Zur Erkennung verformbarer Objekte anhand ihrer Teile

Vom Fachbereich 12 Elektrotechnik und Informatik

der Universität Siegen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
(Dr.-Ing.)

genehmigte Dissertation

von
Dipl.-Ing. Martin Stommel

1. Gutachter: Prof. Dr.-Ing. Klaus-Dieter Kuhnert
2. Gutachter: Prof. Dr. Volker Blanz

Tag der mündlichen Prüfung: 1.7.2010

Danksagung

Ich danke Li Zhang (张俐) für die Aufbereitung der Stichprobe und Ting Yuan (袁婷), Lu Xi (奚璐) und Bo Lou (娄渤) für die Programmierung der Farbklassifikation.

Zusammenfassung

Die Erkennung verformbarer Objekte mit den Mitteln der digitalen Bildverarbeitung ist ein drängendes, aber bisher weitgehend ungelöstes Problem. In vielen industriellen und anderen Bereichen besteht ein großer Bedarf, Abläufe zu automatisieren, die in einer sich verändernden oder nicht vollständig kontrollierbaren Umgebung stattfinden. Technische Systeme folgen jedoch derzeit in der Regel starren Abläufen, ohne mit ihrer Umgebung zu interagieren. Das Hauptproblem liegt dabei in der Interpretation der Kameradaten. Die existierenden Verfahren zur Erkennung von Objekten funktionieren nur in einfachen Spezialfällen.

In dieser Arbeit wird daher ein neuartiger Ansatz untersucht, der sowohl eine Klassifikation als auch eine Lokalisation von Objekten im Bild ermöglicht. Dazu wird ein kompositionelles Modell eingeführt, bei dem ein Objekt als Hierarchie von Teilen und Unterteilen in geometrischen Beziehungen beschrieben wird. Ein besonderer Schwerpunkt liegt dabei auf der Untersuchung, welches Verhältnis zwischen der Ausprägung und der Position lokaler Merkmale besteht. Da gerade verformbare Objekte in ihrer Erscheinung stark variieren, speichert das Modell mehrere Objektansichten. Dies unterscheidet den vorliegenden Ansatz von vielen anderen.

Das Modell wird mittels einer Stichprobe von Beispielbildern trainiert. Dies umfaßt sowohl die automatische Wahl geeigneter Teile als auch die Identifikation charakteristischer Ansichten. Die Teilemengen auf verschiedenen Hierarchieebenen werden aufgrund unterschiedlicher Randbedingungen individuell optimiert. Über eine Erkennungsmethode, die sowohl zur Hough-Transformation als auch zu Radialen Basisfunktionen Ähnlichkeiten besitzt, wird das Modell mit den Bildern verglichen.

Die Leistungsfähigkeit des entwickelten Verfahrens wird am Beispiel einer Cartoon-Datenbank gezeigt. Dazu werden unterschiedliche Modellkonfigurationen vorgestellt, die bei einer Korrektklassifikationsrate von mindestens 78 Prozent entweder einen positiven Vorhersagewert von 97 Prozent oder eine Sensitivität von 93 Prozent erreichen.

Abstract

The recognition of deformable objects by the means of digital image processing is a crucial, but widely unsolved problem yet. In many industrial and other areas there is a strong need to automate processes which take place in a changing or not completely controllable environment. However, technical systems are presently characterised by fixed operational procedures and little interaction with their environment. The main problem lies in the interpretation of the camera data. The existing object recognition methods work only in simple special cases.

Therefore, in this thesis a novel approach is studied which allows for a simultaneous classification and localisation of the objects present in an image. To this end, a compositional model is introduced which describes an object as a hierarchy of parts and sub-parts. Between parts, geometrical relationships are modelled. A major emphasis is placed on the analysis of the relationship between the position and the attributes of parts. To account for the strongly varying appearance of deformable objects, the model stores multiple views. This is in contrast to many other recent approaches.

The model is build by analysing sample images of the objects to be recognised. The training comprises both the automatic selection of appropriate parts as well as the identification of characteristic views. Due to differing boundary conditions, the resulting sets of parts are optimised individually for every level in the hierarchy. The comparison between the model and a test image is done by a voting method which has similarities to the Hough transform or to radial basis functions.

