Informatik Forsch. Entw. (2002) 17: 21–40

Informatik
Forschung und Entwicklung

Erscheinungsbasierte statistische Objekterkennung

Josef Pösl*, **Heinrich Niemann**

Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen
(e-mail: J.Poesl@fh-amberg-weiden.de, niemann@informatik.uni-erlangen.de)

Eingegangen am 5. April 2000 / Angenommen am 20. Juli 2001

Zusammenfassung. Die automatische Erkennung und Lokalisation von Objekten in digitalen Bildern ist ein wesentlicher Bestandteil vieler praktisch relevanter Anwendungen. In diesem Artikel wird ein erscheinungsbasiertes Verfahren zur Erkennung starrer zwei- oder dreidimensionaler Objekte vorgestellt, dem eine statistische Modellierung zugrundeliegt. Im Gegensatz zu segmentierungsbasierten Verfahren, wie sie vor allem im Bereich der 3D-Objekterkennung eingesetzt werden, ermöglicht der erscheinungsbasierte Ansatz aufgrund der Modellierung der Intensitätswerte oder davon abgeleiteter lokaler Merkmale eines Bildes die Erkennung komplexer Objekte. Die statistische Formulierung der Problemstellung bildet den mathematischen Kontext zur Bestimmung optimaler Lösungen.

Die Form der Modellierung erlaubt neben der Einzelobjekterkennung auch die Berücksichtigung von heterogenem Bildhintergrund und Mehrobjektszenen. Die dazu benötigten lokalen Merkmale entstehen durch räumlich begrenzte Transformationen des Bildes, wie beispielsweise Gabor- oder Wavelet-Transformationen. Die statistische Modellierung beschreibt die Verteilung dieser lokalen Merkmale anhand einer Dichtefunktion, die sich bei der Hintergrund- und Mehrobjektmodellierung als Mischungsverteilung der Einzelobjektverteilungen ergibt. Die Aufgabenstellungen des Erlernens und Erkennens von Objekten sind damit als Parameterschätzprobleme formal darstellbar. Dabei werden im einen Fall die Modellparameter und im anderen Fall die Lageparameter beziehungsweise die Klassen von Objekten geschätzt. Die experimentelle Überprüfung des Ansatzes anhand realer Objektaufnahmen durch CCD-Kameras zeigt seine Brauchbarkeit zur Erkennung von 2D- und 3D-Objekten bei homogenem und heterogenem Hintergrund.

Schlüsselwörter: Objekterkennung, Wavelets, statistische Modellierung

Abstract. The automatic recognition and localization of objects in digital images is an important part of many real world applications. In this article we present an appearance based statistical approach for the recognition of two and three dimensional objects. In contrast to segmentation based methods, the appearance based approach allows the recognition of complex objects. This is a consequence of modeling image intensity values or derived local features. The statistical formulation of the recognition problem defines the mathematical context for its optimal solution.

The model allows the recognition of single objects in images with homogeneous background as well as objects within heterogeneous background or multi object scenes. Local features for this model are derived from local transformations of the image. Examples for such transformations are Gabor or Wavelet transforms. The statistical model describes the distribution of the local features with a density function, which is defined as mixture density in case of background and multi object scenes. The object recognition and localization task can be formulated as parameter estimation problem, where the model parameters, the object pose and the object class are estimated respectively. We finally show experimental results of the presented approach, which demonstrate, that it can be utilized to successfully localize and recognize two and three dimensional objects in CCD-camera images with homogeneous as well as heterogeneous background.

Keywords: Object recognition, Wavelets, Statistical models

CR Subject Classification: I.4.8, I.4.7, I.5.3, G.3

1 Einleitung und Motivation

Die Erkennung und die Lokalisation von Objekten in Bildaufnahmen sind aktuelle und in ihrer Allgemeinheit bisher ungelöste Probleme, die in vielen praktisch relevanten Anwendungsszenarien zu lösen sind. Beispielhaft für die Vielzahl der Anwendungen seien hier die Erkennung von Bauteilen bei der Fertigung oder von Gegenständen in Haushalts- oder Serviceumgebungen durch kamerageführte Roboter genannt.

* Die vorliegende Arbeit entstand im Rahmen der Mitgliedschaft des Autors im Graduiertenkolleg für 3D-Bildanalyse und -synthese, das durch die Deutsche Forschungsgemeinschaft getragen wird.

In derartigen Szenarien werden zunächst mit einer oder mehreren Kameras Aufnahmen der zu analysierenden Umgebung gemacht. Dabei werden oft, beispielsweise aus Kostengründen, keine Aufnahmegeräte eingesetzt, die ein dreidimeinsionales Abbild des aufgenommenen Bereichs erzeugen. Stattdessen wird die reale dreidimensionale Welt durch CCD-Kameras auf ein oder mehrere zweidimensionale Bilder projiziert. Je nach Anwendung können entweder zu bestimmten Zeitpunkten Einzelaufnahmen einer Szene gemacht oder ganze Bildfolgen aufgenommen werden. In diesem Artikel wird auf Einzelbilder einer einzelnen Kamera fokussiert, in denen vorgegebene Objekte zu erkennen sind. Allgemein ist eine Übertragung eines Erkennungssystems für Einzelaufnahmen auf Bildfolgen möglich, indem das System einfach auf jedes Bild der Folge angewandt wird, wie unter anderem in [14] gezeigt wird. Allerdings gibt es für die Analyse von Bildfolgen auch alternative Ansätze, welche die speziellen Eigenschaften aufeinanderfolgender Bilder ausnutzen (siehe beispielsweise [7]). Ähnliches gilt für Aufnahmen mit mehreren Kameras, wie etwa Stereoaufnahmen, welche die gleiche Szene zum selben Zeitpunkt aus verschiedenen Blickrichtungen zeigen.

Liegt nur ein einzelnes Bild einer Szene beispielsweise mit Fertigungsstücken vor, so besteht die Schwierigkeit der Erkennung der Objekte darin, dass die dreidimensionale Objektstruktur nicht mehr unmittelbar und vollständig erkennbar ist. Stattdessen zeigt das zweidimensionale Bild eine Menge von Helligkeitswerten. Bereits die große Zahl der Bildpunkte stellt ein wesentliches Problem dar. So bestehen beispielsweise die Bilder, die in den Experimenten dieser Arbeit vorliegen, bei einer horizontalen und vertikalen Auflösung von je 256 Bildpunkten aus insgesamt über 60000 einzelnen Bildpunkten. Hinzu kommt, dass die Helligkeitswerte der Punkte von den Beleuchtungsbedingungen und anderen Umgebungseigenschaften abhängig sind. Eine weitere Schwierigkeit besteht darin, dass bei vielen realen Anwendungen keine Information über die genaue Lage des Objekts vorliegt. Da bereits die Lage eines Objekts in einem Bild durch drei Dimensionen beschrieben wird (Drehung und horizontale/vertikale Position) und außerdem das Objekt im Raum gedreht sein oder in unterschiedlicher Entfernung zur Kamera liegen kann, resultiert im allgemeinsten Fall ein hochdimensionaler Suchraum für die Objektposition, was zu entsprechenden Rechenzeiten führt.

Die meisten der bisher bekannten Lösungsansätze sind nur auf Objekte und eine Umgebung mit eng vorgegebenen Rahmenbedingungen anwendbar. Das bedeutet beispielsweise, dass die Beleuchtung nur in einem sehr engen Rahmen schwanken darf oder nur Objekte mit wenigen „scharfen“ Kanten bezüglich des Helligkeitskontrasts in den Bildern erlaubt sind. Nur wenige Ansätze sind auf komplexe Objekte beziehungsweise Szenarien mit einer größeren Variabilität in der möglichen Erscheinungsform von Objekten und Bildhintergrund skalierbar. Die Schwierigkeit ergibt sich in der Regel aus dem Versuch, die Geometrie der betrachteten Objekte direkt im Modell darzustellen, oder durch die Verwendung einer rein heuristischen und zumeist fest vorgegeben Metrik zur Charakterisierung von Objektähnlichkeiten im Merkmalsraum, in den die Objekte zur Vereinfachung der Erkennung abgebildet werden. Wird etwa versucht, jedes Objekt mit einem dreidimensionalen Gitter-Modell aus seinen Kanten darzustellen, so wird diese Vorgehensweise bereits bei einem Objekt mit gekrümmter glatter Oberfläche versagen.

Generell lassen sich zwei Ansätze zur Objekterkennung in Bildern unterscheiden: der segmentierungsbasierte und der erscheinungsbasierte Ansatz. Beim segmentierungsbasierten Ansatz werden bedeutungstragende Bestandteile, wie etwa Ecken oder Kanten, aus einem Bild extrahiert und zur Objektmodellierung und Objekterkennung verwendet. Nachteilig an diesem Ansatz ist zum einen, dass Segmentierungsfehler die Erkennung selbst erschweren, und zum anderen der Informationsverlust durch die Beschränkung der Erkennung auf die Ergebnisse einer vorgelagerten Phase. Zu diesen beiden Nachteilen kommt erschwerend hinzu, dass die Segmentierung in der Regel ohne Anpassung an verschiedene Objektteile gleichförmig im gesamten Bild durchgeführt wird.

In diesem Artikel wird ein Verfahren zur Erkennung starrer zwei- oder dreidimensionaler Objekte entwickelt, das erscheinungsbasiert ist und auf statistischen Modellen aufbaut. Ein Objekt wird dabei als zweidimensional bezeichnet, wenn es nur senkrecht zur Bildebene gedreht werden kann, also in allen relevanten Aufnahmen keine dreidimensionale Struktur erkennen lässt. Im Gegensatz zu segmentierungsbasierten Verfahren, wie sie vor allem im Bereich der 3D-Objekterkennung auf der Grundlage von Geometriemodellen der Objekte eingesetzt werden, ermöglicht der erscheinungsbasierte Ansatz durch die Modellierung der Intensitätswerte oder daraus abgeleiteter lokaler Merkmale eines Bildes die Erkennung komplexer Objekte. Außerdem kann damit ein Großteil des Informationsgehalts eines Bildes zur Erkennung eingesetzt werden und geht nicht, wie möglicherweise bei einer vorherigen Segmentierung, verloren.

Die statistische Modellierung ergibt eine mathematische Beschreibung der beobachteten Merkmalswerte und liefert damit die Grundlage für die Lösung der Objekterkennungsaufgaben durch Bestimmung optimaler Lösungen von Parameterschätzproblemen. Die Parameter können hierbei die Objektlage oder der -typ sein.

Der hauptsächliche Nachteil des statistischen Ansatzes besteht in der großen Menge von Bilddaten, die zum Training der Modellparameter erforderlich sind. Darauf wird in Abschnitt 6.1 noch näher eingegangen. Dies ist letztlich eine Folge der wenigen Annahmen, die über die zu berücksichtigenden Objekte gemacht werden, so dass im Modell Intensitätswerte abgebildet werden und natürlich in verschiedensten möglichen Erscheinungsformen – abhängig beispielsweise von den Umgebungsbedingungen – bereits zum Training verfügbar sein müssen. Um hier zumindest nicht alle möglichen lokalen Beleuchtungsschwankungen berücksichtigen zu müssen, bietet sich die Extraktion und Modellierung lokaler, teilweise beleuchtungsinvarianter Merkmale anstatt der ursprünglichen Intensitätswerte des Bildes an. Einige Möglichkeiten dafür werden in Abschnitt 4 aufgezeigt. Ein weiteres Problem statistischer Ansätze ist generell die Tatsache, dass selbst bei Modellen, die auf einfachen Einzeldichtefunktionen basieren, das Gesamtdichtemodell durch eine aufwändig berechenbare Funktion beschrieben wird. Diese Problematik lässt sich allerdings beispielsweise durch eine hierarchische Vorgehensweise, bei der

zunächst schnell berechenbare Funktionen für eine Grobanalyse des Bildes eingesetzt werden, entschärfen.

2 Literatur

Ein Überblick über die in der Literatur bekannten Verfahren zur Objekterkennung wird in [16, 19, 21] gegeben. In der Literatur finden sich auf dem Gebiet der erscheinungsbasierten Verfahren neben den klassischen korrelationsbasierten Verfahren Verallgemeinerungen in Form der Eigenraummethoden ([2, 15]). Bei beiden Verfahrenstypen wird ein Bild zunächst als Vektor dargestellt, in dem die Intensitätswerte aller Bildpunkte einfach in einer vorgegebenen Reihenfolge als Komponenten eingetragen sind, wobei beispielsweise zunächst die Bildpunkte der ersten Zeile von links nach rechts aufgeführt werden, dann die der zweiten Zeile usw. Die Grundidee bei korrelationsbasierten Verfahren ist die Beobachtung, dass der Betrag des Skalarprodukt zweier normierter Vektoren dann sein Maximum annimmt, wenn die beiden Vektoren übereinstimmen, sodass sich das Skalarprodukt zur Abstandsmessung von Vektoren eignet. Deshalb werden zu einem gegebenen Bild, das zu analysieren ist, beziehungsweise dem zugehörigen normierten Vektor, die Skalarprodukte mit einer Menge von Objektaufnahmen gebildet, die jeweils die zu erkennenden Objekte enthalten. Die Objektaufnahmen, die bereits im Vorfeld der eigentlichen Erkennung in einer sogenannten Trainingsphase angefertigt werden, werden aufgrund ihrer Funktion auch als Objektschablonen bezeichnet. Liegt das größte Skalarprodukt über einem vorgegebenen Schwellwert, so gilt das jeweilige Objekt als erkannt. Diese Vorgehensweise ist aber offensichtlich nur bei einigen wenigen möglichen Ansichten eines Objektes sinnvoll.

Kann ein dreidimensionales Objekt aus allen möglichen Blickrichtungen betrachtet werden, so liegen pro Objekt eine große Menge einzelner Trainingsbilder vor, die das Objekt für jede mögliche Drehrichtung im Raum zeigen. Eine Korrelation des zu analysierenden Bildes mit allen Modellbildern ist in diesem Fall zu aufwändig. Aus diesem Grund wird bei den Eigenraummethoden eine Approximation dieser Abstandsmessung, die wie oben beschrieben über das Skalarprodukt erfolgt, eingesetzt. Dazu wird eine lineare Eigenraumtransformation bestimmt, die den Abstand der transformierten Bilder einer Trainingsmenge minimiert. Entscheidend hierbei ist, dass die hochdimensionalen Bildvektoren durch diese lineare Abbildung in niedrigdimensionale Vektoren transformiert werden, wobei allerdings aufgrund der Konstruktion der Transformation die Ähnlichkeit von Bildern mit denen der Trainingsmenge auch an den jeweils transformierten, wesentlich kleineren Bildern immer noch erkennbar ist. Das bedeutet, dass die aufwändige Abstandsmessung im Originalraum durch eine schnell durchführbare approximative Messung im Eigenraum, also im Raum weniger beschreibender Merkmale ersetzt wird. Trotz dieser Vereinfachung ist eine Abstandsmessung zu den Merkmalsvektoren aller Trainingsbilder immer noch sehr aufwändig, sodass in einem weiteren Schritt die Menge der Punkte im Merkmalsraum, die durch die Merkmalsvektoren beschrieben wird, für die Analyse durch eine approximative Hyperfläche ersetzt wird. Da die aufgrund der Transformation gewonnenen Merkmale meist nicht unimodal verteilt sind, wird dazu unter anderem eine parametrische Mannigfaltigkeit im Eigenraum bestimmt, die eine Abstandsmessung und damit die Bewertung einer konkreten Objektlage oder eines Typs ermöglicht. Problematisch sind bei dem dabei in der Regel verwendeten quadratischen Abstandsmaß einzelne Ausreißer in den Bilddaten oder Objektverdeckungen.

Andere Verfahren messen die Abstände zu einer Menge von Objektschablonen, wobei sie statistische Abweichungen der Intensitätswerte berücksichtigen, modellieren die Verteilung von Intensitätswerten eines Objekts durch Histogramme oder verwenden Mischungsverteilungen zur Modellierung der Intensitätsverteilungen eines Bildes (siehe beispielsweise [11, 26, 28]). Ist die Zahl der Objektschablonen sehr groß, so bietet sich bei den entsprechenden Verfahren die Auswahl brauchbarer Teilmengen und die Definition eines geeigneten Abstandsmaßes an („Support Vector“-Maschinen, siehe beispielsweise [17, 3]).

Statistische Verfahren werden zur Einzelobjekterkennung – vor allem auf der Basis nicht-segmentierter Bilddaten – selten eingesetzt. In [11] wird beispielsweise eine Mischungdichte der Grauwerte von Objektaufnahmen definiert. Der EM-Energieterm (Erwartungswert-Maximierung, „Expectation Maximization“), der im Rahmen einer POEM-Schätzung (perzeptuell organisierte EM) eingesetzt wird, wird dabei jedoch um eine heuristische Komponente ergänzt (zum EM-Algorithmus, einem stochastischen Parameterschätzverfahren, siehe [6]). Außerdem ist die Auswertung sehr komplex, da jede Objektposition als verborgene Variable modelliert wird. Im Bereich der Erkennung handgeschriebener Buchstaben oder Ziffern finden statistische Klassifikatoren häufiger ihre Anwendung. Beispielhaft sei hier auf [5] verwiesen. Dort wird eine Mischungsverteilung für die unterschiedlichen auftretenden Objekttypen definiert, wobei die Dichtefunktion jedoch aufgrund der hohen Dimensionalität nicht auf die ursprünglichen Bilder (16×16 Bildpunkte) angewendet wird, sondern auf Merkmale nach einer Eigenraumtransformation. Durch den Einsatz eines statistischen Modells für die Eigenraummerkmale ist dieser Ansatz wesentlich flexibler auf die Verteilung der Merkmale trainierbar als die teilweise in der Literatur gewählten parametrisierten Mannigfaltigkeiten. Nachteilig daran ist allerdings, dass das Modell keine genaue Information über die Position des Objekts im dreidimensionalen Raum enthält. Überdies wird die Objektposition im Bild als bekannt vorausgesetzt. Dies ist für die Zeichenerkennung durchaus realistisch, da hier geeignete Verfahren zur Isolierung der einzelnen Zeichen zur Verfügung stehen. Für die Erkennung in Aufnahmen realer Objekte ist dies aber nur in wenigen Situationen so vorgegeben. Ein statistisches Objektmodell auf der Basis von Segmentierungsprimitiven wie Objekteckpunkten oder -kanten wird in [9] vorgestellt. In [4] wird ein statistisches Objektmodell zur Gesichtserkennung auf der Basis markanter Stellen im Gesicht, wie etwa der Position der Mundwinkel oder der Augen, präsentiert, das für alle möglichen Ansichten – vom Profil bis zur Frontalansicht – eingesetzt werden kann.

