladimir Kolmogorov, und Andrew Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. DOI 10.1145/1015706.1015720.
- Jean Roy, Nicolas Duclos-Hindie, und Dany Dessureault. Efficient cluster management algorithm for multiple-hypothesis tracking. In *Signal and Data Processing of Small Targets, Conf. on*, Bd. 3163, S. 301–313, SPIE, 1997. DOI 10.1117/12.279526.
- Florian Schmidt und Stefan Hinz. A scheme for the detection and tracking of people tuned for aerial image sequences. In *Photogrammetric Image Analysis*, number 6952 in LNCS, S. 257–270, Springer, Heidelberg, 2011. DOI 10.1007/978-3-642-24393-6_22.
- K. Shafique und M. Shah. A noniterative greedy algorithm for multiframe point correspondence. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 27(1):51–65, 2005. DOI 10.1109/TPAMI.2005.1.
- K. Shafique, Mun Wai Lee, und N. Haering. A rank constrained continuous formulation of multi-frame multi-target tracking problem. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 1–8, 2008. DOI 10.1109/CVPR.2008.4587577.
- G. Sharma, C. J. Merry, P. Goel, und M. McCord. Vehicle detection in 1-m resolution satellite and airborne imagery. *International Journal of Remote Sensing*, 27(4):779–797, 2006. DOI 10.1080/01431160500238901.
- Beril Sirmacek und Peter Reinartz. Kalman filter based feature analysis for tracking people from airborne images. In *ISPRS Hannover Workshop 2011: High-Resolution Earth Imaging for Geospatial Information*, Bd. XXXVIII-4/W19 aus IAPRS, S. 1–6, 2011.
- Robert W. Sittler. An optimal data association problem in surveillance theory. *Military Electronics, IEEE Trans. on*, 8(2):125–139, 1964. DOI 10.1109/TME.1964.4323129.
- I. Smal, M. Loog, W. Niessen, und E. Meijering. Quantitative comparison of spot detection methods in fluorescence microscopy. *Medical Imaging, IEEE Trans. on*, 29(2):282–301, 2010. DOI 10.1109/TMI.2009.2025127.

Literaturverzeichnis

- G. Keith Still. *Crowd Dynamics*. Dissertation, Department of Mathematics, University of Warwick (UK), 2000.
- Paul Suetens, Pascal Fua, und Andrew J. Hanson. Computational strategies for object recognition. *ACM Comput. Surv.*, 24:5–62, 1992. DOI 10.1145/128762.128763.
- K.-K. Sung und T. Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 20(1):39–51, 1998. DOI 10.1109/34.655648.
- I. Szotzka und M. Butenuth. Tracking multiple vehicles in airborne image sequences of complex urban environments. In *Joint Urban Remote Sensing Event*, S. 13–16, 2011. DOI 10.1109/JURSE.2011.5764707.
- U. Thomas, D. Rosenbaum, F. Kurz, S. Suri, und P. Reinartz. A new software/hardware architecture for real time image processing of wide area airborne camera images. *Journal of Real-Time Image Processing*, 4(3):229–244, 2008. DOI 10.1007/s11554-008-0109-6.
- Markus Ulrich. *Hierarchical Real-Time Recognition of Compound Objects in Images*. Dissertation, Technische Universität München, 2003.
- Koen E.A. van de Sande, Theo Gevers, und Cees G.M. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 32(9):1582–1596, 2010. DOI 10.1109/TPAMI.2009.154.
- Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer, 2000. ISBN 978-0-387-98780-4.
- C.J. Veenman, M.J.T. Reinders, und E. Backer. Resolving motion correspondence for densely moving points. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 23(1):54–72, 2001. DOI 10.1109/34.899946.
- Claude Vidal, Jean-Guy Boureau, Nicolas Robert, Nicolas Py, Josiane Zerubia, Xavier Descombes, und Guillaume Perrin. Automatic crown cover mapping to improve forest inventory. In *Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium*, S. 333–340, 2006.
- P. Viola und M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, Bd. 1, S. 511–518, 2001. DOI 10.1109/CVPR.2001.990517.
- P. Viola, M. J. Jones, und D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005. DOI 10.1007/s11263-005-6644-8.
- A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. DOI 10.1214/aoms/1177731118.
- John R. Werthmann. Step-by-step description of a computationally efficient version of multiple hypothesis tracking. In *Signal and Data Processing of Small Targets, Conf. on*, Bd. 1698, S. 288–300, SPIE, 1992. DOI 10.1117/12.139379.
- Bo Wu und Ram Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75:247–266, 2007. DOI 10.1007/s11263-006-0027-7.
- Zheng Wu, T.H. Kunz, und M. Betke. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 1185–1192, 2011. DOI 10.1109/CVPR.2011.5995515.
- Jiangjian Xiao, Hui Cheng, Feng Han, und H. Sawhney. Geo-spatial aerial video processing for scene understanding and object tracking. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 1–8, 2008a. DOI 10.1109/CVPR.2008.4587434.
- Jiangjian Xiao, Changjiang Yang, Feng Han, und Hui Cheng. Vehicle and person tracking in aerial videos. In *Multimodal Technologies for Perception of Humans*, Bd. 4625 aus LNCS, S. 203–214, Springer, 2008b. DOI 10.1007/978-3-540-68585-2_18.