The performance of the new methods is demonstrated by the recognition of a character from a cartoon data-base with strongly varying appearance. Two model configurations are presented achieving either a precision of 97 percent or a recall of 93 percent with a general accuracy of at least 78 percent for both cases.

Inhaltsverzeichnis

Kapitel 1

Objekterkennung und Lokalisation

In vielen Anwendungsbereichen besteht ein großes Bedürfnis, Abläufe zu automatisieren, die in einer sich verändernden oder nicht vollständig kontrollierbaren Umgebung stattfinden. Diese betreffen in industriellen Anwendungen vor allem Prüf-, Sicherheits- und Fertigungsaufgaben. Ein mobiler Roboter könnte beispielsweise automatisch ein größeres Fabrikareal abfahren und Rohrleitungen auf Lecks prüfen oder den Gebäudezustand begutachten. Ein solcher Roboter könnte auch Montagearbeiten unterstützen und auf Anweisung Bauteile holen oder andere Handreichungen durchführen, ohne daß man ihm den damit verbundenen Arbeitsablauf detailliert vorgibt. Einen noch höheren Grad an Autonomie erfordern Service-Roboter, die beispielsweise im Haushalt eingesetzt werden sollen. Für solche Roboter wird gefordert, daß sie flexibel mit Menschen kommunizieren können und mit alltäglichen Gegenständen umgehen. Ein Kennzeichen dieser Einsatzgebiete ist, daß sie sich nicht gestalten lassen und ein hohes Maß an Unvorhersehbarkeit enthalten.

In der Praxis sind technische Systeme jedoch derzeit durch starre Arbeitsabläufe und eine geringe Interaktion mit ihrer Umgebung gekennzeichnet. Die Kommunikation geschieht in der Regel einseitig durch die Vorgabe von Bewegungen mittels Tastern oder Computerprogrammen. Die Arbeitsabläufe müssen Bewegung für Bewegung vorgegeben werden. Eine Umgebungserkennung findet höchstens in der Form statt, daß bestimmte Parameter ansonsten bekannter Objekte vermessen werden oder ein unbekanntes Objekt abhängig von bestimmten Sensorwerten einer von wenigen bekannten Klassen von Objekten zugeordnet wird. Da die Leistungsfähigkeit aktueller Systeme für eine Reaktion auf veränderliche Einsatzbedingungen nicht ausreicht, wird die Umgebung so vorhersehbar wie möglich gestaltet.

Das Hauptproblem ist, daß die Systeme ihre Umgebung nur begrenzt wahrnehmen. Es stehen zwar viele Sensoren zur Verfügung, mit denen die Umgebung vermessen werden kann, die Interpretation der Sensorwerte ist allerdings

für den Fall einer allgemeinen Erkennung von Umgebungsobjekten weitgehend ungelöst. Auf der Sensorseite können beispielsweise Berührungssensoren, Kameras, Laserscanner, Ultraschall- und Radarsensoren eingesetzt werden. Von diesen Sensoren haben Kameras theoretisch den größten Einsatzbereich, da sie sehr schnell sind, die meisten Oberflächen erkennen und berührungslos sowohl nahe als auch ferne Objekte aufnehmen. Andererseits ist die Interpretation von Kamerabildern schwierig, da die meisten Objekte in einer Vielzahl von Erscheinungen auftreten können. Die Erscheinung eines Objekts hängt von der Beleuchtung, der Objektoberfläche, soweit diese überhaupt vorhanden ist, und der Lage des Objekts ab. Eine besondere Schwierigkeit stellen insbesondere auch verformbare Objekte dar, da diese besonders stark in ihrer Erscheinung variieren. Einen weiteren Einfluß stellt die Kamera selbst dar, die geometrische Verzerrungen, Rauschen, Zeilenverschiebungen, Helligkeitsänderungen oder Kompressionsartefakte bewirken kann. Diese Einflüsse sind allerdings in der Regel sehr viel schwächer als die Variabilität der Objekte an sich.