Statt einer statistischen Modellierung werden in der Literatur unter anderem auch neuronale Netze zur Objekterkennung eingesetzt. Als Beispiel sei hier auf [8] verwiesen. Dort wird jedes Objekt durch einen Vektor aus Merkma-

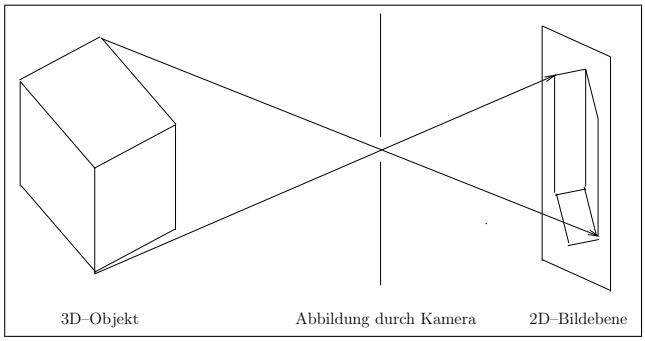
len beschrieben, die durch Gabor-Filterung des Bildes gewonnen werden. Die Klassifikation von Objekten in einem Bild erfolgt mit einem neuronalen Netz, das auf derartige Merkmalsvektoren trainiert ist. Da die Merkmale bei diesem System das Gesamtobjekt beschreiben und nicht einzelnen Bereichen zugeordnet sind, ist eine Robustheit gegenüber Objektverdeckungen nur schwer erreichbar. Außerdem muss die Objektposition zur Klassifikation bereits vorher bekannt sein. Hierzu wird eine Bildsegmentierung vorgeschaltet. Deutlich wird in [8] die Objektbeschreibungsfähigkeit von Gabor-Merkmalen. Ein etwas anderer Weg zur Gesichtserkennung mit Gabor-Merkmalen (in Kombination mit einer Wavelet-Analyse) wird in [10] eingeschlagen. Dort ist aufgrund der Dynamik in den menschlichen Gesichtszügen nur der Einsatz von lokalen Merkmalen angebracht. Diese werden allerdings in einer Vorphase zunächst nur an markanten Stellen aus dem Bild segmentiert, sodass der nachfolgende neuronale Ansatz jeweils nur eine überschaubare Zahl von dynamisch zugeordneten Merkmalen zu verknüpfen hat. Sehr interessant an diesem Verfahren ist, dass es auch auf nicht starre Objekte anwendbar ist (siehe auch [4]). Allerdings ist wie bei [8] keine genaue Bestimmung der Objektlage eines dreidimensionalen Objekts (abgesehen von seiner Position im Bild) möglich, da in der Regel nur eine Objektansicht modelliert wird und diese unabhängig von der genauen Lage von Teilmerkmalen ist.

Das in dieser Arbeit vorgestellte Verfahren grenzt sich von den Eigenraummethoden unter anderem durch seine statistisch fundierte Konzeption und von anderen statistischen Verfahren durch die Modellierung von 3D-Objekten und eine spezielle Art der Parametrisierung der Dichtefunktionen ab. Außerdem wird lokal begrenztes, aber nicht notwendigerweise stationäres Rauschen in Bezug auf die Bilddaten berücksichtigt und eine explizite Hintergrund-/Mehrobjektmodellierung durchgeführt.

Generell ist noch anzumerken, dass bei vielen Verfahren, die in der Literatur präsentiert werden, das Objekterkennungsproblem im Vordergrund steht. Deshalb wird dort die Objektlage im Bild als bekannt vorausgesetzt. Auf diese Voraussetzung wird in der vorliegenden Arbeit verzichtet. Stattdessen wird vor einer Objekterkennung zunächst die Objektposition bestimmt. Aufgrund der Struktur der Modellierung ist es trotz der dadurch wesentlich aufwändigeren Aufgabenstellung möglich, brauchbare Rechenzeiten zu erzielen (siehe Abschnitt 7).

3 Problembeschreibung und Systemübersicht

Wie bereits in der Einleitung beschrieben, ist die Zielsetzung des in diesem Artikel präsentierten Verfahrens die Erkennung und Positionsbestimmung von Objekten in einzelnen Bildern. Um die eigentliche Aufgabenstellung besser zu verstehen, soll zunächst der Vorgang der Bildentstehung betrachtet werden: Bei der Bildaufnahme wird der dreidimensionale reale Raum auf die zweidimensionale Bildebene projiziert (siehe Abb. 1). Physikalisch gesehen besteht das Bild aus den Helligkeitswerten in der Bildebene. Diese kontinuierliche Helligkeitsfunktion wird durch die Kamera nur an diskreten Punkten abgetastet, sodass für die eigentliche



3D-Objekt Abbildung durch Kamera 2D-Bildebene

Abb. 1. Durch die Kameraaufnahme wird ein Objekt der dreidimensionalen realen Welt auf eine zweidimensionale Bildebene abgebildet

Analyse nur ein Raster von Helligkeitswerten zur Verfügung steht.

Eingabe für das vorliegende System ist also ein derart abgetastetes Bild. Als Ergebnis der Analyse wird in der einfachsten Form der Typ eines einzelnen im Bild vorliegenden Objekts und/oder die Lage des Objekts, also seine genaue Position und Drehlage im Raum, erwartet.

Um diese Aufgabenstellung mit einem statistischen Ansatz zu lösen, wird – vereinfacht ausgedrückt – eine statistische Dichtefunktion bestimmt, welche die Wahrscheinlichkeit beschreibt, dass ein vorgegebenes Bild ein bestimmtes Objekt in einer bestimmten Lage enthält. Die Lösung des Problems besteht dann einfach darin, den Objekttyp und seine Lage mit der größten Wahrscheinlichkeit zu berechnen. Offensichtlich ergibt sich für jedes Objekt eine andere Dichtefunktion. Wählt man für alle möglichen Objekte die gleiche Familie von Dichtefunktionen, also beispielsweise allgemein Normalverteilungen, so wird jedes spezifische Objekt durch einen konkreten Parametersatz – seine Modellparameter – charakterisiert. Erst mit den Modellparametern ergibt sich jeweils eine spezifische Dichtefunktion, welche die Erscheinung eines Objekts beschreibt.

Generell lassen sich aufgrund der statistischen Natur des Systems zwei Phasen unterscheiden: die Trainingsphase und die Erkennungsphase. In einer Trainingsphase werden zunächst die Modellparameter geschätzt und anschließend in der Erkennungsphase, die beliebig oft wiederholt werden kann, die Lage beziehungsweise der Typ von Objekten aufgrund der geschätzten Parameter bestimmt.

Um eine für beide Fälle geeignete Dichtefunktion zu erhalten, werden die Helligkeitswerte des Bildes nicht unmittelbar verwendet. Stattdessen werden zunächst aus den Helligkeitswerten Merkmale abgeleitet, deren Auftreten dann statistisch erfasst wird. Sowohl die Trainings- als auch die Erkennungsphase lassen sich also in die Schritte Merkmalsextraktion und Parameterschätzung unterteilen. Zur Modellierung werden lokale Merkmale verwendet, die einen räumlich begrenzten Bildbereich beschreiben. Dazu werden näherungsweise rotationsinvariante Merkmale definiert, die teilweise beleuchtungsunabhängig sind. In diesem Artikel werden Merkmale der diskreten Wavelet-Transformation und Merkmale der Gabor-Transformation vorgestellt. Die diskrete Wavelet-Transformation, für die eine schnelle Berechnungsvorschrift ähnlich der schnellen Fourier-Trans-

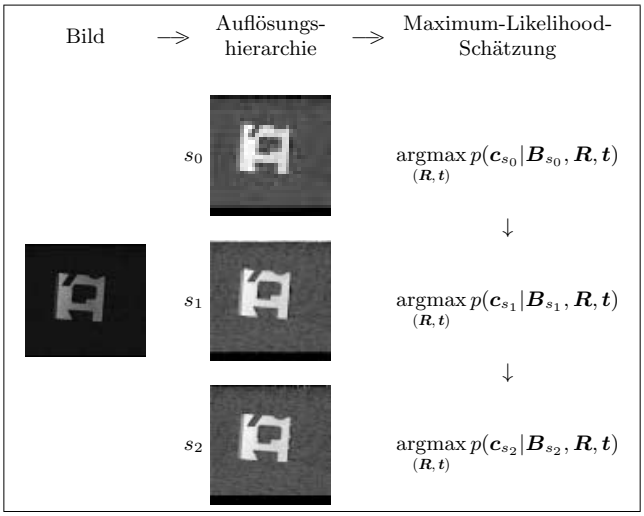


Abb. 2. Hierarchische Lageschätzung

formation existiert, ergibt mit den Tief- und Hochpassanteilen einer Auflösungsebene geeignete Merkmale. In dieser Arbeit werden zur Berechnung von Merkmalen unter anderem das Haar- und das Johnston-Wavelet verwendet. Gabor-Funktionen haben die Eigenschaft, nicht nur im Ortsbereich des Bildes, sondern auch gleichzeitig im Frequenzbereich einen lokal begrenzten Raumbereich zu beschreiben. Das bedeutet konkret, dass eine Gabortransformierte, die für eine vorgegebene Bildposition und Frequenz berechnet wird, nur durch die Intensitätswerte in unmittelbarer Nähe der Bildposition, also in einem lokalen Fenster um die Bildposition, beeinflusst wird, und dies auch nur dann, wenn die Änderung im Frequenzsprektrum des Fensterbereichs ebenfalls nahe bei der vorgegebenen Frequenz liegt. Gabor-Wavelets, zu denen die Transformation in Abschnitt 4.2 angegeben ist, parkettieren Orts- und Frequenzebene logarithmisch. Durch eine Zusatztransformation können damit Merkmale extrahiert werden, welche die Variation eines Signals über der Richtung im Zweidimensionalen rotations- und beleuchtungsinvariant beschreiben.

Bei der statistischen Modellierung wird eine Dichtefunktion definiert, welche die Wahrscheinlichkeit der Merkmale eines Bildes beschreibt. Die Parameter dieser Dichtefunktion werden in einer Trainingsphase geschätzt. Sowohl die Bestimmung der Dichteparameter als auch die Bestimmung von Objektlage und -typ erfolgen mit einer Maximum-Likelihood-Schätzung. Um die Objektlageschätzung effizient durchführen zu können, wird eine Auflösungshierarchie verwendet. In Abb. 2 ist dieses Vorgehensprinzip dargestellt. Auf der linken Seite ist ein Originalbild zu sehen, das zu analysieren ist. Zu diesem Originalbild werden Merkmale verschiedener Auflösungsebenen berechnet, die das Objekt unterschiedlich detailliert beschreiben. Die mittlere Spalte zeigt dies exemplarisch, indem ein einzelnes Merkmal pro Bildposition (pro Bildposition werden im allgemeinen mehrere Merkmale berechnet!) für drei Auflösungen in Grauwertbildern dargestellt wird. Die rechte Spalte von Abb. 2 schließlich zeigt die Positionsbestimmung (siehe Abschnitt 6). Ausgehend von einer globalen Lageschätzung auf der gröbsten Auflösungsebene s_0 wird die Lageschätzung auf den nachfolgenden Ebenen $s_1, \ldots$ schrittweise verfeinert. Auf jeder einzelnen Ebene erfolgt dabei eine Maximum-Likelihood-Schätzung der Positionsparameter des Objekts. Dabei beschreibt die Dichtefunktion $p(\boldsymbol{c}_{s_i}|\boldsymbol{B}_{s_i},\boldsymbol{R},\boldsymbol{t})$ die Wahrscheinlichkeit der beobachteten Merkmale $\boldsymbol{c}_{s_i}$ auf der Auflösungsebene s_i ($i = 0, 1, \ldots$) unter Annahme der Objektrotation $\boldsymbol{R}$ und -translation $\boldsymbol{t}$ für ein Objekt mit dem Modellparametersatz $\boldsymbol{B}_{s_i}$. Die Werte von Objektrotation und -translation werden bei der Schätzung jeweils so bestimmt, dass die Dichtefunktion für die Beobachtung, also die aus dem Bild abgeleiteten Merkmale, maximal wird.

Die Dichtefunktion der Merkmale einer Auflösungsebene wird für 2D-Objekte innerhalb eines Objektfensters als Normalverteilung der interpolierten Merkmalswerte modelliert, die sich aufgrund der Transformation in der Bildebene ergeben. Außerhalb des Objektfensters im Hintergrundbereich wird eine Gleichverteilung der Merkmale angenommen. Für 3D-Objekte werden die funktionalen Dichteparameter durch eine Linearkombination von einfachen Basisfunktionen dargestellt. Um die Dichteparameter sinnvoll schätzen zu können und bei der Lageschätzung brauchbare Rechenzeiten zu erhalten, werden anstatt einer vollständigen Merkmalsabhängigkeit nur die Unabhängigkeit oder lokale Abhängigkeit der Merkmale betrachtet. Die Dichtefunktion wird dabei als Produkt einer Menge von Teildichten dargestellt. Beispiele für lokale Abhängigkeiten sind Zeilen- oder Spaltenabhängigkeiten. Bei Spaltenabhängigkeiten werden nur die Merkmale jeweils einer Spalte als voneinander abhängig modelliert, die Merkmale unterschiedlicher Spalten werden als unabhängig angenommen. Die Gesamtdichte ergibt sich damit als Produkt der Dichtewerte der einzelnen Spalten. Außerdem gilt in diesem Fall, dass jedes Merkmal einer Spalte nur von seinen unmittelbaren Nachbarn abhängig ist. Die Dichte einer Spalte ergibt sich somit in einer etwas vereinfachten Schreibweise für die Merkmale $c_1,\ldots,c_S$ einer Spalte zu $p(c_1,\ldots,c_S) = p(c_1)\prod_{i=2}^{S} p(c_i|c_{i-1})$.

Bei der Hintergrund- und Mehrobjektmodellierung wird angenommen, dass jede Merkmalsposition entweder einem Objekt oder dem Hintergrund zugeordnet werden kann, sodass sich die Dichte als Mischungsverteilung über den möglichen Zuordnungen der Positionen ergibt.

4 Merkmale

4.1 Merkmalsextraktion allgemein

Aus theoretischer Sicht sind bei Erfassung der vollständigen statistischen Information des möglichen Bildmaterials die Ausgangsbildddaten und alle umkehrbar eindeutig daraus berechenbaren Merkmale gleichwertig für die Objekterkennung geeignet. Aufgrund der Rechenzeiten bei der Erkennung und der im Vergleich zu den möglichen Variationen der Umgebungsbedingungen kleinen Zahl von Trainingsaufnahmen erfolgt in der Praxis jedoch eine Einschränkung des statistischen Modells auf eine Familie von einfach berechenbaren parametrisierten Dichtefunktionen, die mit einer kleinen Zahl von Trainingsaufnahmen geschätzt werden können. Eine Folge davon ist, dass verschiedene lokale Transformationen der Bilddaten im Rahmen des Modells unterschiedlich gut für die Erkennung geeignet sind.

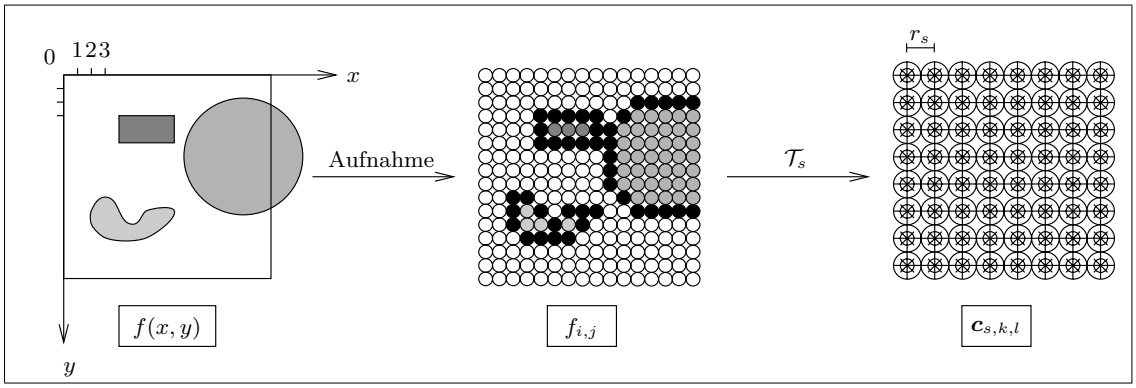


Abb. 3. Extraktion lokaler Merkmale: Durch die Abtastung bei der Aufnahme wird das kontinuierliche Bildsignal $f(x, y)$ in das diskrete Muster $\boldsymbol{f} = (f_{i,j})$ umgewandelt. Das Muster wird durch Transformationen $\mathcal{T}_{s,n}$ auf verschiedenen Auflösungsebenen s in lokale Merkmalswerte transformiert. In der Abbildung ist diese Transformation für eine Auflösung skizziert

Unter lokalen Merkmalstransformationen werden dabei Transformationen verstanden, die jeweils im Wesentlichen nur einen räumlich begrenzten Bildbereich zur Merkmalsberechnung verwenden. Die lokalen Merkmale beschreiben damit – im Gegensatz zu globalen Merkmalen des gesamten Bildes – einen Ausschnitt des Bildes beziehungsweise Objekts. Nur mit lokalen Merkmalen kann räumlich unterschiedliches Rauschen beim Training und Erkennen berücksichtigt werden.

Zur Extraktion lokaler Merkmale eignen sich sowohl klassische Verfahren der Merkmalsberechnung, wie beispielsweise die lokale Fourier-Transformation, als auch Wavelet- oder Gabor-Transformationen (verschiedene Transformationen werden in [19] dargestellt). Wie bereits erwähnt, werden dabei Merkmale verschiedener Auflösungsebenen berechnet, die Informationen über unterschiedlich große Bildbereiche beinhalten.

Wird mit einer Kamera ein zweidimensionales Bild aufgenommen, so wird die kontinuierliche Helligkeitsverteilung in der Bildebene an diskreten Positionen abgetastet und durch diskrete Helligkeitswerte dargestellt, das heißt quantisiert. Ein zweidimensionales Bild lässt sich also zunächst durch eine Funktion f beschreiben, welche die Helligkeitswerte an den kontinuierlichen Bildpositionen angibt. Durch das Aufnahmegerät wird dieses Bild an diskreten Positionen innerhalb eines Rechtecks abgetastet, sodass zur eigentlichen Verarbeitung eine diskretisierte Version des Bildes zur Verfügung steht. Ein Bild f ist dabei durch seine diskreten Abtastwerte $f_{i,j}$ mit $i \in \{0, 1, \ldots, D_x - 1\}$ und $j \in \{0, 1, \ldots, D_y - 1\}$ gegeben (siehe Abb. 3). Unter geeigneten Rahmenbedingungen, die unter anderem das Abtasttheorem beschreibt, lässt sich das ursprüngliche kontinuierliche Signal im Aufnahmebereich bei Bedarf wieder nahezu vollständig aus dem tatsächlich vorliegenden diskretisierten Bild zurückgewinnen.