- Jiangjian Xiao, Hui Cheng, H. Sawhney, und Feng Han. Vehicle detection and tracking in wide field-of-view aerial video. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 679–684, 2010. DOI 10.1109/CVPR.2010.5540151.
- Alper Yilmaz, Omar Javed, und Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):1–45, 2006. DOI 10.1145/1177352.1177355.
- Qian Yu und G. Medioni. Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, S. 2671–2678, 2009. DOI 10.1109/CVPRW.2009.5206541.
- Qian Yu, I. Cohen, G. Medioni, und Bo Wu. Boosted markov chain monte carlo data association for multiple target detection and tracking. In *Pattern Recognition, IEEE Int. Conf. on*, Bd. 2, S. 675–678, 2006. DOI 10.1109/ICPR.2006.336.
- Chang Yuan, G. Medioni, Jinman Kang, und I. Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 29(9):1627–1641, 2007. DOI 10.1109/TPAMI.2007.1084.
- Beibei Zhan, Dorothy N. Monekosso, Paolo Remagnino, Sergio A. Velastin, und Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19:345–357, 2008. DOI 10.1007/s00138-008-0132-4.
- Tao Zhao und Ram Nevatia. Car detection in low resolution aerial images. *Image and Vision Computing*, 21(8):693–703, 2003. DOI 10.1016/S0262-8856(03)00064-7.

A. Anhang

A.1. Umformung der Hypothesenwahrscheinlichkeit

Nachfolgend werden die im Abschnitt 4.2.1 präsentierten Formeln ausführlich hergeleitet. Ausgangspunkt ist die von Reid (1979) entwickelte Gleichung zur Ermittlung der Wahrscheinlichkeit einer bestimmten Hypothese Θ_i^t zum Zeitpunkt t unter Berücksichtigung sämtlicher Beobachtungen $Z^{1:t}$ im Untersuchungszeitraum:

$$P(\Theta_i^t | Z^{1:t}) = \frac{1}{c_0} \cdot P(Z^t | \theta_h^t, \Theta_g^{t-1}, Z^{1:t-1}) \cdot P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) \cdot P(\Theta_g^{t-1} | Z^{1:t-1}) \quad (\text{A.1})$$

Der konstante Faktor c_0 spielt keine Rolle bei der Ermittlung der wahrscheinlichsten Hypothese und wird daher oft vernachlässigt. Er dient allein zur Normalisierung und ergibt sich aus der Summe der Einzelwahrscheinlichkeit aller zum Zeitpunkt t berücksichtigten Hypothesen Θ^t :

$$c_0 = \sum_{i=1}^{N_{\Theta^t}} P(\Theta_i^t | Z^{1:t}) \quad (\text{A.2})$$

Darüber hinaus besteht Gleichungen A.1 aus folgenden drei Teilen:

1. $P(Z^t | \theta_h^t, \Theta_g^{t-1}, Z^{1:t-1})$, der Wahrscheinlichkeit die Beobachtungen $Z^t = \{z_j^t\}$ zum Zeitpunkt t zu erhalten, gegeben aller bisherigen Beobachtungen $Z^{1:t-1}$ und Zuordnungsergebnisse $\Theta_g^{t-1} = \{\theta^1, \dots, \theta^{t-1}\}$ inklusive des aktuellen θ_h^t ,
2. $P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1})$, der Wahrscheinlichkeit des aktuellen Zuordnungsergebnisses θ_h^t , gegeben aller bisherigen Ereignisse Θ_g^{t-1} und Beobachtungen $Z^{1:t-1}$ sowie
3. $P(\Theta_g^{t-1} | Z^{1:t-1})$, der Wahrscheinlichkeit der Elternhypothese Θ_g^{t-1} zum vorhergehenden Zeitpunkt.