Die Erkennung deformierbarer Objekte in digitalen Bildern wird in dieser Arbeit genauer untersucht. Um die Wirksamkeit der entwickelten Methoden zu prüfen, wird prototypisch eine Software zur Objekterkennung erstellt und auf eine Bilderdatenbank angewandt. Damit die Fragestellung präzisiert werden kann, wird als nächstes der allgemeine Aufbau eines Objekterkennungssystems beschrieben.

1.1 Aufbau von Objekterkennungssystemen

Ein Rechner erkennt Objekte anhand bestimmter *Merkmale*. Häufige Typen von Merkmalen sind Bildpunkte, Kanten, Ecken und Farbverläufe. Merkmale können sich neben dem Typ auch in ihrer Ausprägung unterscheiden. Bei Kantenmerkmalen kann beispielsweise die Orientierung der Kante verschieden ausgeprägt sein oder bei Bildpunkten die Farbe. Merkmale repräsentieren typische Eigenschaften von Objekten und erlauben es, wenn sie gut gewählt sind, von dem vorliegenden Bildmaterial auf ein bestimmtes Objekt zu schließen. Um diese Zuordnung zu erreichen, müssen zu jedem Objekt, daß erkannt werden soll, die typischen Merkmale gespeichert werden. Diese Informationen stellen das *Modell* dar.

Üblicherweise sollen nicht einzelne Objekte erkannt werden, sondern verschiedene *Klassen* von Objekten. Eine Klasse umfaßt dabei mehrere gleichartige Objekte. Das Modell speichert in dem Fall Merkmale, die so allgemein sind, daß sie in allen Objekten einer Klasse auftreten, sich für verschiedene Klassen aber unterscheiden.

Um die in einem Bild dargestellten Objekte zu erkennen, führt das Objekterkennungssystem zuerst eine Merkmalsextraktion durch. Während dieses Schrittes wird berechnet, welche Merkmale in welchen Ausprägungen im Bild vorliegen. Die extrahierten Merkmale werden nun mit dem Modell verglichen. Wenn zwischen den extrahierten und den im Modell gespeicherten Merkmalen einer bestimmten Klasse eine ausreichend hohe Übereinstimmung besteht, wird

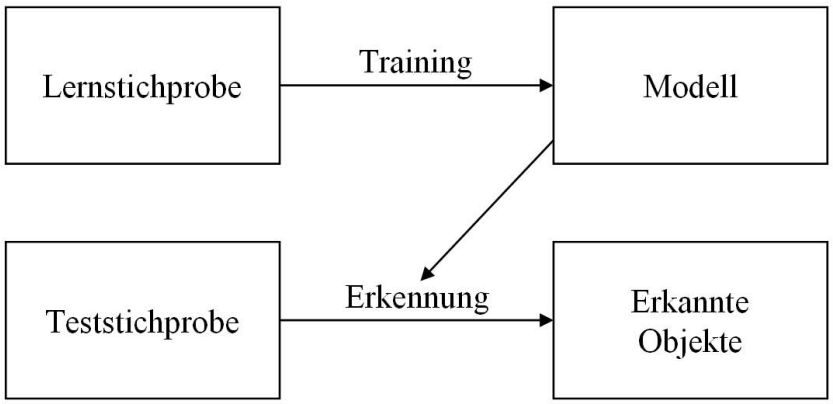


Abbildung 1.1: **Prinzipieller Ablauf der Objekterkennung:** Während der Trainingsphase wird eine Stichprobe der zu erkennenden Objekte analysiert. Die objekttypischen Eigenschaften werden in einem Modell zusammengefaßt. In der Testphase werden Bilder einer zweiten Stichprobe mit dem Modell verglichen. Die Übereinstimmungen zwischen dem Modell und der Stichprobe stellen das Ergebnis der Objekterkennung dar und werden ausgegeben.

entschieden, daß es sich bei den extrahierten Merkmalen um die betreffende Klasse handelt. Diese Zuordnung ist die *Klassifikation*.

Für viele Anwendungsbereiche ist es nicht nur interessant, welche Klassen von Objekten in einem Bild vorliegen, sondern auch wo die Objekte im Bild liegen. Die Bestimmung der Objektpositionen im Bild ist die *Lokalisation*. Der Begriff *Objekterkennung* bezeichnet im folgenden die Kombination aus Klassifikation und Lokalisation.