Die Extraktion lokaler Merkmale lässt sich nun auf jeder Auflösungsebene s durch eine Transformationsfunktion $\mathcal{T}_s$ darstellen. Im nächsten Abschnitt wird auf einige spezielle Realisierungen dieser Transformation eingegangen. Generell gilt, dass die Auflösung des erzeugten Merkmalsraums bei $1 : r_s$ (Auflösung $r_s \in \mathbb{R}^+$) im Vergleich zum Originalbild liegt. Das bedeutet auch, dass die Zahl der Merkmalspositionen in jeder Bilddimension um den Faktor r_s kleiner ist als die Zahl der Bildpunkte. In Bezug auf das Originalbild sind die Merkmalspositionen um r_s Bildpunkte voneinander entfernt und beschreiben einen entsprechend großen Bildbereich. Hierbei wird allerdings nicht nur ein Merkmal pro Bildbereich extrahiert. Im Allgemeinen liefert die Transformation $\mathcal{T}_s$ einen Vektor von Einzelmerkmalen pro Merkmalsposition. Die Transformation setzt sich also aus mehreren Teiltransformationen $\mathcal{T}_{s,n}$ ($n = 0, 1, \ldots$) zusammen, die jeweils eine Merkmalskomponente pro Position liefern.

Im Folgenden werden die diskreten Bildpositionen der Auflösungsebene s und Auflösung r_s, an denen die Merkmale berechnet werden (Merkmalspositionen), mit $\boldsymbol{x}_{k,l} = (kr_s, lr_s)$ bezeichnet, wobei für die beiden Indizes k und l der beiden Bilddimensionen gilt, dass $k \in 0, \ldots, \lfloor \frac{D_x}{r_s} - 1 \rfloor$ und $l \in 0, \ldots, \lfloor \frac{D_y}{r_s} - 1 \rfloor$. Zur Vereinfachung der Schreibweise werden die beiden Indizes durch einen Index $m \in 0 \ldots M_s - 1$ aufgezählt, wobei $\boldsymbol{x}_m = \boldsymbol{x}_{k,l}$. Alle Merkmalspositionen werden damit ebenfalls durch m aufgezählt, wobei beispielsweise beginnend bei der ersten Spalte alle Spalten der Reihe nach von oben nach unten durchlaufen werden, sodass sich folgende Aufzählung ergibt: $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots = \boldsymbol{x}_{0,0}, \boldsymbol{x}_{0,1}, \boldsymbol{x}_{0,2}, \ldots \boldsymbol{x}_{1,0}, \boldsymbol{x}_{1,1}, \boldsymbol{x}_{1,2}, \ldots$. Die Zahl der Merkmalspositionen auf der Skalierungsebene s ist hierbei durch $M_s - 1$ gegeben. Die lokalen Merkmalsvektoren an diesen diskreten Bildpositionen $\boldsymbol{x}_m$ werden mit $\boldsymbol{c}_s(\boldsymbol{x}_m) = \boldsymbol{c}_{s,m} = \left(c_{s,m,0}, \ldots, c_{s,m,N-1}\right)^{\mathrm{T}}$, bezeichnet. Dabei ist N die Dimension der lokalen Merkmalsvektoren. Pro Merkmalsposition liegen also N unterschiedliche Merkmale vor. Zur Erzeugung einer Hierarchie von Merkmalen wird dabei die Auflösung r_s mit steigendem Index s der Auflösungsebenen kleiner ($r_{s+1} < r_s$). In den Experimenten in Abschnitt 7 sind die verwendeten Auflösungen beispielsweise $r_0 = 8$ und $r_1 = 4$.

4.2 Spezielle Verfahren zur Merkmalsextraktion

Aufgrund ihrer guten Eignung für die Objekterkennung [1, 8, 10, 22, 19] werden in diesem Artikel logarithmierte Koeffizienten von diskreten Wavelet-Transformationen und Gabor-Transformierte als Merkmale $\boldsymbol{c}_s$ vorgestellt.

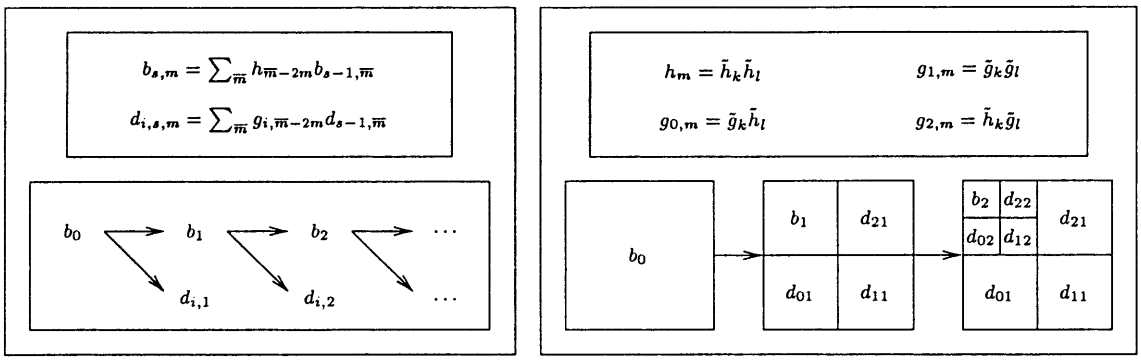


Abb. 4. Schnelle diskrete Wavelet-Transformation (linkes Bild) und die ersten zwei Analyseschritte bei Tensorproduktwavelets (rechtes Bild). Die Skalierungskoeffizienten, welche die Wavelet-Transformation beschreiben, sind dabei mit h_m und $g_{i,m}$ bezeichnet. Sie ergeben sich über die Gleichungen rechts oben aus den Koeffizienten $\tilde{h}_k$ und $\tilde{g}_i$ der eindimensionalen Transformation

Merkmale der diskreten Wavelet-Transformation. Die diskrete Wavelet-Transformation ist gut für die Extraktion lokaler Merkmale geeignet, da sie – im Gegensatz zur Fourier-Transformation – eine lokale Frequenzanalyse beinhaltet und eine schnelle Berechnung möglich ist. Im Gegensatz zur Fourier-Transformation, die in genau einer Form festgelegt ist, gibt es unterschiedliche diskrete Wavelet-Transformationen. Dabei erfolgt die Transformation selbst stets nach dem gleichen Prinzip, ist aber mit unterschiedlichen Koeffizienten parametriert. Da die Wavelet-Transformation eine Auflösungshierarchie erzeugt, werden diese Koeffizienten auch als Skalierungskoeffizienten bezeichnet. Im Fall eines zweidimensionalen Bildes werden die Koeffizienten ebenfalls in zwei Dimensionen angegeben. Um eine Symmetrie in der Behandlung der beiden Dimensionen zu gewährleisten und zudem die Berechnung der Transformation zu vereinfachen, bietet sich eine spezielle Form der Wavelet-Transformation an, nämlich die Tensorprodukt-Wavelet-Transformation (siehe Abb. 4). Dabei lässt sich die Transformation auf mehrere eindimensionale Transformationen zurückführen. Bereits die Skalierungskoeffizienten ergeben sich aus denen einer eindimensionalen Transformation.

Seien die Koeffizienten der zugehörigen eindimensionalen Transformation mit $\tilde{h}_k$ (Tiefpass) und $\tilde{g}_k$ (Hochpass) bezeichnet. Bei praktisch relevanten Transformationen sind nur circa 2 bis 20 dieser Koeffizienten verschieden Null. Der Index k ist damit bei der Berechnung nur für wenige Werte zu durchlaufen. Aus diesen eindimensionalen Koeffizienten ergeben sich wie in Abb. 4 rechts oben dargestellt die zweidimensionalen, indem das Tensorprodukt gebildet wird. Auf diese Weise ergeben sich ein „Satz“ Tiefpasskoeffizienten h_m und drei Sätze Hochpasskoeffizienten $g_{i,m}$ ($i = 0, 1, 2$). Die zweidimensionalen Koeffizienten sind durch das Indexpaar $\boldsymbol{m} = (k, l)^{\mathrm{T}} \in \mathbb{Z}^2$ indiziert. Die Werte, die sich aus der Wavelet-Transformation mittels dieser Skalierungskoeffizienten ergeben, bezeichnet man als Hochpass- und Tiefpasskoeffizienten der Transformation. Seien im Folgenden $b_{s,m}$ die Tiefpass- und $d_{i,s,m}$ mit $i \in \{0, 1, 2\}$ die Hochpasskoeffizienten einer derartigen diskreten Tensorprodukt-Wavelet-Transformation. Dabei bezeichnet $\boldsymbol{m} = (k, l) \in \mathbb{Z}^2$ die Ortskoodinate der Koeffizienten. Die Wavelet-Koeffizienten werden nach der in Abb. 4 links oben dargestellten Berechnungsvorschrift beginnend mit den Bilddaten $b_{0,m} = f_{kl}$ als feinste Auflösungsebene ($s = 0$) schrittweise für alle gröberen Auflösungen ($s = 1, 2, \ldots$) berechnet (siehe dazu im Detail [12, 19]). Schrittweise ergeben sich auf jeder Stufe Hoch- und Tiefpasskoeffizienten und die Tiefpasskoeffizienten werden in der jeweils nächsten Stufe weiterverarbeitet. Da die Auflösung der Koeffizienten auf jeder Stufe um den Faktor 2 abnimmt, lassen sich die Koeffizienten für alle durchzuführenden Transformationsstufen in einer Bildmatrix gleicher Größe darstellen, wie in Abb. 4 schematisch gezeigt. Hingewiesen sei an dieser Stelle noch auf die Tatsache, dass sich bei Tensorprodukt-Wavelets die Transformation auf mehrere eindimensionale Transformationen zurückführen lässt, was die Berechnung erheblich vereinfacht.

Aus den Koeffizienten der Wavelet-Transformation eines Bildes ergeben sich folgende Merkmale der diskreten Wavelet-Transformation, wobei multiplikative Beleuchtungsanteile durch Logarithmieren in additive Anteile umgewandelt werden und die Richtungsabhängigkeit der Hochpasskoeffizienten durch Mittelung abgeschwächt wird:

$$\begin{aligned} c_{s,m,0} &= \mathcal{T}_{s,0}^{DISK}\{f\}(\boldsymbol{x}_m) := \log|b_{s,\boldsymbol{m}}| \\ c_{s,m,1} &= \mathcal{T}_{s,1}^{DISK}\{f\}(\boldsymbol{x}_m) := \log\left(\sum_{i=0,1,2}|d_{i,s,\boldsymbol{m}}|\right). \end{aligned} \tag{1}$$

Merkmale der Gabor-Transformation. Die Motivation für den Einsatz der Gabor-Transformation liegt in ihrer physiologischen und theoretischen Bedeutung einerseits ([22]) und ihren erfolgreichen Ergebnissen bei der Merkmalsextraktion in verschiedenen Bereichen der Mustererkennung (siehe [19]) andererseits. Wird ihre Position $\boldsymbol{\omega}_0$ im Frequenzraum in Polarkoordinaten $(\omega_0, \phi_0) \in (\mathbb{R} \times [0, 2\pi[)^{\mathrm{T}}$ angegeben, dann ist die Darstellung der isotropen separierbaren Gabor-Funktion zur Frequenz ω_0 in der Richtung ϕ der Modulation mit der Fensterbreite σ gleich

$$\begin{aligned} &g_R(\boldsymbol{x}|\boldsymbol{x}_0, \omega_0, \phi, \sigma) \\ &\quad = g(\boldsymbol{x}|\boldsymbol{x}_0, \boldsymbol{R}(\phi)(\omega_0, 0)^{\mathrm{T}}, (\sigma, \sigma)^{\mathrm{T}}) \\ &\quad = \exp\left(-\left(\frac{(x-x_0)^2}{2\sigma^2} + \frac{(y-y_0)^2}{2\sigma^2}\right)\right. \\ &\qquad \left. +\mathrm{i}\left(\omega_0(\cos\phi)(x-x_0) - \omega_0(\sin\phi)(y-y_0)\right)\right). \end{aligned} \tag{2}$$

Da die Gabor-Funktionen keinen verschwindenden Mittelwert haben, das heißt ihre Fourier-Transformierte an der Stelle $\omega_0 = 0$ nicht gleich Null ist, ist die Zulässigkeit als Wavelets für den $L^2(\mathbb{R})$ nicht erfüllt. Dabei ist die Zulässigkeit eine notwendige Forderung an Wavelets, die vor allem gewährleistet, dass aus den Koeffizienten der Transformation das ursprüngliche Signal wieder rekonstruiert werden kann. Letztlich führt die Zulässigkeitsbedingung zu brauchbaren Koeffizienten, welche die ursprüngliche Bildinformation noch vollständig enthalten. Dies wird natürlich auch durch die Forderung erreicht, dass die Wavelet-Funktionen der verschiedenen Skalierungsebenen eine Basis des betrachteten Bildraums bilden. Betrachtet man räumlich- und bandbegrenzte Signale, wie sie im Fall eines diskret abgetasteten Bildes angenommen werden, so erfüllen Gabor-Wavelets, die sich aus dem Basis-Wavelet (siehe auch [13, 18] und Gleichung (2))

$$g_B(\boldsymbol{x}) = g_R(\boldsymbol{x}|\mathbf{0}, \omega_0, 0, \sigma_0) \tag{3}$$

mit $\sigma_0 = 1/\sqrt{2}$ durch Rotation, Dilatation und Translation ergeben, die notwendigen Basiseigenschaften. Die Zulässigkeitsbedingung kann außerdem durch einen meist vernachlässigbaren, additiven Zusatzterm erreicht werden (Morlet-Wavelet [1]). In der Literatur werden teilweise auch die Funktionen der gefensterten Fourier-Transformation als Gabor-Wavelets bezeichnet [27].

Bei den Gabor-Funktionen ist die Wavelet-Transformation der Skalierungsebene $s \in \mathbb{Z}$ einer zweidimensionalen Funktion f gegeben durch

$$w_s(\boldsymbol{x}|\theta) = \int f(\boldsymbol{x}_0) g_B\left(d^{-s}\boldsymbol{R}(\theta)(\boldsymbol{x} - \boldsymbol{x}_0)\right) d\boldsymbol{x}_0, \tag{4}$$

mit $d \in \mathbb{R}$, $d > 1$, $s \in \mathbb{Z}$ und $\theta \in \{\theta_\tau = \frac{\tau\pi}{N_\tau}\}_{\tau=0,\ldots,N_\tau-1}$ (zur diskreten Version siehe [19]). Hierbei bezeichnet $\boldsymbol{R}(\theta)$ die zweidimensionale Rotationsmatrix zum Winkel θ und s einen Zahlenwert, mit dem die Wavelet-Funktionen der einzelnen Ebenen skaliert werden, sodass sich eine Menge von im Vergleich zueinander gedrehten und unterschiedlich skalierten Wavelet-Transformationsfunktionen ergibt. Dies scheint zunächst eine völlig andere Form der Berechnung zu sein, wie sie im diskreten Fall im vorigen Abschnitt angegeben wurde. Zwischen beiden Varianten besteht aber eine sehr enge Verbindung, wozu hier allerdings auf die Literatur (beispielsweise [12]) verwiesen sei. Merkmale der Skalierungsebene s bei der Position $\boldsymbol{x}$ ergeben sich mit den Wavelet-Koeffizienten zu

$$\begin{aligned} c_{s,m,n} &= \mathcal{T}_{s,n}^{GABW}\{f\}(\boldsymbol{x}_m) \\ &:= \left| \underset{\rho=1}{\overset{N_\tau-1}{DFT}} \left\{ \left| \underset{\tau=0}{\overset{N_\tau-1}{DFT}} \{\log|w_s(m|\theta_\tau)|\}_\rho \right| \right\}_n \right|, \end{aligned} \tag{5}$$

wobei $N = \lfloor \frac{N_\tau+1}{2} \rfloor$ und

$$\underset{\rho=\rho_0}{\overset{\rho_1}{DFT}}\{f_\rho\}_\tau = \sum_{\rho=\rho_0}^{\rho_1} f_\rho \exp\left(-\frac{2\pi \mathrm{i}\rho\tau}{N_\tau}\right) \tag{6}$$

plus1pt minus2ptdie diskrete Fourier-Transformation ist. Die Zusatztransformation, die auf die Koeffizienten $w_s(m|\theta_\tau)$ angewandt wird und die in Analogie zur Berechnung des Cepstrums in der Spracherkennung zu sehen ist, ist dabei ähnlich motiviert wie im Fall der diskreten Wavelet-Transformation: Der Logarithmus der Ausgangswerte führt zur Umwandlung multiplikativer Beleuchtungsanteile in additive Anteile, die durch Vernachlässigung des konstanten Terms ($\rho = 0$) bei der DFT-Rücktransformation (äußerer DFT-Term) vollständig wegfallen. Dadurch werden die Merkmale robust gegenüber Beleuchtungsschwankungen. Außerdem führt die Summation über der Ausrichtung θ_τ zu weitgehend richtungsunabhängigen lokalen Merkmalen. Hingewiesen sei noch darauf, dass die DFT-Transformation und ihre Inverse nahezu identisch sind, weswegen bei der Rücktransformation die DFT selbst angegeben ist. Die DFT und ihre Inverse unterscheiden sich generell nur in der Normierung der resultierenden Koeffizienten und beim Vorzeichen im Exponenten, also bei der Form der Berechnung des komplexen Sinus. Da im vorliegenden Fall aufgrund der Betragsbildungen nur reelle Werte transformiert werden, spielt das Vorzeichen beim komplexen Sinus keine Rolle, sodass der Unterschied von DFT und ihrer Inversen nur noch in der Normierung liegt.

5 Statistische Modellierung

5.1 Überblick

Zielsetzung des vorliegenden Systems ist die Klassifikation und Lageschätzung eines starren zwei- oder dreidimensionalen Objekts in einer einzelnen zweidimensionalen Aufnahme. Dazu werden in diesem Abschnitt die statistischen Dichtefunktionen definiert, die das statistische Verhalten der aus den Bilddaten abgeleiteten Merkmale beschreiben.

Um den Zeitaufwand für die Lösung der Schätzprobleme auf ein praktikables Maß zu reduzieren, wird eine hierarchische Vorgehensweise eingeschlagen (Abb. 2). Dichtefunktionen werden dazu auf verschiedenen Auflösungsebenen der Merkmale definiert. Die Parameterschätzungen werden zunächst auf der gröbsten Auflösungsebene durchgeführt und anschließend auf detaillierteren Ebenen verfeinert.