Der erste Teil der Gleichung, die bedingte Wahrscheinlichkeit die aktuellen Messungen zu erhalten, lässt sich entsprechend des zugehörigen Zuordnungsergebnisses θ_h^t faktorisieren. Dieses weist allen Beobachtungen eine der drei möglichen Optionen zu: Falschalarm (FA), neues Objekt (NT) oder wiedererkanntes Objekt (DT).

$$P(Z^t | \theta_h^t, \Theta_g^{t-1}, Z^{1:t-1}) = \prod_j^{N_{NT}} p_{NT}(z_j^t) \cdot \prod_j^{N_{FA}} p_{FA}(z_j^t) \cdot \prod_j^{N_{DT}} f_{DT}(z_j^t | \theta_h^t, Z^{1:t-1}) \quad (\text{A.3})$$

Die Wahrscheinlichkeitsverteilungen eines einzelnen Falschalms, neuen Objekts und wiedererkannten Objekts sind mit p_{FA} , p_{NT} und f_{DT} bezeichnet. Der zweite Teil von Gleichung A.1, die bedingte Wahrscheinlichkeit des aktuellen Zuordnungsergebnisses, lässt sich ebenfalls zerlegen. Unter der Annahme, dass die Anzahl der neuen Objekte N_{NT} und Falschalarme N_{FA} poissonverteilt ist und die Anzahl der wiedererkannten Objekte N_{DT} binomialverteilt, ergibt sich folgendes (für Details s. Reid (1979)):

$$P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) = \frac{N_{FA}! N_{NT}!}{N_Z!} \cdot P_D^{N_{DT}} \cdot (1 - P_D)^{N_{LT}} \cdot P_{\hat{N}_{FA}}(N_{FA}) \cdot P_{\hat{N}_{NT}}(N_{NT}) \quad (\text{A.4})$$

A. Anhang

Die zu erwartende Anzahl an Falschalarmen und neuer Objekten ist mit \hat{N}_{FA} bzw. \hat{N}_{NT} bezeichnet, die Anzahl der aktuellen Messungen mit N_Z . Fügt man nun die Formel der Poisson-Verteilung ein, ergibt sich:

$$P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) = \frac{N_{FA}! N_{NT}!}{N_Z!} \cdot P_D^{N_{DT}} \cdot (1 - P_D)^{N_{LT}} \cdot \frac{(\hat{N}_{FA})^{N_{FA}}}{N_{FA}!} e^{-\hat{N}_{FA}} \cdot \frac{(\hat{N}_{NT})^{N_{NT}}}{N_{NT}!} e^{-\hat{N}_{NT}} \quad (\text{A.5})$$

Kürzt man und ordnet konstante Terme nach vorn, folgt:

$$P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) = \frac{e^{-\hat{N}_{FA}} \cdot e^{-\hat{N}_{NT}}}{N_Z!} \cdot P_D^{N_{DT}} \cdot (1 - P_D)^{N_{LT}} \cdot (\hat{N}_{FA})^{N_{FA}} \cdot (\hat{N}_{NT})^{N_{NT}} \quad (\text{A.6})$$

Kombiniert man dieses Ergebnis mit Gleichung A.3, lassen sich einige Terme zusammenfassen:

$$P(Z^t | \theta_h^t, \Theta_g^{t-1}, Z^{1:t-1}) \cdot P(\theta_h^t | \Theta_g^{t-1}, Z^{1:t-1}) = \frac{e^{-\hat{N}_{FA}} \cdot e^{-\hat{N}_{NT}}}{N_Z!} \cdot \prod_{i=1}^{N_{LT}} (1 - P_D) \cdot \prod_j^{N_{NT}} [\hat{N}_{NT} \cdot p_{NT}(z_j^t)] \cdot \prod_j^{N_{FA}} [\hat{N}_{FA} \cdot p_{FA}(z_j^t)] \cdot \prod_j^{N_{DT}} [f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P_D] \quad (\text{A.7})$$

Nutzt man dieses Ergebnis, um die entsprechenden Terme in Gleichung A.1 zu ersetzen und berücksichtigt zusätzlich den Zusammenhang in Gleichung 4.17, erhält man die Formel 4.8, in welcher die Anteile der verschiedenen Zuordnungsoptionen an der Hypothesenwahrscheinlichkeit deutlich zum Ausdruck kommen:

$$P(\Theta_i^t | Z^{1:t}) = \frac{1}{c_1} \cdot \prod_j^{N_{NT}} f_{NT}(z_j^t) \cdot \prod_j^{N_{FA}} f_{FA}(z_j^t) \cdot \prod_j^{N_{DT}} [f_{DT}(z_j^t | \Theta_i^t, Z^{1:t-1}) P_D] \cdot \prod_{i=1}^{N_{LT}} (1 - P_D) \cdot P(\Theta_g^{t-1} | Z^{1:t-1}) \quad (\text{A.8})$$

Die Konstante c_1 fasst alle hypothesenunabhängigen Terme zusammen:

$$c_1 = c_0 \cdot \frac{N_Z!}{e^{-\hat{N}_{FA}} \cdot e^{-\hat{N}_{NT}}} \quad (\text{A.9})$$