Vor der Objekterkennung muß das Modell erzeugt werden. Dieser Schritt wird als *Training* bezeichnet. Während des Trainings wird mittels einer *Trainingsstichprobe* von repräsentativen Objektdarstellungen analysiert, anhand welcher Merkmale die verschiedenen Klassen erkannt werden können. Traditionell wird jedem Element der Trainingsstichprobe vor dem Training manuell die jeweilige Klassenbezeichnung zugeordnet. In dem Fall handelt es sich um *überwachtes* Training. Da dieser Schritt sehr aufwendig sein kann, wurden Methoden entwickelt, bei denen die vorherige Zuordnung nur unvollständig geschieht oder vollständig ausgelassen wird. Diese Methoden werden *schwach überwacht* bzw. *unüberwacht* genannt.

Nach dem Training des Modells muß die Güte des Modells und die der Objekterkennung geprüft werden. Dazu wird eine Objekterkennung auf einer *Teststichprobe* durchgeführt.

Gute Ergebnisse der Teststichprobe zeigen an, daß das trainierte Modell repräsentativ für die in Frage kommenden Klassen sind. Gute Ergebnisse zeigen sich in vielen richtigen und wenigen falschen Treffern. Wenn die Ergebnisse

		Tatsächliche Klasse	
		Positiv	**Negativ**
Erkannte Klasse	**Positiv**	Richtig Positive	Falsch Positive
	Negativ	Falsch Negative	Richtig Negative

Tabelle 1.1: Vertauschungsmatrix für Zwei-Klassen-Probleme. Die Vertauschungsmatrix gibt für die Klassen "Positiv" und "Negativ" an, wieviele Stichprobenelemente in welche Klasse eingeordnet wurden.

schlecht sind, kann dies daran liegen, daß das Modell zu selektiv ist. In dem Fall ergeben sich sowohl wenige falsche als auch wenige richtige Ergebnisse. Eine andere Ursache für schlechte Ergebnisse kann eine zu hohe Verallgemeinerbarkeit sein. Dann ergeben sich zwar viele richtige Treffer, aber auch viele falsche.

Für kleine Bilder, die entweder vollständig von einem zu erkennenden Objekt ausgefüllt sind oder nur den Hintergrund zeigen, können die Ergebnisse systematisch in Form einer Vertauschungs- oder Wahrheitsmatrix (engl. *Confusion Matrix*) dargestellt werden. Diese gibt die Anzahl an richtig und falsch erkannten Stichprobenbildern wieder. Tabelle 1.1) zeigt eine Vertauschungsmatrix für die Unterscheidung von zwei Klassen. Im Idealfall ergeben sich nur Ergebnisse auf der Hauptdiagonalen. Im ungünstigsten Fall sind die Ergebnisse gleichmäßig über die ganze Matrix verteilt, sodaß noch nicht einmal durch die Negation der Ergebnisse ein Erkenntnisgewinn erzielt werden kann. Aus der Vertauschungsmatrix lassen sich weitere Qualitätskriterien ableiten. Die wichtigsten sind die der positiver Vorhersagewert (engl. *Precision*), die Sensitivität (engl. *Recall*) und die Korrektklassifikationsrate (engl. *Accurracy*). Der

$$\text{positive Vorhersagewert} = \frac{\text{Anzahl richtig Positiver}}{\text{Anzahl der richtigen und falschen Positiven}} \quad (1.1)$$

gibt an, mit welcher Wahrscheinlichkeit ein detektiertes Objekt korrekt erkannt wurde. Im Idealfall liegt der Wert bei Eins. In dem Fall ist jedes Objekt, das vom System erkannt wird, richtig. Bei kleineren Werten werden zunehmend Teile des Hintergrundes fälschlicherweise als Objekt erkannt. Die Optimierung des positiven Vorhersagewertes kann jedoch für manche Anwendungen zu konservativ sein, da insgesamt nur in sehr sicheren Fällen überhaupt Detektionen stattfinden. Viele Objekte bleiben möglicherweise unentdeckt. Die

$$\text{Sensitivität} = \frac{\text{Anzahl richtig Positiver}}{\text{Anzahl der richtig Positiven und falsch Negativen}}$$