Für die Bestimmung der Dichtefunktion der Merkmale einer Auflösungsebene gibt es unterschiedliche Vorgehensweisen. Die Schwierigkeit liegt prinzipiell in der Tatsache begründet, dass aus praktischen Gründen immer nur relativ wenige Trainingsbilder zur Schätzung des statistischen Verhaltens der Merkmale verfügbar sind und die eigentliche Erkennung so schnell wie möglich ablaufen soll. Dies führt dazu, dass die Modellierung mit einer Familie von Funktionen durchgeführt wird, die nur wenige Parameter aufweisen und schnell berechenbar sind. Aufgrund ihrer Einfachheit und der Tatsache, dass sie theoretisch sehr gut untersucht ist, bietet sich hier zunächst die Normalverteilung als Funktionenfamilie an. Sie weist im eindimensionalen Fall nur zwei Parameter, nämlich Mittelwert und Varianz auf und eignet sich zumindest zur Modellierung unimodaler Verteilungen, so wie sie bei den Merkmalen einer Objektaufnahme vorliegen, bei der sich nur additive Schwankungen beispielsweise aufgrund von Beleuchtungsschwankungen (siehe Merkmalsberechnung) ergeben. Dies zeigen die Merkmalsverteilungen in den Trainingsmengen. Letztlich sind aber die Ergebnisse

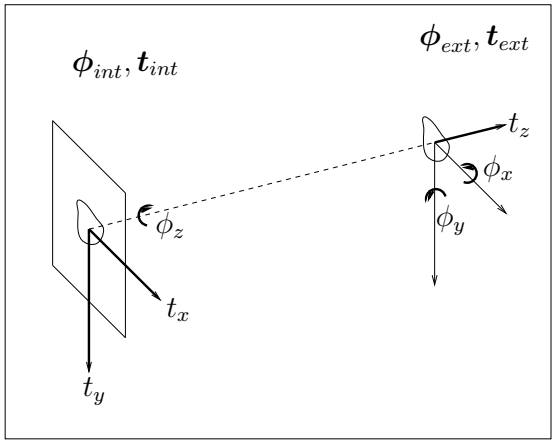


Abb. 5. 3D-Transformationsparameterraum unterteilt in interne Transformation (ϕ_{int}, $\boldsymbol{t}_{int}$) mit ϕ_{int} = (ϕ_z) und $\boldsymbol{t}_{int}$ = (t_x, t_y, 0)$^{\mathrm{T}}$ und externe Transformation (ϕ_{ext}, $\boldsymbol{t}_{ext}$) mit ϕ_{ext} = (ϕ_y, ϕ_x) und $\boldsymbol{t}_{ext}$ = (0, 0, t_z)$^{\mathrm{T}}$

der Erkennung (siehe Abschnitt 7) das entscheidende Kriterium für die Beurteilung der Modellierung mit einer Normalverteilung. Werden in einer Aufnahme mehrere Objekte, Verdeckungen oder einfach nur andere Ansichten ein- und desselben Objekts zugelassen, dann reicht diese Form der Modellierung nicht mehr aus. In diesem Fall gibt es weitere Einflussfaktoren, wie die Objektlage oder die Kombinationsmöglichkeiten von Objekten in einem Bild, welche die Verteilung der Merkmale bestimmen. Um auch derartige Modifikationen zu erfassen, muss ihre Auswirkung auf die Merkmale untersucht werden. Dies führt zu entsprechenden Erweiterungen des statistischen Modells.

Seien $\boldsymbol{c}_s$ der Vektor der konkatenierten Merkmalsvektoren der Auflösungsebene s, $\boldsymbol{B}_s$ die Modellparameter eines Objekts und $\boldsymbol{R}, \boldsymbol{t}$ die 3D Rotationsmatrix beziehungsweise der Translationsvektor, welche die Lage des Objekts bezüglich einer Referenzlage beschreiben. Mit der Definition der lokalen Merkmalsvektoren an den Bildpositionen $\boldsymbol{x}_m$ ($m \in 0 \ldots M_s - 1$) als $\boldsymbol{c}_s(\boldsymbol{x}_m) = \boldsymbol{c}_{s,m} = \left(c_{s,m,0}, \ldots, c_{s,m,N-1}\right)$ nach Abschnitt 4.1 ergibt sich der Gesamtmerkmalsvektor $\boldsymbol{c}_s$ durch Konkatenation zu

$$\begin{aligned}\boldsymbol{c}_s = \big(&c_{s,0,0}, \ldots, c_{s,0,N-1},\\ &c_{s,1,0}, \ldots, c_{s,1,N-1},\\ &\ldots,\\ &c_{s,M_s-1,0}, \ldots, c_{s,M_s-1,N-1}\big)^T .\end{aligned} \tag{7}$$

Die Rotation $\boldsymbol{R}$ ist durch die Rotationswinkel ϕ_x, ϕ_y und ϕ_z um die x-, y- beziehungsweise z-Achse definiert (siehe Abb. 5).

Die Modellparameter $\boldsymbol{B}_s$ bestehen aus geometrischer Information, wie den Positionen der lokalen Wahrscheinlichkeitsdichten und anderen Dichteparametern.

Um nun auf Basis der Dichtefunktion $p(\boldsymbol{c}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$ der Merkmale einer Auflösungebene die Lage zu schätzen, werden die Lageparameter so bestimmt, dass die Dichte maximal wird. Es wird also der wahrscheinlichste Lageparametersatz gewählt. Dies wird in Chapter 6 detaillierter ausgeführt. Hier sei nur noch die formale Notation dieser Form der Generierung von Objektlagehypothesen mit einer sogenannten Maximum-Likelihood-Schätzung der Lageparameter angegeben:

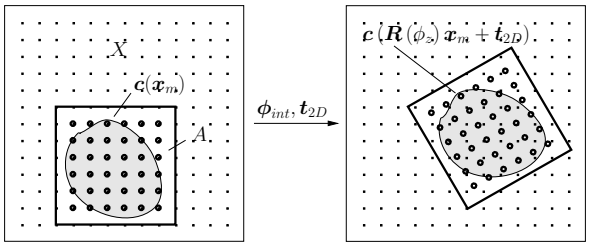


Abb. 6. Das Objekt wird durch ein Gitter lokaler Merkmale überdeckt

$$\left(\widehat{\boldsymbol{R}}_s, \widehat{\boldsymbol{t}}_s\right) = \underset{\boldsymbol{R},\boldsymbol{t}}{\operatorname{argmax}}\, p(\boldsymbol{c}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t}). \tag{8}$$

5.2 Einzelobjektdichte

In diesem Abschnitt wird eine Dichtefunktion für ein einzelnes Objekt definiert. Zur Vereinfachung der Notation wird der Index s weggelassen, da die Ableitung der Dichte für alle Auflösungsebenen analog erfolgt. Außerdem werden im Folgenden alle Dichten für eindimensionale Merkmalsvektoren $\boldsymbol{c}(\boldsymbol{x}_m) = c_m$ abgeleitet. Das heißt, es wird für jede Merkmalsposition genau ein reeller Merkmalswert bei der Modellierung berücksichtigt. Die Formulierung der Theorie für Vektoren mit mehreren Dimensionen ist analog möglich (siehe [19]).

Das Modellobjekt wird durch ein rechteckiges Gitter lokaler Merkmale (allgemein: Merkmalsvektoren, siehe obige Einschränkung) überdeckt (siehe Abb. 6). Die Gitterauflösung entspricht dabei der Auflösung der Merkmale auf der jeweiligen Auflösungsebene. Sei $A \subset \mathbb{R}^2$ ein kleiner Bereich (beispielsweise rechteckig), der die Projektion des Objekts auf die Bildebene für alle möglichen Objektrotationen ϕ_{ext} = (ϕ_y, ϕ_x) außerhalb der Bildebene überdeckt. Sei $X = \{\boldsymbol{x}_m\}_{m=0,\ldots,M-1}$, $\boldsymbol{x}_m \in \mathbb{R}^2$ die Menge der Gitterpositionen und $\boldsymbol{c}(\boldsymbol{x})$ der Merkmalsvektor an der Position $\boldsymbol{x}$. Außerdem werde in diesem Abschnitt angenommen, dass die Hintergrundmerkmale gleichverteilt und unabhängig von den Objekmerkmalen sind. Dann ist es ausreichend, die Dichte $p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t})$ zu betrachten, wobei $\boldsymbol{c}_A$ aus der Untermenge der Merkmale von $\boldsymbol{c}$ besteht, deren Positionen durch A überdeckt werden.

Die Merkmalsvektoren werden als normalverteilt mit unabhängigen Komponenten angenommen. Bezeichne $\mathcal{N}(\boldsymbol{c}|\mu, \boldsymbol{\Sigma})$ die Normalverteilung, wobei $\mu = (\mu_m)_m^{\mathrm{T}}$ der Mittelwertvektor mit konkatenierten lokalen Merkmalsmittelwerten μ_m ist und $\boldsymbol{\Sigma} = \left(\sigma_{m,\overline{m}}\right)_{m,\overline{m}}$ die Kovarianzmatrix mit den Elementen $\sigma_{m,\overline{m}} = \operatorname{cov}(c_m, c_{\overline{m}})$.

Wäre die Position des Objekts in Bezug auf die Kamera unveränderlich, so könnten die Objektmerkmale des Bildes mit einer „Standard-Normalverteilung“ modelliert werden, bei der Mittelwertvektor und Kovarianzmatrix Konstanten sind. Die Statistik würde in diesem Fall vor allem durch Helligkeitsschwankungen bei der Aufnahme, Rauschen der Sensoren des Aufnahmegeräts und Ähnliches erzeugt werden. Bei veränderlicher Objektlage, speziell bei einer unterschiedlichen Ansicht des gleichen Objekts, ergeben sich

allerdings selbst bei gleichen Umgebungsbedingungen unterschiedliche Merkmale. Die Dichteparameter sind deshalb eine Funktion der externen Transformation (ϕ_{ext}, $\boldsymbol{t}_{ext}$) dreidimensionaler Objekte, die ja die Objektansicht bestimmen, sodass

$$\begin{aligned} & p(\boldsymbol{c}_A|\boldsymbol{B},\phi,\boldsymbol{t}) \\ & \quad = p(\boldsymbol{c}_A|(\mu(\phi_{ext},\,\boldsymbol{t}_{ext}),\boldsymbol{\Sigma}(\phi_{ext},\,\boldsymbol{t}_{ext})),\phi_z,\boldsymbol{t}_{2D}) \\ & \quad = \mathcal{N}(\boldsymbol{c}_A(\phi_z,\boldsymbol{t}_{2D})|\mu(\phi_{ext},\,\boldsymbol{t}_{ext}),\boldsymbol{\Sigma}(\phi_{ext},\,\boldsymbol{t}_{ext})), \end{aligned} \tag{9}$$

wobei $\boldsymbol{c}_A(\boldsymbol{R}(\phi_z)\,,\boldsymbol{t}_{2D})$ die konkatenierten lokalen Merkmale $c(\boldsymbol{R}(\phi_z)\,\boldsymbol{x}_m + \boldsymbol{t}_{2D})$ innerhalb des Objektfensters sind, mit der 2D-Rotationsmatrix $\boldsymbol{R}(\phi_z)$ für die Rotation und der Translation $\boldsymbol{t}_{2D} = (t_x,\,t_y)^{\mathrm{T}}$ in der Bildebene. Da bei Rotation und Translation des Objekts in der Bildebene die Merkmalspositionen im Bild entsprechend rotiert beziehungsweise verschoben sind, lässt sich die Transformation in der Bildebene einfach durch Extraktion der Merkmale an diesen transformierten Positionen modellieren. Die Merkmale an den transformierten 2D-Positionen werden durch lineare Interpolation berechnet, sodass sich mit der transformationsabhängigen Interpolationsmatrix $\boldsymbol{W}$ für die Gesamtbeobachtung $\boldsymbol{c}$ folgende Darstellung ergibt:

$$\boldsymbol{W}(\phi_z,\boldsymbol{t}_{2D})\boldsymbol{c} =: \tilde{\boldsymbol{c}},\ \text{womit}\ \ \tilde{\boldsymbol{c}}(\boldsymbol{x}) = \boldsymbol{c}(\boldsymbol{R}(\phi_z)\boldsymbol{x} + \boldsymbol{t}_{2D}). \tag{10}$$

Angemerkt sei an dieser Stelle, dass die lineare Interpolation bei der Berechnung der Funktion in der Praxis natürlich nicht mit einer derartig aufwändigen Matrixmultiplikation durchgeführt wird. Die Merkmale werden einfach aus wenigen Nachbarmerkmalen interpoliert. Die angegebene Matrixdarstellung ist allerdings notwendig, um die mathematische Formulierung als Dichtefunktion anzugeben.

Unter der Annahme der Stetigkeit können die Funktionen μ_m und $\boldsymbol{\Sigma}$ durch eine Basismenge $\{v_r\}_{r=0,\ldots,\infty}$ des Funktionenraums der Funktionen auf dem Definitionsbereich von $\boldsymbol{W}$ mit den Koordinaten $a_{m,r}, b_{m,\overline{m},r} \in \mathbb{R}$ ($r = 0,\ldots,\infty$) dargestellt werden in der Form

$$\mu_m = \sum_{r=0}^{\infty} a_{m,r} v_r, \quad \tilde{\sigma}_{m,\overline{m}} = \sum_{r=0}^{\infty} b_{m,\overline{m},r} v_r, \tag{11}$$

wobei $\tilde{\sigma}_{m,\overline{m}}$ die Elemente der inversen Kovarianzmatrix $\boldsymbol{\Sigma}^{-1}$ sind.

Anstatt der angegebenen exakten Darstellung werden die Funktionen durch eine Untermenge $\{v_r\}_{r=0,\ldots,L-1}$ der vollständigen Basismenge approximiert. Die Theorie der Taylorentwicklungen zeigt, dass der Approximationsfehler prinzipiell durch großes L beliebig verkleinert werden kann. In der Praxis beschränkt neben der Rechenzeit für die Berechnung der Dichte allerdings vor allem die Größe der Trainingsmenge die Zahl der schätzbaren Parameter. Diese Form der Basisdarstellung erlaubt eine schnelle Berechnung der Dichtefunktion und eine Maximum-Likelihood-Schätzung der Basiskoeffizienten (siehe dazu allgemein [20]). Die Schätzung ergibt sogar analytische Terme für die Koeffizienten, wenn $\tilde{\sigma}_{m,\overline{m}}$ als konstant angenommen wird (siehe Abschnitt 6.1). Bezeichnet man die Zahl der Basisfunktionen für die Elemente der Kovarianzmatrix mit L_σ und die des Mittelwertvektors mit L_μ, dann gilt in diesem Fall $L_\sigma = 1$. Für den Mittelwertvektor ist im 2D-Fall ebenfalls $L_\mu = 1$ und ansonsten $L_\mu > 1$ (siehe Abschnitt 7.3).

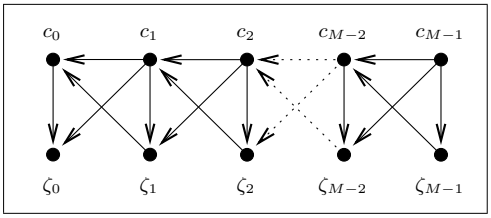


Abb. 7. Abhängigkeitsstruktur der Mischungsverteilung

In Bezug auf die Abhängigkeitsstruktur der lokalen Merkmale wird in den Experimenten in Abschnitt 7 die Modellierung unabhängiger Merkmale und die Berücksichtigung von Spaltenabhängigkeiten untersucht. In beiden Fällen ist die inverse Kovarianzmatrix nur dünn besetzt, sodass eine schnelle Berechnung der Dichtewerte möglich ist.

5.3 Hintergrundmodell

Im vorigen Abschnitt wurde eine statistische Dichtefunktion eines einzelnen Objekts definiert. Implizit ist darin bereits ein Hintergrundmodell enthalten, da zum einen die Merkmale des Objekts durch ein Objektfenster abgetrennt werden und zum anderen auch die Merkmale innerhalb des Objektfensters bei Vorliegen ausreichenden Trainingsmaterials als Hintergrundmerkmale trainiert werden können. In diesem Abschnitt wird der Hintergrund explizit in die Modellierung mit aufgenommen. Dabei wird in diesem Artikel ein beliebiger Hintergrund angenommen. Diese Form des Hintergrunds liegt vor, wenn mehrere Objekte oder allgemein „Struktur“ im Hintergrund vorhanden ist, ohne die genaue Position oder den Typ zu kennen. Zu anderen Hintergrundtypen siehe [19].

Die einzelnen Merkmale des Bildes werden je nach Zugehörigkeit zu Objekt oder Hintergrund in zwei Klassen eingeteilt. Bezeichne Ω_1 die Objekt- und Ω_0 die Hintergrundklasse. Von der Hintergrundklasse wird angenommen, dass sie dieselbe Verteilung an jeder Bildposition aufweist und deshalb keine Positionsparameter besitzt. Es wird angenommen, dass jede Bildposition entweder zum Hintergrund oder zum Objekt gehört, wobei die Zugehörigkeit einer Bildposition im Allgemeinen für ein vorgegebenes Bild allerdings nicht bekannt ist. Sie wird deshalb durch eine Funktion modelliert. Sei $\zeta : X \to \{0,1\}$ die Zuordnungsfunktion, welche die verborgene Information beinhaltet, zu welcher Klasse $\Omega_{\zeta(\boldsymbol{x}_m)}$ die Position $\boldsymbol{x}_m$ gehört. Zur Vereinfachung der Notation wird $\zeta_m = \zeta(m) := \zeta(\boldsymbol{x}_m)$ geschrieben. Wie später noch gezeigt wird, lässt sich aus dem bisherigen Einzelobjektmodell und einem einfachen Modell für Hintergrundmerkmale die Verbundwahrscheinlichkeit $p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\phi,\boldsymbol{t})$ von Bildmerkmalen und Zuordnung ableiten, die durch die Objektlage bedingt ist. Die Mischungsverteilung der Bildmerkmale ergibt sich daraus zu

$$p(\boldsymbol{c}|\boldsymbol{B},\phi,\boldsymbol{t}) = \sum_{\boldsymbol{\zeta}} p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\phi,\boldsymbol{t}) \tag{12}$$

mit $\boldsymbol{\zeta} = (\zeta(m))_{\boldsymbol{x}_m \in X}$. Da die Berücksichtigung lokaler Abhängigkeiten eine Verallgemeinerung der Unabhängigkeitsannahme darstellt, sind im Allgemeinen bessere Ergebnisse zu erwarten, wenn die Trainingsmenge ausreichend

groß ist (siehe dazu auch [20]). Aus diesem Grund werden die folgenden Überlegungen für Spaltenabhängigkeiten der lokalen Merkmale und Unabhängigkeit der lokalen Zuordnungen $\zeta(m)$ ausgeführt (siehe Abb. 7). Die Theorie lässt sich einfach auf beliebige Nachbarschaftssysteme erweitern (siehe [19]). Bei Spaltenabhängigkeiten werden nur Abhängigkeiten von Merkmalen betrachtet, die in einer Spalte untereinander liegen. Merkmale unterschiedlicher Spalten werden als unabhängig angenommen. Für die Merkmale einer Spalte wird jeweils eine Abhängigkeit nur vom unmittelbaren Vorgänger (also dem Merkmal, das „oberhalb“ liegt), seiner Zuordnung und der eigenen Zuordnung angenommen, sodass sich die in Abb. 7 angegebene Abhängigkeitsstruktur ergibt.

Die Dichte der Objektklasse Ω_1 kann direkt aus dem Einzelobjektmodell (siehe Gleichung (9)) übernommen werden. Die Hintergrunddichte (Hintergrundklasse Ω_0) wird aus unabhängigen und für jede Position identischen Komponenten $p(c_m|\Omega_0, \boldsymbol{B}, \phi, \boldsymbol{t}) = p(c_m|\Omega_0, \boldsymbol{B})$ zusammengesetzt:

$$p(\boldsymbol{c}|\Omega_0, \boldsymbol{B}, \phi, \boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in X} p(c_m|\Omega_0, \boldsymbol{B}). \tag{13}$$

Seien die Merkmalspositionen m bezüglich ihrer Spaltenabhängigkeit geordnet. Dann ergibt sich für die Terme $p(\boldsymbol{c}, \boldsymbol{\zeta}|\boldsymbol{B}, \phi, \boldsymbol{t}) = p(\boldsymbol{c}, \boldsymbol{\zeta})$ der Mischungsverteilung nach Gleichung (12), bei denen zur Vereinfachung der Schreibweise der Bedingungsteil weggelassen wird:

$$\begin{aligned} p(\boldsymbol{c}, \boldsymbol{\zeta}) &= p(\boldsymbol{c}_0, \boldsymbol{\zeta}_0) \prod_{m=1,\ldots,M-1} p(\boldsymbol{c}_m, \zeta_m|\boldsymbol{c}_{m-1}, \zeta_{m-1}) \\ &= \left(\prod_{m=0,\ldots,M-1} p(\zeta_m) \right) \\ &\quad \left(p(\boldsymbol{c}_0|\zeta_0) \prod_{m=1,\ldots,M-1} \frac{p(\boldsymbol{c}_{m-1}, \boldsymbol{c}_m|\zeta_{m-1}, \zeta_m)}{p(\boldsymbol{c}_{m-1}|\zeta_{m-1})} \right). \end{aligned} \tag{14}$$

Wenn zwei Nachbarpositionen derselben Klasse zugeordnet werden, ist der Term $p(c_m, c_{m-1}|\zeta(m), \zeta(m-1))$ gleich $p(c_m, c_{m-1}|\Omega_{\zeta(m)})$. Andernfalls wird er als $p(c_m|\Omega_{\zeta(m)})p(c_{m-1}|\Omega_{\zeta(m-1)})$ angenommen und ergibt sich damit auch direkt aus den Größen des Einzelobjektmodells.

5.4 Mehrobjektmodell

Die Modellierung der Kombination von Hintergrund und Objekt durch eine Zuordnungsfunktion lässt sich in analoger Weise auf mehrere Objekte erweitern. Hierzu kann eine Zuordnungsfunktion definiert werden, die jede Merkmalsposition entweder dem Hintergrund oder einer von mehreren Objektklassen zuordnet. Die Berechnung der Dichte lässt sich analog dem Hintergrundmodell durchführen. Siehe dazu im Detail [19].

6 Erlernen und Erkennen von Objekten

6.1 Training der Modellparameter

Für das Training der Modellparameter steht eine Menge $\{^{\rho}\boldsymbol{f}\}$ von Objektaufnahmen $^{\rho}\boldsymbol{f}$ mit $\rho = 0, \ldots, N_{\rho} - 1$ zur Verfügung, die das Objekt in unterschiedlichen Lagen, Beleuchtungsverhältnissen und mit unterschiedlichem Hintergrund zeigen. Die daraus abgeleitete Menge $\{^{\rho}\boldsymbol{c}\}$ von Merkmalsvektoren wird zur Parameterschätzung verwendet. Eine grundlegende Annahme dabei ist, dass die Einzelbeobachtungen unabhängig voneinander sind, sodass sich mit den Modellparametern $\boldsymbol{B}$ und den beobachtungsspezifischen Lageparametern $^{\rho}\boldsymbol{a} = (^{\rho}\boldsymbol{R}, ^{\rho}\boldsymbol{t})$ aus der Dichte $p(\boldsymbol{c}|\boldsymbol{B}, \boldsymbol{a})$ einer Beobachtung die Gesamtdichte

$$p\left((^{\rho}\boldsymbol{c})_{\rho=0,\ldots,N_{\rho}-1}|\boldsymbol{B}, (^{\rho}\boldsymbol{a})_{\rho=0,\ldots,N_{\rho}-1}\right) = \prod_{\rho=0,\ldots,N_{\rho}-1} p(^{\rho}\boldsymbol{c}|\boldsymbol{B}, ^{\rho}\boldsymbol{a}) \tag{15}$$

ergibt. Sind die Objektlageparameter $^{\rho}\boldsymbol{a}$ für jede Beobachtung bekannt, so ist die ML-Schätzung $\widehat{\boldsymbol{B}}$ der Modellparameter gleich:

$$\widehat{\boldsymbol{B}} = \underset{\boldsymbol{B}}{\operatorname{argmax}} \prod_{\rho=0,\ldots,N_{\rho}-1} p(^{\rho}\boldsymbol{c}|\boldsymbol{B}, ^{\rho}\boldsymbol{a}). \tag{16}$$

Generell stellt sich hier bei der Parameterschätzung die Frage, wieviele Trainingsaufnahmen notwendig sind, um die Parameter sicher schätzen zu können. Die Schätztheorie kann zumindest einen Hinweis auf eine vernünftige Zahl von Aufnahmen in Form von Konfidenzbereichen liefern. So lassen sich etwa Mittelwert und Streuung einer einzelnen Normalverteilung mit einer Sicherheit von 90% bei einer maximalen Abweichung von 30% in Bezug auf die Streuung schätzen, wenn mindestens circa 40 Aufnahmen vorliegen. 80 Aufnahmen reduzieren die maximale Abweichung bei gleicher Sicherheit auf etwas über 20%. Werden alle Merkmale als unabhängig angenommen, so liefert diese Form der Abschätzung für zweidimensionale Objekte zumindest einen Anhaltspunkt für eine brauchbare Kardinalität der Trainingsmenge. Liegen mehr Freiheitsgrade, das heißt Parameter im Verhältnis zu den Merkmalspositionen, etwa aufgrund einfacher Abhängigkeitsannahmen oder aufgrund der Berücksichtigung mehrerer Objektansichten vor, so ist eine analytische Behandlung der Konfidenzbereiche nicht mehr so leicht möglich. Als einfache Abschätzung bietet sich aber eine Vervielfachung der Trainingsmenge um die relative Zahl der zusätzlichen Parameter an. Es sei an dieser Stelle außerdem angemerkt, dass nicht nur eine ausreichend große Zahl von Trainingsaufnahmen, sondern auch eine sinnvolle Überdeckung aller möglichen Objektansichten notwendig ist. Letztlich sind aber insbesondere hier für die Beurteilung die praktischen Ergebnisse relevant. Denn bereits bei der Bewertung der Zahl der Trainingsaufnahmen für die Güte der Schätzung ist die Bedeutung einer 20- oder 30- prozentigen Abweichung der Parameter in Bezug auf die Erkennungsergebnisse aufgrund der Komplexität der involvierten mathematischen Funktionen nicht mehr analytisch darstellbar und aus der Theorie heraus einfach berechenbar.

Für die Einzelobjektdichte in Form einer parametrisierten Normalverteilung nach Gleichung (9) und (11) sind die Parameter $a_{m,r}$ und $b_{m,\overline{m},r}$ zu schätzen. Sei die Kovarianzmatrix der Verteilung zunächst konstant bezüglich der externen Transformation (ϕ_{ext}, $\boldsymbol{t}_{ext}$). Das heißt, dass $L_{\sigma} = 1$ und damit $\boldsymbol{\Sigma}^{-1} = (\tilde{\sigma}_{m,\overline{m}})_{m,\overline{m}} = (b_{m,\overline{m},0})_{m,\overline{m}}$ ist. Dies ist beispielsweise beim 2D-Objektmodell der Fall, kann aber auch für das 3D-Objektmodell sinnvoll sein, wenn die Varianz der Merkmale unabhängig von der externen Transformation ist oder die Abhängigkeit aufgrund einer kleinen

Stichprobenmenge nur unzureichend geschätzt werden kann. Für die Schätzung ergibt sich mit den lageabhängigen Interpolationsmatrizen ${}^{\rho}\boldsymbol{W} = \boldsymbol{W}({}^{\rho}\phi_z, {}^{\rho}\boldsymbol{t}_{2D})$ (siehe Gleichung (10)), der Parametermatrix $\boldsymbol{A} = (a_{m,r})_{m,r}$ und dem Vektor ${}^{\rho}\boldsymbol{v} = (v_r({}^{\rho}\phi_{ext}, {}^{\rho}\boldsymbol{t}_{ext}))_{r=0,\ldots,L_\mu-1}^{\mathrm{T}}$ der Basisfunktionswerte aus Gleichung (16), (9) und (11):

$$\begin{aligned}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Sigma}}) &= \underset{(\boldsymbol{A},\boldsymbol{\Sigma})}{\operatorname{argmax}} \prod_{\rho=0,\ldots,N_\rho-1} \mathcal{N}\left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} |\ \boldsymbol{A}{}^{\rho}\boldsymbol{v}, \boldsymbol{\Sigma}\right) \\ &= \underset{(\boldsymbol{A},\boldsymbol{\Sigma})}{\operatorname{argmin}} N_\rho \log\left(|\det 2\pi\boldsymbol{\Sigma}|\right) \\ &\quad + \sum_{\rho=0,\ldots,N_\rho-1} \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \\ &\qquad \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right)\end{aligned} \tag{17}$$

Wegen der Beschränkung auf Unabhängigkeit beziehungsweise Spaltenabhängigkeiten bei den Merkmalskomponenten (siehe Abschnitt 5.2) ist die inverse Kovarianzmatrix nur dünn besetzt, wie in [19] nachgewiesen wird. Für Spaltenabhängigkeiten sind beispielsweise alle Einträge außer der Haupt- und der ersten Nebendiagonale gleich Null. Die Nulleinträge der inversen Kovarianzmatrix sind die Nebenbedingungen, unter denen die Maximierung durchzuführen ist. Sei $\boldsymbol{\Lambda}$ eine Matrix mit Lagrange-Multiplikatoren $\lambda_{m,\overline{m}} \in \mathbb{R}$ an genau den Positionen, an denen die inverse Kovarianzmatrix Nulleinträge hat. An den übrigen Stellen sei $\boldsymbol{\Lambda}$ mit Nullen besetzt. Dann ergibt sich mit der Ableitung der zu optimierenden Funktion

$$\begin{aligned}L(\boldsymbol{A}, \boldsymbol{\Sigma}) &= N_\rho \log\left(|\det 2\pi\boldsymbol{\Sigma}|\right) \\ &\quad + \sum_{\rho=0,\ldots,N_\rho-1} \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right) \boldsymbol{\Sigma}^{-1} \\ &\qquad \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right)\end{aligned} \tag{18}$$

nach den Schätzgrößen:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{A}}(\boldsymbol{A}, \boldsymbol{\Sigma}) &= \sum_{\rho=0,\ldots,N_\rho-1} -2\boldsymbol{\Sigma}^{-1}\left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right) {}^{\rho}\boldsymbol{v}^{\mathrm{T}} \overset{!}{=} \boldsymbol{0} \\ \frac{\partial L}{\partial \boldsymbol{\Sigma}^{-1}}(\boldsymbol{A}, \boldsymbol{\Sigma}) &= -N_\rho\boldsymbol{\Sigma} + \sum_{\rho=0,\ldots,N_\rho-1} \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right) \\ &\qquad \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \boldsymbol{A}{}^{\rho}\boldsymbol{v}\right)^{\mathrm{T}} \overset{!}{=} \boldsymbol{\Lambda}.\end{aligned} \tag{19}$$

Die Schätzterme sind deshalb

$$\boxed{\begin{aligned}\widehat{\boldsymbol{A}} &= \left(\sum_\rho {}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c}{}^{\rho}\boldsymbol{v}^{\mathrm{T}}\right)\left(\sum_\rho {}^{\rho}\boldsymbol{v}{}^{\rho}\boldsymbol{v}^{\mathrm{T}}\right)^{-1} \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{N_\rho}\sum_\rho \left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \widehat{\boldsymbol{A}}\,{}^{\rho}\boldsymbol{v}\right)\left({}^{\rho}\boldsymbol{W}{}^{\rho}\boldsymbol{c} - \widehat{\boldsymbol{A}}\,{}^{\rho}\boldsymbol{v}\right)^{\mathrm{T}} + \tilde{\boldsymbol{\Lambda}},\end{aligned}} \tag{20}$$

mit $\tilde{\boldsymbol{\Lambda}} = \frac{-1}{N_\rho}\boldsymbol{\Lambda}$.

In Bezug auf die Kovarianzmatrix sind nur die Einträge $\sigma_{m,\overline{m}}$ der Kovarianzmatrix aus der Schätzung zu bestimmen, für welche die inverse Matrix keine Nulleinträge aufweisen soll ($\tilde{\sigma}_{m,\overline{m}} \neq 0$). Dies ist eine Folge der additiven Lagrange-Matrix $\tilde{\boldsymbol{\Lambda}}$, aufgrund derer die übrigen Einträge $\sigma_{m,\overline{m}}$ so zu wählen sind, dass die Nebenbedingung von Nulleinträgen $\tilde{\sigma}_{m,\overline{m}} = 0$ in der Inversen erfüllt ist.

Bei der Berechnung der Normalverteilungsdichte nach Gleichung (9) (siehe auch Gleichung (17)) wird die inverse Kovarianzmatrix benötigt. Anstatt das Gleichungssystem (20) unter den gegebenen Nebenbedingungen für $\boldsymbol{\Sigma}$ zu lösen und die numerisch problematische Invertierung durchzuführen, können die Einträge der dünn besetzten Inversen direkt berechnet werden. Dazu werden zunächst die Kovarianzschätzterme

$$\sigma_{m,\overline{m}} = \frac{1}{N_\rho}\sum_\rho \left({}^{\rho}\tilde{c}_m - \widehat{a}_m\right)\left({}^{\rho}\tilde{c}_{\overline{m}} - \widehat{a}_{\overline{m}}\right) \tag{21}$$

nach Gleichung (20) für alle $m, \overline{m}$ berechnet, an denen die inverse Kovarianzmatrix Nulleinträge $\tilde{\sigma}_{m,\overline{m}} = 0$ aufweist. Daraus lassen sich bei lokalen Abhängigkeitsstrukturen die inversen Einträge durch eine einfache Berechnungsvorschrift ableiten (siehe [19]).

Ist die Kovarianzmatrix in Bezug auf die externe Transformation keine Konstante, so können keine analytischen Schätzterme für die Dichteparameter angegeben werden. Numerische Optimierungsverfahren ermöglichen aber auch in diesen Fällen das Auffinden lokaler Maxima der Schätzfunktion, deren Optimalität bei Unabhängigkeit der lokalen Merkmalsvektoren nachgewiesen werden kann (siehe [19]).

Im Falle des Hintergrundmodells (siehe Gleichung (14)) sind neben den Einzelobjektdichten die a-priori-Zuordnungswahrscheinlichkeiten $p(\zeta(m)|\boldsymbol{B}, \phi, \boldsymbol{t})$ (Terme des ersten Produkts in der zweiten Darstellung in Gleichung (14)) zu schätzen. Eine erwartungstreue Schätzung ergibt sich bei Annahme der Unabhängigkeit von der externen Transformation und bei Unabhängigkeitsannahme der Merkmale über

$$\begin{aligned}\widehat{p}(\zeta_m = 1|({}^{\rho}\boldsymbol{c})_\rho) &= \frac{1}{N_\rho}\sum_\rho p(\zeta_m = 1|{}^{\rho}\boldsymbol{c}_m) \\ &= \frac{1}{N_\rho}\sum_\rho \frac{p({}^{\rho}\boldsymbol{c}_m|\kappa = 1)}{p({}^{\rho}\boldsymbol{c}_m|\kappa = 0) + p({}^{\rho}\boldsymbol{c}_m|\kappa = 1)}.\end{aligned} \tag{22}$$

Dieser Schätzterm lässt sich für den Fall abhängiger Merkmale geeignet verallgemeinern [19]. Dies wird in diesem Artikel nicht weiter ausgeführt, da in den hier vorgestellten Experimenten die a-priori-Zuordnungswahrscheinlichkeiten als gleichverteilt angenommen werden.

6.2 Lokalisation von Objekten

6.2.1 Einzelobjektmodell

Die Lokalisation einzelner Objekte erfolgt durch eine ML-Schätzung nach Gleichung (8). Dabei wird angenommen, dass sich genau ein Objekt mit bekannten Modellparametern im Bild befindet. Ziel ist die Bestimmung der Lageparameter $(\phi, \boldsymbol{t})$ des Objekts. Die Lösungen der Schätzung können im Allgemeinen nicht analytisch bestimmt werden. Ist bereits ein guter Schätzwert – beispielsweise aufgrund einer hierarchisch strukturierten Suche – bekannt, so eignen sich numerische lokale Suchverfahren wie der Downhill-Simplex-Algorithmus (siehe [23]) zur Auffindung des lokalen Optimums.

In einer ersten Stufe des angewendeten Suchverfahrens ist allerdings eine globale Suche zur Bestimmung des globalen Optimums erforderlich. Die Dichtefunktion (siehe Gleichung (9)) der Beobachtung muss dazu für sehr viele Parameter ausgewertet werden. Die Parameter, für welche die Funktion ausgewertet wird, können entweder deterministisch vorgegeben oder probabilistisch ermittelt werden. In dieser Arbeit wird eine deterministische Gittersuche verwendet. Die spezielle Anordnung der zu durchsuchenden Parameter bei der Gittersuche ermöglicht dabei eine gegenüber vielen Einzelfunktionsauswertungen beschleunigte Berechnung der Dichtewerte. Sei $D_{\phi,\boldsymbol{t}}$ die Zahl der auszuwertenden Transformationsparameter. Dann ist die Zeitkomplexität der Gittersuche von der Ordnung $\mathcal{O}(|A|LD_{\phi,\boldsymbol{t}}$ mit der Zahl L verwendeter Basisfunktionen, wenn nur Merkmalsabhängigkeiten einer begrenzten Nachbarschaft von Merkmalspositionen berücksichtigt werden.

Um die Darstellung zu vereinfachen, werden die folgenden Ableitungen für $\phi_z = 0$ und $\boldsymbol{t}_{ext} = \mathbf{0}$ angegeben und die damit entfallenden Funktionsargumente weggelassen, also beispielsweise $c_m(\boldsymbol{t}_{2D}) := c_m(0, \boldsymbol{t}_{2D})$ gesetzt. Die Verallgemeinerung ergibt sich analog. Zur Lageschätzung wird die Funktion

$$\begin{aligned} &p(\boldsymbol{c}_A|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \\ &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(\frac{-1}{2}\sum_{m,\overline{m}}\tilde{\sigma}_{m,\overline{m}}\,(c_m(\boldsymbol{t}_{2D})-\mu_m(\phi_{ext}))\right. \\ &\qquad\left.(c_{\overline{m}}(\boldsymbol{t}_{2D})-\mu_{\overline{m}}(\phi_{ext}))\right) \end{aligned} \tag{23}$$

nach Gleichung (9) mit $\mu_m(\phi_{ext}) = \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{v}(\phi_{ext})$ und den Merkmalskomponenten $c_m(\boldsymbol{t}_{2D}) = c(\boldsymbol{x}_m + \boldsymbol{t}_{2D})$ bezüglich $\boldsymbol{R},\boldsymbol{t}$ maximiert. Im Folgenden wird angenommen, dass die Kovarianzmatrix unabhängig von der externen Transformation ist, um die Darstellung zu vereinfachen. Eine analoge Ableitung ist bei variabler Kovarianzmatrix möglich. Durch Anwendung des Logarithmus ergibt sich folgende zu minimierende Funktion:

$$\begin{aligned} h(\phi,\boldsymbol{t}) = \sum_{m,\overline{m}}\tilde{\sigma}_{m,\overline{m}}\,&\left(c_m(\boldsymbol{t}_{2D}) - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{v}(\phi_{ext})\right) \\ &\left(c_{\overline{m}}(\boldsymbol{t}_{2D}) - \boldsymbol{a}_{\overline{m}}^{\mathrm{T}}\boldsymbol{v}(\phi_{ext})\right). \end{aligned} \tag{24}$$

Mit $\phi = (\phi_x,\phi_y,\phi_z) = (\phi_x,\phi_y,0)$ und den Funktionen

$$h_1(\phi,\boldsymbol{t}) = \sum_{m,\overline{m}} c_m(\boldsymbol{t}_{2D})c_{\overline{m}}(\boldsymbol{t}_{2D})\tilde{\sigma}_{m,\overline{m}} \tag{25}$$

$$h_{2,r}(\phi,\boldsymbol{t}) = \sum_{m,\overline{m}} c_m(\boldsymbol{t}_{2D})a_{\overline{m},r}\tilde{\sigma}_{m,\overline{m}} \tag{26}$$

$$h_3(\phi,\boldsymbol{t}) = \sum_{m,\overline{m}} \left(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{v}(\phi_{ext})\right)\left(\boldsymbol{a}_{\overline{m}}^{\mathrm{T}}\boldsymbol{v}(\phi_{ext})\right)\tilde{\sigma}_{m,\overline{m}}, \tag{27}$$

ist die Summe

$$(h_1 - 2\boldsymbol{v}(\phi_{ext})^{\mathrm{T}}h_2 + h_3)(\phi,\boldsymbol{t}) \tag{28}$$

zu minimieren. Bei der globalen Suche werden die Funktionswerte an allen Gitterpositionen $(\phi,\boldsymbol{t}) = (\phi_{i,j},\boldsymbol{t}_{2D,k,l})$ des Suchgitters (siehe dazu auch nachfolgende Definition) ausgewertet, was zu der angegebenen Komplexität führt.

Die Funktionen h_1 und $h_{2,r}$ haben beide die Form

$$\tilde{h}(\boldsymbol{t}_{2D}) = \sum_{m,\overline{m}} f(\boldsymbol{x}_m + \boldsymbol{t}_{2D}, \boldsymbol{x}_{\overline{m}} + \boldsymbol{t}_{2D})w_{m,\overline{m}} \tag{29}$$

für ein festes ϕ. Dabei ist für h_1 beispielsweise $f(\boldsymbol{x}_m + \boldsymbol{t}_{2D}, \boldsymbol{x}_{\overline{m}} + \boldsymbol{t}_{2D}) = c_m(\boldsymbol{t}_{2D})c_{\overline{m}}(\boldsymbol{t}_{2D})$ und $w_{m,\overline{m}} = \tilde{\sigma}_{m,\overline{m}}$. Aufgrund der dünn besetzten inversen Kovarianzmatrix bei einfachen Abhängigkeitsstrukturen sind die meisten Faktoren $w_{m,\overline{m}}$ gleich Null. Seien die Vorgänger des Abhängigkeitsnetzes gleichförmig auf dem ganzen Netz durch die Menge S der relativ abhängigen Positionen definiert, sodass $\overline{\mathcal{P}}(\boldsymbol{x}_m) = \{\boldsymbol{x}_{(k,l)-s}|s \in S\}$ die Vorgänger der Position $\boldsymbol{x}_m = \boldsymbol{x}_{k,l}$ sind. Diese Menge ist bei Unabhängigkeitsannahme gleich $S = \emptyset$ und bei Spaltenabhängigkeiten gleich $S = \{(0,1)\}$. Dann kann obige Gleichung umgeschrieben werden:

$$\tilde{h}(\boldsymbol{t}_{2D}) = \sum_{s\in S_0}\sum_m \bar{f}_s(\boldsymbol{x}_m + \boldsymbol{t}_{2D})w_{m,s(m)}, \tag{30}$$

mit $s(\boldsymbol{x}_m) = \boldsymbol{x}_{(k,l)-s}$, $\bar{f}_s(\boldsymbol{x}_m) = f(\boldsymbol{x}_m, s(\boldsymbol{x}_m))$ und $S_0 = S\cup\{(0,0)\}$. Die Summation ist dabei für den gesamten gültigen Bereich der Nachbarschaften durchzuführen. Wenn das Auswertungsgitter $\{(\phi_{i,j},\boldsymbol{t}_{2D,k,l})\}$ der Transformationsparameter $(\phi,\boldsymbol{t}_{2D}) \in \{(\phi_{i,j},\boldsymbol{t}_{2D,k,l})\}$ als Erweiterung des Gitters X auf den möglichen Parameterbereich gewählt wird, sodass

$$\phi_{i,j} = (\phi_{x,0} + i\Delta\phi_x, \phi_{y,0} + j\Delta\phi_y, 0) = (\phi_{x,i}, \phi_{y,j}, 0) \tag{31}$$

$$\boldsymbol{t}_{2D,k,l} = -(t_{x,0} + k\Delta t_x, t_{y,0} + l\Delta t_y)^{\mathrm{T}} = -(t'_{x,k}, t'_{y,l}), \tag{32}$$

kann die zweite Summe über m in (30), beziehungsweise über dem Wertebereich X von $\boldsymbol{x}_m$, als Faltung interpretiert werden. Die Schrittweite der Auswertung auf dem Translationsraum muss dazu dem Gitterabstand $\Delta x_s = \Delta y_s = r_s$ der Merkmalspositionen entsprechen, sodass $\Delta t_x = \Delta t_y = r_s$. Entsprechendes gilt für die Basiswerte der Translation, womit $t_{x,0} = t_{y,0} = 0$. Das Suchgitter der Winkelpositionen kann beliebig gewählt werden. In den Experimenten wird der Winkelabstand der Auswertung auf zirka 10° gesetzt.

Das sich ergebende Auswertungsschema ist in Abb. 8 dargestellt. Es wird zur globalen Suche in den Experimenten in Abschnitt 7 eingesetzt. Die Linearkombination, die am Schluss des Auswertungsschemas durchzuführen ist, ergibt sich durch die Anwendung von (28) auf die einzelnen zum Index r von $h_{2,r}$ gehörigen Basisfunktionen. Die Verwendung der FFT erlaubt die Berechnung der Faltung von Gleichung (30) und damit die Berechnung jeder der $L + 1$ Funktionen h_1 und $h_{2,r}$ mit einem Zeitaufwand der Ordnung $\mathcal{O}(D_t \log(D_{t_x})\log(D_{t_y}))$. Der Zeitaufwand für h_3 ist von der Ordnung $\mathcal{O}\left(D_{\phi_{ext}}\right)$. Dies ergibt für die gesamte Suche eine Komplexität der Ordnung $\mathcal{O}\left(LD_{\phi,\boldsymbol{t}}\log(D_{t_x})\log(D_{t_y})\right)$, wobei die Berechnung von h aus den einfachen Funktionen sehr schnell durchgeführt werden kann. Die Suche durch Auswertung jeder einzelnen Dichte hat im Vergleich dazu – wie bereits angegeben – eine Komplexität von $\mathcal{O}\left(|A|\,LD_{\phi,\boldsymbol{t}}\right)$.

6.2.2 Hintergrundmodell

Wird eine Hintergrundmodellierung, wie in Abschnitt 5.3 beschrieben, vorgenommen, so kann die Objektlokalisation ebenso wie beim Einzelobjektmodell über eine ML-Schätzung nach Gleichung (8) erfolgen. Problematisch an

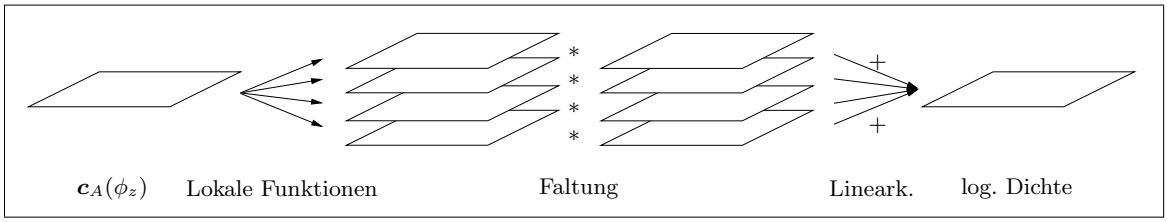


Abb. 8. Schnelle Gittersuche zur Einzelobjektlokalisation durch Ausnutzung der Redundanzen in der Berechnung mehrerer Gitterwerte: Der aus einem Bild berechnete und entsprechend der internen Rotation interpolierte Merkmalsvektor wird durch lokale Funktionen, die jeweils nur von benachbarten lokalen Merkmalsvektoren abhängig sind, in mehrere unterschiedliche Ebenen lokaler Vektoren transformiert. Auf jeder Ebene wird eine schnelle Faltung (FFT) mit Koeffizienten aus dem Objektmodell durchgeführt. Die lokalen Vektoren der verschiedenen Ebenen werden anschließend in einer Linearkombination, deren Gewichtsterme von der externen Rotation abhängig, aber für alle Positionen die gleichen sind, mit den Dichtewerten abhängig von der internen Translation verknüpft

einer direkten ML-Schätzung ist allerdings der Aufbau der Dichte nach Gleichung (12) und ihrer speziellen Realisierungen nach (14). Die Darstellung als Summe von Produkten lokal unterschiedlicher Funktionen führt dazu, dass die Einzelauswertungen der Dichtefunktion numerisch problematischer sind als beim Einzelobjektmodell und keine analoge schnelle Durchführung der globalen Suche möglich ist.

Deshalb wird im Folgenden zusätzlich eine alternative Schätzung, die der EM-Schätzung (siehe [6]) ähnlich ist, angegeben. Dazu wird bei der Kullback-Leibler-Statistik auch der Erwartungswert zum aktuellen Schätzwert berechnet und damit

$$\left(\widehat{\boldsymbol{R}},\widehat{\boldsymbol{t}}\right) = \underset{\boldsymbol{R},\boldsymbol{t}}{\operatorname{argmax}}\, \mathcal{E}_{\boldsymbol{\zeta}}\left(\log p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})|\boldsymbol{c},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})\right) \quad (33)$$

geschätzt. Dabei bezeichnet $\mathcal{E}_{\boldsymbol{\zeta}}$ den Erwartungswert bezüglich $\boldsymbol{\zeta}$ (der Erwartungswert einer Funktion f bezüglich einer diskreten Größe x mit Wahrscheinlichkeitsverteilung p(x) ist definiert als $\mathcal{E}_x(f) = \sum_x f(x)p(x)$). Die EM-Schätzung berechnet an den Übergängen vom ersten zum zweiten Schätzschritt identische Terme. Trotzdem sind die Schätzungen nicht äquivalent, da die Maximierung für eine andere Funktion durchgeführt wird, womit das EM-Optimum nicht notwendig erreicht wird. Der Vorteil der Vorgehensweise nach (33) liegt jedoch darin, dass der zum aktuellen Schätzwert des Transformationsparametersatzes berechnete Erwartungswert jeweils die aktuelle Schätzung der Zuordnung berücksichtigt und nicht wie beim Standard-EM-Ansatz die Schätzung nach dem jeweils letzten Parametersatz der Iteration, die vor allem am Anfang der Suche von einem Iterationsschritt zum nächsten stark abweichen kann und initial aufgrund der beobachtungsunabhängig vorbelegten Zuordnungswahrscheinlichkeiten im Prinzip nur eine leicht modifizierte Einzelobjektsuche durchführt. Nachteilig ist zwar der erhöhte Rechenaufwand für jede einzelne Funktionsauswertung, da die Zuordnung bei jeder Auswertung zu berechnen ist. Dafür ist aber aufgrund der Aktualität der berechneten Zuordnungswahrscheinlichkeit und damit der größeren Ähnlichkeit der Gütefunktion zur ML-Gütefunktion eine schnellere Konvergenz zu erwarten. Der Nachteil der nicht garantierten Lokalisation des EM- oder ML-Optimums kann durch eine anschließende lokale ML-Suche kompensiert werden. Der Vorteil gegenüber der ML-Schätzung liegt darin, dass sich bei Abhängigkeit auf Merkmalsebene und Unabhängigkeit auf Zuordnungsebene ein schnelles globales Suchverfahren ableiten lässt, das bei der ML-Schätzung nur im Falle unabhängiger Merkmale und Zuordnungen möglich ist.

Dieses globale Suchverfahren wird im Folgenden allgemein vorgestellt. Voraussetzung dafür ist, dass die Gütefunktion q, die zur Suche eingesetzt wird, von der Form

$$q(\phi,\boldsymbol{t}|\boldsymbol{c}) = \sum_m q_m(\phi,\boldsymbol{t}|\boldsymbol{c}_{N(m)}) \quad (34)$$

ist, wobei N eine Menge von relativen Positionen ist, $\boldsymbol{c}_{N(m)} = ((\boldsymbol{c}_{\overline{m}}^{\mathrm{T}})_{\overline{m}-m\in N(m)})^{\mathrm{T}}$ ist und q_m lokale Funktionen sind, die jeweils nur einen lokalen Bereich von Merkmalsvektoren bewerten, der in seiner Struktur jedoch überall gleichförmig durch die Positionsmenge N gegeben ist. Es sei weiterhin angenommen, dass sich viele dieser lokalen Funktionen ähnlich sind.

Um eine schnelle näherungsweise Berechnung der Gesamtdichte zu ermöglichen, ist es sinnvoll, sie durch die Linearkombination

$$q_m \approx \widehat{q}_m = \sum_k d_{m,k} q_k \quad (35)$$

einer kleinen Teilmenge $\{q_k\} \subset \{q_m\}$ zu approximieren. Die Approximation wird in dieser Form durchgeführt, um die Ableitung eines schnellen Suchverfahrens zu ermöglichen. Die Gewichtsterme $d_{m,k} \in \mathbb{R}$ werden dabei so bestimmt, dass der quadratische Approximationsfehler auf der Menge $\{{}^{\rho}\boldsymbol{c}_{\overline{m}}\}$ aller oder zumindest einer großen Zahl lokaler Trainingsvektoren minimal wird:

$$\sum_{\rho}\sum_{\overline{m}} \left(\widehat{q}_m(\boldsymbol{c}_{N(\overline{m})}) - q_m({}^{\rho}\boldsymbol{c}_{N(\overline{m})})\right)^2 \to \min\,. \quad (36)$$

Die Minimierung über alle Trainingsvektoren ist notwendig, da eine Funktionsapproximation durchgeführt werden soll, für welche die jeweils aufgrund ihrer Verteilung eingeschränkten lokalen Trainingsvektoren nicht ausreichend sind. Meist kann allerdings aus Rechenzeitgründen nur eine Teilmenge der Trainingsvektoren verwendet werden.

Zur Auswahl einer geeigneten Teilmenge $\{q_k\}$ von Basisfunktionen wird durch Summation des quadratischen Fehlers (36) über einer Untermenge der zu approximierenden Funktionen eine Gütefunktion definiert, die beginnend bei einer einelementigen Menge $\{q_k\}$ schrittweise um jeweils die Funktion erweitert wird, die den Gesamtapproximationsfehler minimiert.

Sind externe Freiheitsgrade der Transformationsparameter gegeben, so sind die Koeffizienten $d_{m,k}$ keine Konstanten

mehr, sondern Abhängige der externen Parameter. Mit einer Basisdarstellung

$$d_{m,k} = \sum_{r=0}^{L_d-1} d_{m,k,r} v_r \tag{37}$$

durch L_d Basisfunktionen analog der Darstellung der Normalverteilungsparameter nach (11) ergibt sich als zu optimierende Gesamtgütefunktion:

$$\boxed{\sum_{m\in A}\sum_{k} q_k \sum_{r=0}^{L_d-1} d_{m,k,r} v_r.} \tag{38}$$

Das ist aber bei der Berechnung auf einem Translationsgitter wie beim Einzelobjektmodell im vorigen Abschnitt eine Faltung von lokal mit q_k transformierten Dichtewerten und modellabhängigen Koeffizienten. Das Berechnungsschema ist somit auch durch Abb. 8 gegeben. Der Unterschied besteht nur darin, dass die lokalen Funktionen eine andere Form haben und das Ergebnis der Berechnung nicht notwendigerweise eine logarithmierte Dichte ist, sondern beispielsweise ein EM-basierter Schätzterm.

Brauchbare Gütefunktionen lassen sich bei der EM-basierten Vorgehensweise nach Gleichung (33) für einfache Abhängigkeitsstrukturen (siehe Abschnitt 5.3) über den Erwartungswert

$$\begin{aligned}\tilde{q}^{EM}(\boldsymbol{R},\boldsymbol{t}|\boldsymbol{c}) &:= \mathcal{E}_{\zeta}\left(\log p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})|\boldsymbol{c},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}\right)\\ &= \sum_{\zeta}\log p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) p(\boldsymbol{\zeta}|\boldsymbol{c},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})\end{aligned} \tag{39}$$

ableiten. Bei unabhängigen Merkmalen und Zuordnungen führt die ML-Schätzung ebenfalls zu brauchbaren separierbaren Schätztermen.

7 Experimente

7.1 Allgemeines

Um die Brauchbarkeit des vorgestellten Ansatzes bewerten zu können, wurden zu mehreren zwei- und dreidimensionalen Objekten roboterunterstützt Aufnahmeserien generiert. Die Aufnahmeserien zeigen die Objekte jeweils in unterschiedlichen Lagen und werden in disjunkte Trainings- und Testmengen zerlegt. Unterschieden werden aufgrund der unterschiedlichen Objektmodelle außerdem Experimentkategorien für zweidimensionale und dreidimensionale Objekt vor homogenem Hintergrund und Experimente vor heterogenem Hintergrund. Nach dem Training der jeweiligen Modellparameter für eine Kategorie von Experimenten werden Experimente zur Bewertung der Lokalisationsgenauigkeit bei der Lagebestimmung der Objekte und die Erkennungsleistung des Systems durchgeführt.

Bei den Experimenten zur Lagebestimmung wird der Objekttyp jeweils als bekannt vorausgesetzt und mit Hilfe der zugehörigen Dichtefunktion die Position des Objekts geschätzt. Kriterium für die Beurteilung ist hier der Anteil der Objekte, deren Position richtig bestimmt wurde, und beispielsweise die mittlere Abweichung der Schätzung von der korrekten Position für die richtig analysierten Aufnahmen. Ob eine Position richtig ist, wird dabei durch einen Schwellwert der erlaubten Abweichung festgelegt. Die eigentliche Objektposition ist aufgrund von automatisch generierten Aufnahmen oder aufgrund einer manuellen Positionsbestimmung bekannt. Bei den Bildern der Trainingsmenge werden die Objektpositionen benötigt, um die Modellparameter zu schätzen. Bei den Lokalisationsexperimenten auf der Testmenge wird dem System keine Information über die Objektposition vorgegeben. Erst nach Abschluss der Experimente werden die tatsächlichen Objektpositionen zur Bewertung der Ergebnisse eingesetzt.

Bei den Experimenten zur Erkennungsleistung werden zu einer beliebigen Aufnahme zunächst alle Dichtefunktionen bezüglich der Objektlage maximiert. Das heißt, jedes trainierte Objekt wird zunächst im Bild gesucht. Auch hier ist keine Information über eine eventuelle tatsächliche Objektposition vorgegeben. Anschließend wird der Objekttyp als erkannt angegeben, dessen Dichtewert maximal ist (Bayes-Klassifikator bei Annahme gleichwahrscheinlicher Objekte).

Neben der grundsätzlichen Zielsetzung, die Leistung des Systems zu messen, gibt es noch einige weitere Fragestellungen, die durch die Experimente beantwortet werden sollen. Diese beziehen sich auf die Freiheitsgrade des Systems einerseits und seine Robustheit andererseits. In Bezug auf die Freiheitsgrade ist hier insbesondere interessant, welche der vorgestellten Merkmale besonders gut für die Systemleistung geeignet sind und wie sich die einzelnen Modelltypen und die Zahl ihrer Parameter (bspw. die Zahl der Basisfunktionen beim dreidimensionalen Modell) auswirken. Um dies beurteilen zu können, müssen jeweils Experimente mit unterschiedlichen Einstellungen dieser Parameter durchgeführt und verglichen werden. Die Robustheit des Verfahrens wird anhand zweier wichtiger Kriterien validiert: der Empfindlichkeit gegenüber Beleuchtungsschwankungen und den Erkennungsergebnissen bei ähnlich aussehenden Objekten. Zu diesem Zweck enthält die Testmenge Bilder mit größeren Variationen in den Helligkeiten und es wurden ähnliche Objekte in die Experimente aufgenommen.

Ein weiteres wichtiges Kriterium zur praktischen Beurteilung des Verfahrens ist die Rechenzeit zur Lösung der einzelnen Aufgabenstellungen. Entsprechende Zeiten werden zu allen Experimenten angegeben. An dieser Stelle sei darauf hingewiesen, dass für die Experimente eine SGI O2 Workstation (R10000 Prozessor, 150 MHz, 128 MB Hauptspeicher) ohne Spezialhardware eingesetzt wurde.

7.2 2D-Einzelobjekterkennung

Die zehn Objekte der 2D-Experimente sind in Abb. 9 beispielhaft dargestellt. Sie bestehen aus einer Gruppe von fünf relativ ähnlichen Objekten und fünf unterschiedlichen Objekten. Zu jedem Objekt sind sechs Aufnahmesequenzen verfügbar, in denen das Objekt jeweils auf einem Drehteller rotiert vor homogenem Hintergrund in verschiedenen Drehlagen vorliegt (siehe zweite Zeile von Abb. 9). Der Winkelbereich von 0 bis 360° wird dabei durch jeweils 35 Positionen gleichmäßig überdeckt. Die Beleuchtung ist bei jeder Sequenz anders eingestellt, wie in der zweiten Zeile von Abb. 9 bei verschiedenen Drehlagen exemplarisch dargestellt. Für

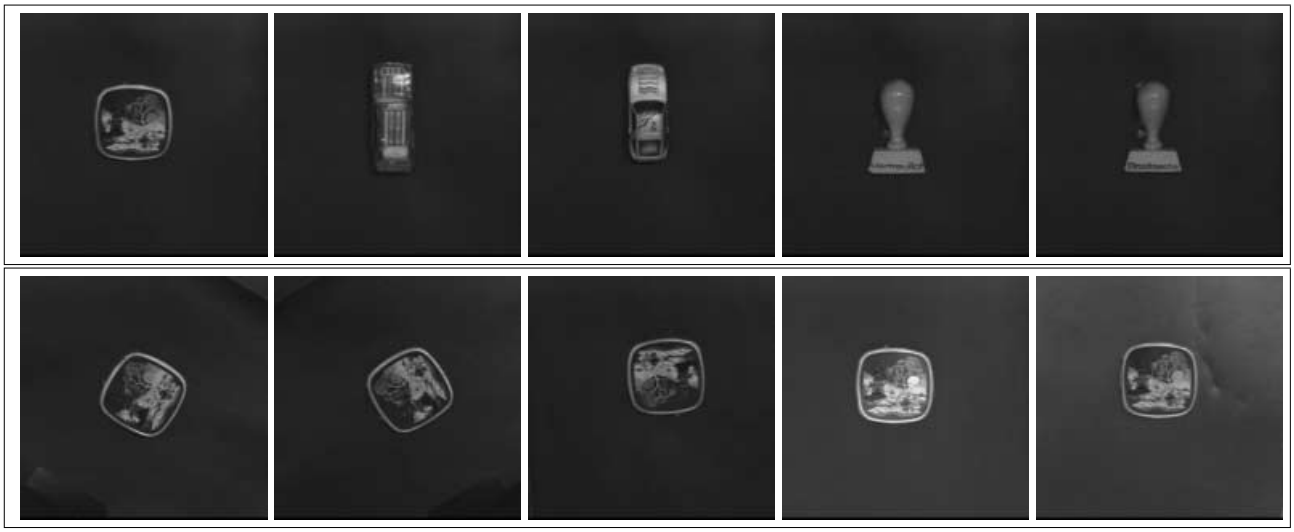


Abb. 9. Beispiele der fünf unterschiedlichen 2D-Objekte (linke drei Bilder der ersten Zeile) und der ähnlichen 2D-Objekte vom Typ *Stempel* (rechte zwei Bilder der ersten Zeile mit den ähnlichsten Objekten *Stempel2* und *Stempel3*. In der zweiten Zeile sind unterschiedliche Rotationslagen und Beleuchtung der 2D-Objekte am Beispiel des Objekts *Deckel* dargestellt. Bei 2D-Objekten ist dabei die Ansicht eines Objekts in allen zugehörigen Aufnahmen die gleiche, es variiert nur die Drehlage und Position des Objekts im Bild

Tabelle 1. Lokalisationsergebnisse der zehn 2D-Objekte (siehe Abb. 9) bei Unabhängigkeitsannahme und Spaltenabhängigkeit der Merkmale. Neben dem prozentualen Anteil der fehlerhaft lokalisierten Objekte und den mittleren und maximalen Abweichungen der geschätzten Positionen von den tatsächlichen Positionen ist der mittlere logarithmierte Dichtewert $\log p_A$ angegeben. Alle Werte beziehen sich auf die Gesamtheit der zehn Objekttests pro Merkmalstyp, bei denen jeweils 35 Testbilder ausgewertet wurden (also auf insgesamt 350 Einzeltests). Abweichungen bei der Translation sind in Bildpunkten, abgekürzt mit „Pix“, angegeben

Merkmale	Ergebnisse bei Unabhängigkeit						Ergebnisse bei Spaltenabhängigkeit					
	Fehler (%)	Abweichung				Dichte	Fehler (%)	Abweichung				Dichte
		Transl. (Pix)		Rotation (°)				Transl. (Pix)		Rotation (°)		
		Mittel	Max	Mittel	Max	$\log p_A$		Mittel	Max	Mittel	Max	$\log p_A$
Wavelet Johnston	0	0.6	2.7	0.7	4.3	−126	0	0.6	3.1	0.7	6.0	374
Wavelet Andrew	0	0.6	2.7	0.7	4.5	−136	0	0.6	3.0	0.7	5.9	369
Wavelet Haar	0	0.6	2.5	0.6	4.5	−145	0	0.6	2.9	0.6	5.4	396
Gabor-Wavelet	0	0.4	2.4	0.6	4.3	−24505	0	0.4	2.2	0.6	3.9	−23723

jedes Objekt werden fünf der vorliegenden Sequenzen (175 Bilder) zum Training eingesetzt und eine Sequenz zum Test (35 Bilder).

Die Ergebnisse der Lokalisationsexperimente auf der zweiten, feinsten Auflösungsebene s_1 (r_{s_1} = 4) sind in Tabelle 1 zusammengefasst. Dabei wird der Translationsparameterraum bei der globalen Suche auf der ersten Ebene gitterförmig entsprechend der Auflösung, also mit einem 8 × 8-Bildpunkt-Gitter, gleichmäßig überdeckt und der Rotationsparameterraum vollständig mit 35 Winkeln. Das Suchgitter des Rotationsbereichs ist so gewählt, dass die Randbereiche der Objekte bei einem Objektdurchmesser von 100 Bildpunkten ebenfalls mit einem Abstand von zirka 8 Bildpunkten überdeckt werden.

Generell ist bei den 2D-Experimenten festzustellen, dass die Position der Objekte immer richtig erkannt wird. Beurteilungskriterium für die Güte des Verfahrens sind somit mittlere und maximale Abweichung, beziehungsweise der mittlere Dichtewert für die jeweils gefundenen Positionen. Die Gabor-Wavelet-Merkmale liefern die besten Ergebnisse. Die Gruppe der diskreten Wavelet-Transformationen ergibt ebenfalls gute Resultate, wobei hier nach einer Vorauswahl nur die besten Wavelets angegeben sind (siehe dazu [19]). Die hier vorgestellten diskreten Wavelets sind aufgrund der ähnlichen Lokalisationsgenauigkeit nur schwer zu vergleichen. Wegen des maximalen mittleren Dichtewerts bei der einfachsten Modellannahme (Unabhängigkeit) wird bei den folgenden Experimenten nur das Johnston-Wavelet weiter ausgewertet. Interessant ist hier aber, dass das einfache Haar-Wavelet ähnliche Ergebnisse liefert und bei Berücksichtigung von Spaltenabhängigkeiten sogar einen größeren maximalen Dichtewert aufweist. Dies zeigt, dass in den Merkmalen mehr Redundanzen enthalten sind als beim Johnston-Wavelet, sich diese aber durch Spaltenabhängigkeiten sogar besser modellieren lassen.

Wie Tabelle 1 zeigt, können allgemein die Ergebnisse durch Berücksichtigung von Abhängigkeiten verbessert werden. Dies spiegelt sich vor allem in den größeren mittleren Dichtewerten auf dem Objektfenster wider. Im übrigen liegen die erreichbaren Genauigkeiten bei der Genauigkeit der manuellen Positionsbestimmung.

Die Klassifikation aller 10 Objekte mit den Merkmalen der Johnston- und Gabor-Wavelets liefert in fast allen Fällen die richtige Zuordnung. Der einzige Fehler tritt bei den einfacheren Johnston-Wavelets beim Objekt *Stempel2* auf, das fälschlicherweise in 12 der 35 Aufnahmen der Klasse

Abb. 10. 3D-Objekte des Typs *Katze*, *Büro* und *Tasse* (erste Zeile). Die Aufnahmen der zweiten Zeile zeigen am 3D-Objekt *Tasse4* beispielhaft den externen Rotations- und Translationsbereich bei den Objekttypen *Büro* und *Tasse*. Die Aufnahmen der dritten Zeile zeigen am 3D-Objekt *Katze1* beispielhaft den externen Rotationsbereich beim Objekttyp *Katze*

Stempel3 zugewiesen wird. Dies liegt daran, dass die beiden Objekte bis auf den Schriftzug fast keine Unterschiede aufweisen und dieser auf der Auflösungsebene s_1 bei den Johnston-Wavelets nicht unterschieden werden kann. Die Gabor-Wavelets erweisen sich an dieser Stelle diskriminativer, da sie keinen Klassifikationsfehler liefern.

Die Rechenzeiten der Lokalisation von 2D-Objekten liegen auf einer SGI O2 (R10000 Prozessor, 150 MHz, 128 MB Hauptspeicher) für die zweidimensionalen Johnston-Merkmale bei Spaltenabhängigkeiten bei zirka zwei bis drei Sekunden und für die Gabor-Merkmale bei vier bis fünf Sekunden.

7.3 3D-Einzelobjekterkennung

Die 15 Objekte der 3D-Experimente sind in Abb. 10 beispielhaft dargestellt. Zehn der Objekte (Objekttypen *Büro* und *Tasse*) sind in je zwei Aufnahmesequenzen verfügbar, in denen die Kameraentfernung in einem Bereich von 37,5 bis 47,5 cm variiert (5 Entfernungen: 37,5, 40, 42,5, 45,0, 47,5 cm, siehe Abb. 10) und das Objekt auf einem Drehteller in 71 Winkelpositionen bezüglich der y-Achse vorliegt. Damit ergeben sich pro Sequenz 355 Aufnahmen. Die Aufnahmen beider Sequenzen und aller Kameraentfernungen werden aufgrund der Rotationslage entweder der Trainings- oder der Testmenge zugeordnet. Bei den gegebenen Winkelpositionen von $n(360^\circ/71)$ ($n = 0, \ldots, 70$) werden dazu die Aufnahmen mit geradzahligem n der Trainings- und die Aufnahmen mit ungeradzahligem n der Testmenge zugewiesen. Damit ergeben sich für jedes Objekt eine Trainingsmenge mit 360 Aufnahmen und eine Testmenge mit 350 Aufnahmen. Die Objektpositionen ergeben sich aus der Roboter- und Drehtellersteuerung.

Fünf der Objekte (Objekttyp *Katze*) sind in drei Aufnahmesequenzen verfügbar, in denen die Kameraposition auf einem Kugelbereich um das Objekt liegt (siehe Abb. 10). Der Positionsbereich auf der Kugel wird dabei gleichmäßig durch ein 15×15-Gitter überdeckt. Da der Roboterarm bei der Aufnahme nicht beliebig drehbar ist, ist der interne Drehwinkel ϕ_z verschieden von Null. Die externen Winkel liegen in einem Bereich von $-26^\circ < \phi_x < 26^\circ$ und $-34^\circ < \phi_y < 34^\circ$. Zwei der Sequenzen werden für das Training eingesetzt (450 Aufnahmen) und eine Sequenz (225 Aufnahmen) zum Test. Die Objektpositionen sind auch hier aufgrund der Robotersteuerung bekannt.

Die Lokalisation und Klassifikation der vorliegenden Objekte wird unter Berücksichtigung von Spaltenabhängigkeiten durchgeführt. Die Zahl der Basisfunktionen ist dabei L_μ = 10 (*Katze*) beziehungsweise L_μ = 21. Das entspricht bei den als Basisfunktionen verwendeten Polynomen einem Monomgrad von 3 beziehungsweise 5 in den Variablen der beiden externen Freiheitsgrade. Für den Objekttyp *Katze* wird eine kleinere Zahl von Basisfunktionen gewählt, da der externe Rotationsbereich kleiner ist. Bei den Objekttypen *Büro* und *Tasse* wird eine Normierung der externen Transformationsfreiheitsgrade auf das Intervall $[0, 1]$ vorgenommen, um numerische Probleme aufgrund der hochgradigen Polynome zu vermeiden, die bei der Berechnung der Dichtefunktion auftreten. Die Kovarianz wird in allen 3D-

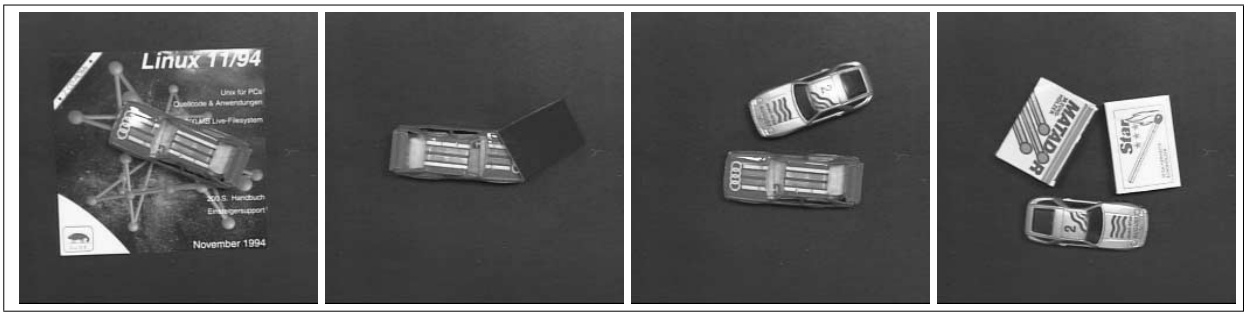


Abb. 11. Aufnahmen zum Test des Hintergrundmodells (von links nach rechts): Objekt mit stark heterogenem Hintergrund, Objekt mit leichten (bis zirka 20%) Verdeckungen, zwei Objekte vor homogenem Hintergrund, drei Objekte vor homogenem Hintergrund

Tabelle 2. Lokalisationsergebnisse für den 3D-Objekttyp *Katze* mit mittleren (∅) und maximalen (Max) Abweichungen. Die Zahl der Merkmalspositionen im Objektfenster liegt abhängig vom Objekt auf der Auf"lösungsebene s_0 im Bereich von 230 bis 300 und auf s_1 im Bereich von 1000 bis 1270

Merkmale	Fehler	Abweichung					
	(%)	t_x, t_y (Pix)		ϕ_z (°)		ϕ_x, ϕ_y (°)	
		∅	Max	∅	Max	∅	Max
Johnston	3	1.7	7	0.6	11	3.0	14
Gabor	0.4	1.7	9.3	0.4	7.9	2.1	12

Tabelle 3. Lokalisationsergebnisse mit Johnston-Merkmalen für die Objekttypen *Büro* und *Tasse*. Die Zahl der Merkmalspositionen im Objektfenster liegt abhängig vom Objekt auf der Auf"lösungsebene s_0 (s_1) im Bereich von 90 bis 420 (360 bis 1720). Eine Lokalisation wird als fehlerhaft eingestuft, wenn die interne Translation um mehr als 10 Bildpunkte oder die Drehlage ϕ_y um mehr als $\phi_{y,\max}$ vom tatsächlichen Wert abweicht

Objekt-	$\phi_{y,\max}$	Fehler	Abweichung					
typ	(°)	(%)	t_x, t_y (Pix)		t_z (cm)		ϕ_y (°)	
			∅	Max	∅	Max	∅	Max
Büro	15	15	1.5	9.9	0.4	3.9	6.0	14
Büro	180	3.6	1.5	9.9	0.4	3.9	7.8	180
Tasse	15	20	1.2	5.9	0.2	1.0	6.1	14
Tasse	180	0	1.2	5.9	0.2	2.6	9.5	170

Experimenten, wenn nicht explizit anders erwähnt, als Konstante modelliert ($L_\sigma = 1$).

Die Ergebnisse der 3D-Lokalisationsexperimente sind in Tabelle 2 und 3 dargestellt. Der Suchraum besteht beim Objekttyp *Katze* im internen Rotations- und Translationsraum, der vollständig durchsucht wird, und im zweidimensionalen externen Rotationsparameterraum, der innerhalb des vorliegenden Bereichs durchsucht wird. Somit ergibt sich ein fünfdimensionaler Suchraum. Da die anderen Objekte in ihrer typischen Lage in der Regel keine interne Rotation aufweisen, werden bei ihnen der interne Translationsparameterraum, der externe Translationsparameterraum, das heißt die Entfernung des Objekts zur Kamera, und der verfügbare externe Rotationsparameter durchsucht. Der Suchraum ist hier somit vierdimensional. In den Tabellen wird ein Objekt als fehlerhaft lokalisiert eingestuft, wenn die interne Translation um mehr als 10 Bildpunkte oder einer der Rotationswerte um mehr als 15° vom tatsächlichen Wert abweicht. Die angegebenen mittleren und maximalen Abweichungen beziehen sich auf die richtig erkannten Objekte.

Die Lokalisationszeiten für den Objekttyp *Katze* liegen bei den Johnston-Merkmalen auf der ersten Auflösungsstufe bei 10 bis 15 Sekunden und auf der zweiten Stufe bei 20 bis 35 Sekunden, sodass sich insgesamt eine Rechenzeit von 30 bis 50 Sekunden ergibt. Bei den anderen Objekttypen liegt die Rechenzeit beider Stufen bei 10 bis 15 Sekunden, sodass insgesamt ein Zeitbedarf von 20 bis 30 Sekunden resultiert.

7.4 Hintergrundmodell

Für die Tests des Hintergrundmodells werden die Objekte der 2D-Experimente verwendet. Sie sind in insgesamt je 20 Aufnahmen (siehe Abb. 11) vor stark heterogenem Hintergrund, mit leichten Verdeckungen und zusammen mit ein beziehungsweise zwei anderen Objekten verfügbar. Damit stehen in der Summe 80 Aufnahmen für die Experimente zur Verfügung.

Da beim Hintergrundmodell für alle Merkmalspositionen des Hintergrunds die gleiche positionsunabhängige Dichtefunktion eingesetzt wird, sind zum Training keine vorsegmentierten Daten nötig. Stattdessen wird die Hintergrunddichte einfach mit allen zur Verfügung stehenden Merkmalen trainiert, unabhängig davon, ob sie zum Objekt gehören oder nicht. Der Grund für diese Vorgehensweise liegt darin, dass der Hintergrund ja auch aus beliebigen Objekten bestehen kann, also entsprechend unspezifisch ist.

Die Ergebnisse der Lokalisation mit Johnston-Merkmalen bei Szenen mit heterogenem Hintergrund oder Verdeckungen sind in Tabelle 4 dargestellt. Die Hintergrunddichte wird dabei entsprechend Abschnitt 5.3 Gleichung (14) durch Spaltenabhängigkeiten auf Merkmalsebene und Unabhängigkeit auf Zuordnungsebene modelliert. Die globale Suche erfolgt allerdings zunächst aufgrund einer Unabhängigkeitsannahme der Merkmale. Das dafür verwendete schnelle Suchverfahren setzt 10 Basisfunktionen zur Approximation der lokalen Dichten ein. Da keine externe Transformation vorliegt, ist $L_d = 1$ in Gleichung (38) zur globalen Suche. Aufgrund der Unabhängigkeitsannahme ist außerdem $N(m) = \{(0,0)\}$ in der Basisapproximation nach (34). Die a-priori-Wahrscheinlichkeiten der Zuordnungen werden als gleichverteilt angenommen, sodass in der aus Gleichung (14) abgeleiteten Gütefunktion nach Gleichung (39) $p(\zeta_m = 0) = (\zeta_m = 1) = 0.5$ (für alle m) ist.

Die Ergebnisse zeigen sowohl in Bezug auf die Fehlerrate als auch in Bezug auf die Lokalisationsgenauigkeit ein verbessertes Verhalten des Hintergrundmodells gegenüber

Tabelle 4. Lokalisationsergebnisse bei Szenen mit heterogenem Hintergrund oder Verdeckungen. Die Ergebnisse sind für alle in Abb. 11 dargestellten Szenentypen zusammengefasst

Modell	Fehler	Abweichung			
	(%)	Transl. (Pix)		Rotation (°)	
		Mittel	Max	Mittel	Max
Einzelobjekt	7	1.3	6	1.6	6.5
Hintergrund	1	1.1	5.4	1.0	4.3

dem Einzelobjektmodell. Die Lokalisationsgenauigkeit ist beim Einzelobjektmodell vor allem bei Objektverdeckungen sehr schlecht und kann durch eine explizite Hintergrundmodellierung verbessert werden. Nachteilig an der expliziten Hintergrundmodellierung ist der erhöhte Zeitbedarf von zirka 23 Sekunden gegenüber 5 Sekunden beim Einzelobjektmodell für die Lokalisation.

8 Ausblick

Das vorangegangene Kapitel hat anhand von Experimenten gezeigt, dass sich das in dieser Arbeit vorgestellte erscheinungsbasierte statistische System zur Lokalisation und Erkennung von starren Objekten in Einzelbildern eignet. Aufgrund seiner allgemeinen probabilistischen Modellierung ist es auf unterschiedlichste Objekttypen anwendbar und ermöglicht überdies die Einbeziehung von heterogenem Hintergrund. Die experimentellen Resultate für die vorliegenden Testdaten sind durchweg als sehr gut einzustufen. Dabei wurden die Testaufnahmen bewusst so angelegt, dass die Robustheit des Verfahrens überprüfbar ist. Benchmarkdatensätze, die diese Forderung erfüllen und gleichzeitig die für den probabilistischen Ansatz benötigte Vielzahl von Aufnahmen jeweils eines Objekts in unterschiedlichen, aber bekannten Lagen beinhalten, gibt es noch nicht. Derzeit wird ein derartiger Datensatz zum Vergleich unterschiedlichster Verfahren am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg zusammengestellt.

Die Voraussetzung umfangreicher Trainingsdatensätze schränkt das Verfahren derzeit noch auf Objekte ein, von denen automatisch ganze Aufnahmeserien gefertigt werden können. Lösungsmöglichkeiten bieten sich hier in der automatischen Schätzung der Objektlageparameter (siehe bspw. weitere Anmerkung unten) beliebiger Aufnahmesequenzen des Objekts an. Beim praktischen Einsatz des Verfahrens sind ferner die Rechenzeiten zu berücksichtigen, die ohne Hardwarebeschleunigung im Sekundenbereich liegen. Das bedeutet, dass das Verfahren zunächst für Einzelbildauswertungen geeignet ist. Die Auswertung größerer Datensätze ist nur dann sinnvoll, wenn keine sofortige Reaktion erforderlich ist, wie etwa bei der Analyse medizinischer Bildfolgen bei der Ganganalyse [14]. Bei der Suche in großen Bildmengen ist ein Einsatz nur dann denkbar, wenn – wie auch bei anderen gängigen Klassifikatoren – ein anderes schnelles Suchverfahren vorgeschaltet wird, das eine Grobauswahl trifft, sodass dem Verfahren nur eine abschließende detailliertere Bewertung bleibt.

Weiterentwicklungsmöglichkeiten des Ansatzes bestehen sowohl innerhalb der Rahmenbedingungen des Verfahrens als auch in Bezug auf allgemeinere Aufgabenstellungen. Bei gleichen Rahmenbedingungen sind beispielsweise

(a) die Bestimmung optimaler Merkmale,
(b) die Verwendung alternativer Basisfunktionen oder
(c) eine stärkere Verknüpfung der Hierarchieebenen

Erfolg versprechende Ansätze, um die Robustheit und Anwendbarkeit des Verfahrens weiter zu verbessern.

Eine der interessantesten Erweiterungen der Rahmenbedingungen des Systems ist der Verzicht auf die Annahme starrer Objekte und auf das Vorliegen der Transformationsparameter der Objekte in der Trainingsmenge. Beim Verzicht auf die zweite Annahme wäre ein wesentlich einfacheres Training möglich, das nicht mehr auf eine aufwendige technische Aufnahmeapparatur angewiesen ist. Eine Lösungsmöglichkeit für die Problematik beider Verallgemeinerungen besteht in der automatischen Bestimmung der Transformationsparameter durch ein anderes Verfahren. Dies kann beispielsweise eine Hauptachsentransformation sein, bei der die Koordinaten bezüglich der Eigenvektoren zu den maximalen Eigenwerten als Parameter gewählt werden. Da das vorliegende System die externen Parameter unabhängig von ihrer geometrischen Bedeutung behandelt, können die so ermittelten Parameter zur Objekterkennung eingesetzt werden.

Eine andere Erweiterungsmöglichkeit besteht in der Verwendung anderer/zusätzlicher Sensordaten zur Objekterkennung. Dazu können beispielsweise mehrere Objektansichten mit bekannten relativen Transformationen oder auch Farbaufnahmen eingesetzt werden. Erste Ergebnisse bei der Erkennung von Objekten, die in einzelnen Ansichten nahezu gleich aussehen, werden in [25] präsentiert. Hierbei wird der vorgestellte Ansatz im Rahmen eines Systems eingesetzt, das ein Objekt mit einer Kamera automatisch aus einer anderen Ansicht aufnimmt, wenn die aktuelle Ansicht keine eindeutige Klassifikation des Objetks zulässt.

Literatur

1. Antoine, J.; Murenzi, R.; Piette, B.; Duval-Destin, M.: *Image Analysis with 2D Continuous Wavelet Transform: Detection of Position, Orientation and Visual Contrast of Simple Objects*, in Meyer, Y. (Ed.): *Wavelets and Applications*, Springer, Marseille, Frankreich, Juni 1992, pp 144–159.
2. Black, M.; Jepson, A.: *EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation*, in *Fourth European Conference on Computer Vision (ECCV)*, No. 1065 in Lecture Notes in Computer Science, Springer, Heidelberg, April 1996, pp 329–341
3. Blanz, V.; Schölkopf, B.; Bülthoff, H.; Burges, C.; Vapnik, V.; Vetter, T.: *Comparison of View-Based Object Recognition Algorithms Using Realistic 3D Models*, in *International Conference on Artificial Neural Networks*, No. 1112 in Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, July 1996, pp 251–256.
4. Cootes, T. F.; Taylor, C. J.: *Statistical Models of Appearance for Computer Vision*, Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, Manchester, Großbritannien, September 1999.
5. Dahmen, J.; Beulen, K.; Güld, M.; Ney, H.: *A Mixture Density Based Approach to Object Recognition for Image Retrieval*, in *6th International RIAO Conference on Content–Based Multimedia Information Access*, Paris, April 2000, pp 1632–1647.

6. Dempster, A.; Laird, N.; Rubin, D.: *Maximum Likelihood from Incomplete Data via the EM Algorithm*, *Journal of the Royal Statistical Society, Series B (Methodological)*, vol 39, No. 1, 1977, pp 1–38.
7. Denzler J.; Niemann. H.: *Real-time Pedestrian Tracking in Natural Scenes* in *Computer Analysis of Images and Patterns (CAIP)*, Springer, Heidelberg, September 1997, pp 42–49.
8. Heidemann, G.; Ritter, H.: *Objekterkennung mit neuronalen Netzen*, in *Report – Situierte Künstliche Kommunikatoren, SFB 360*, Universität Bielefeld, März 1996.
9. Hornegger, J.: *Statistische Modellierung, Klassifikation und Lokalisation von Objekten*, Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, Erlangen, 1996, Shaker Verlag, Aachen, 1996.
10. Krüger, N.; Pötzsch, M.; v.d. Malsburg, C.: *Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs*, Ruhr-Universität Bochum, Januar 1996.
11. Kumar, V.; Manolakos, E.S.: *Unsupervised Model-Based Object Recognition by Parameter Estimation of Hierarchical Mixtures*, in ICIP 96 [24], pp 967–970.
12. Louis, A.K.; Maass, P.; Rieder, A.: *Wavelets*, Teubner, Stuttgart, 1994.
13. Manjunath, B.S.; Shekhar, C.; Chellappa, R.: *A New Approach to Image Feature Detection with Applications*, *Pattern Recognition*, vol 29, No. 4, 1996, pp 627–640.
14. Meyer, D.; Pösl, J.; Niemann, H.: *Gait Classification with HMMs for Trajectories of Body Parts Extracted by Mixture Densities*, in *British Machine Vision Conference (BMVC)*, Southampton, September 1998, pp 459–468.
15. Murase, H.; Nayar, S.K.: *Visual Learning and Recognition of 3–D Objects from Appearance*, *International Journal of Computer Vision*, vol 14, 1995, pp 5–24.
16. Niemann, H.: *Pattern Analysis and Understanding*, vol 4 of *Springer Series in Information Sciences*, Springer, Heidelberg, 1990.
17. Osuna, E.; Freund, R.; Girosi, F.: *Training Support Vector Machines: an Application to Face Detection*, in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society Press, Puerto Rico, June 1997, pp 130–143.
18. Pösl, J.; Niemann, H.: *Statistical 3–D Object Localization Without Segmentation Using Wavelet Analysis*, in *Computer Analysis of Images and Patterns (CAIP)*, Springer, Heidelberg, September 1997, pp 440–447.
19. Pösl., J.: *Erscheinungsbasierte statistische Objekterkennung*, Dissertation, Technische Fakultät, Universität Erlangen–Nürnberg, Erlangen, 1998, Shaker Verlag, Aachen, 1999.
20. Pösl, J.: *Statistical Pose Estimation with Local Dependencies*, in Seidel, H.-P.; Girod, B.; Niemann, H. (Eds.): *3D Image Analysis and Synthesis '97*, Infix, Sankt Augustin, November 1997, pp 147–154.
21. Pope, A.: *Learning to Recognize Objects in Images: Acquiring and Using Probabilistic Models of Appearance*, PhD thesis, University of British Columbia, Vancouver, 1995.
22. Porat, M.; Zeevi, Y.Y.: *The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 10, No. 4, 1988, pp 452–467.
23. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T.: *Numerical Recipes in C – The Art of Scientific Computing*, Cambridge University Press, New York, 1990.
24. *Proceedings of the International Conference on Image Processing (ICIP)*, IEEE Computer Society Press, Lausanne, Schweiz, September 1996.
25. Reinhold, M.; Deinzer, F.; Denzler, J.; Paulus, D.; Pösl., J.: *Active Appearance-Based Object Recognition Using Viewpoint Selection*, in Girod, B.; Greiner, G.; Niemann, H.; Seidel, H.-P. (Eds.): *Vision, Modeling, and Visualization 2000*, Infix, Berlin, November 2000, pp 105–112
26. Schiele, B.; Crowley, J.L.: *Probabilistic Object Recognition using Multidimensional Receptive Field Histograms*, in *Proceedings of the* 13^{th} *International Conference on Pattern Recognition (ICPR)*, IEEE Computer Society Press, Wien, August 1996, pp 50–54.
27. Wickerhauser, M.V.: *A Primer on Wavelet Theory and Its Applications*, in *International Wavelets Conference*, INRIA, Tanger, Marokko, April 1998.
28. Wright, C.R.; Vaz, R.F.; Cyganski, D.: *Establishing Identity and Pose of Objects From a Library Using Reciprocal Basis Set and Direction of Arrival Techniques*, in *Proceedings of the SPIE Symposium on Intelligent Robotic Systems*, Boston, USA, September 1993.

Josef Pösl studierte Informatik an der Universität Erlangen-Nürnberg und erhielt 1993 den Grad eines Diplominformatikers. Von 1994 bis 1995 war er Verantwortlicher für das Software-Systemdesign einer größeren betriebswirtschaftlichen Anwendung bei der Firma WITRON Logistik + Informatik GmbH. Im Rahmen seines Stipendiums des Graduiertenkollegs für 3D-Bildanalyse und -synthese fertigte er am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg von 1996 bis 1998 seine Dissertation zum Thema „Erscheinungsbasierte statistische Objekterkennung“ an und promovierte dort 1998 bei Prof. Dr.-Ing. H. Niemann. Von 1999 bis 2000 übernahm er bei der Firma WITRON Aufgaben im Bereich der Basissoftwareentwicklung, Vorbereitung neuer Software-Technologien, der Ausbildung von Fachinformatikern und war Mitglied im Forschungsinstitut Logistik. Seit Oktober 2000 ist er Professor für Informatik an der Fachhochschule Amberg-Weiden im Fachbereich Elektrotechnik. Er ist Mitglied von GI und IEEE.

Heinrich Niemann ist Professor für Informatik an der Universität Erlangen-Nürnberg und leitet die Forschungsgruppe Wissensverarbeitung am Bayerischen Forschungszentrum für Wissensbasierte Systeme (FORWISS). Davor war er an der Fachhochschule Giessen und am Fraunhofer Institut für Informationsverarbeitung in Technik und Biologie in Karlsruhe tätig. Seine Interessen liegen im Bereich der Verarbeitung und des Verstehens von Bild- und Sprachsignalen sowie in der Anwendung von Methoden der künstlichen Intelligenz in diesem Bereich. Er erhielt die akademischen Grade Dipl.-Ing. (Elektrotechnik) und Dr.-Ing. von der Technischen Universität Hannover.