gibt den Anteil der korrekt erkannten Elemente an der Gesamtmenge der positiven Stichprobenelemente an. Dieser Wert muß optimiert werden, wenn möglichst viele Objekte erkannt werden sollen. Allerdings geht die Anzahl der falschen Treffer auf dem Hintergrund nicht in die Berechnung mit ein. Die

$$\text{Korrektklassifikationsrate} = \frac{\text{Anzahl richtiger Klassifikationen}}{\text{Anzahl aller richtigen u. falschen Klassifikationen}} \quad (1.2)$$

gibt den Anteil der richtigen Klassifikationen an der Gesamtzahl der Klassifikationen dar. Vorder- und Hintergrunddetektionen werden hier gleich gewichtet, sodaß eine Optimierung der Korrektklassifikationsrate einen Mittelweg zwischen der Optimierung der Sensitivität und des positiven Vorhersagewertes darstellt. Abbildung 1.1 faßt den groben Ablauf der Objekterkennung noch einmal zusammen.

1.2 Ziele, Vorgehensweise und Stichprobe

Die Hauptprobleme der Objekterkennung sind die Repräsentation der Klassen, der Vergleich von Testbildern mit dem Modell und die Erzeugung des Modells aus einer Stichprobe. Für den ersten Punkt ist entscheidend, daß das Modell die typischen Eigenschaften eines Objekts widerspiegelt. Dabei ist ein Kompromiß zwischen Selektivität und Verallgemeinerbarkeit zu treffen. Das Ziel für diesen Kompromiß ist, eine möglichst hohe Robustheit gegenüber den Objektvariationen innerhalb der Klassen zu erzielen, dabei aber auch genug Informationen zu speichern, um verschiedene Klassen gut unterscheiden zu können. Wie gut der Kompromiß ausfällt, hängt stark von den gewählten Merkmalen ab. In der Praxis analysiert dazu ein Experte das Datenmaterial und wählt aufgrund der Ergebnisse und seiner Erfahrung sinnvolle Merkmale. Auf diese Weise können sehr gute Modelle entstehen, die allerdings immer speziell auf das jeweilige Anwendungsgebiet angepaßt sind. Für ein anderes Anwendungsgebiet müssen neue Merkmale ausgewählt werden. Da das wiederholte Zusammenstellen von vollständig anwendungsspezifischen Merkmalen zeitaufwendig ist, wird in dieser Arbeit die Erkennung der allgemeineren Klasse der deformierbaren Objekte behandelt. Es ist daher zunächst zu klären, wie ein ausreichend flexibles Modell aussieht, da deformierbare Objekte in ihrer Erscheinung stark variieren. Ein vielversprechender Ausgangspunkt besteht hier sicher in der Kombination mehrerer, unterschiedlicher Merkmale.

Während der Objekterkennung werden Testbilder mit dem Modell verglichen. Dabei sollten Störungen wie Verdeckungen oder geometrische Verformungen toleriert werden. Zu diesem Zweck wird in dieser Arbeit ein auf Teilen basierender Ansatz untersucht: Objekte werden in mehrere Teile aufgespalten, die einzeln mit dem Modell verglichen werden. Wenn eine ausreichende Anzahl von Teilen in einem Bild gefunden wird, gilt ein Objekt als erkannt. Der Teile-Ansatz gilt in der Literatur als tolerant gegenüber Störungen, da einzelne Teile fehlen dürfen und die geometrischen Beziehungen zwischen den Teilen nur grob übereinstimmen müssen.

Die Objekterkennung kann nur dann effizient durchgeführt werden, wenn die Modellrepräsentation auf das Objekterkennungsverfahren optimiert wird. Dabei ergeben sich aus dem Teile-basierten Ansatz eine Reihe von bisher weitgehend unbeantworteten Fragen. Als erstes muß geklärt werden, welche Dinge eines Objekts sich eignen, um als Teil modelliert zu werden und an welchen Stellen Objekte aufgespalten werden können. Um diese Frage zu klären, wird hier insbesondere die Teilegeometrie untersucht. Konsequenterweise wird ein

Teil selbst wieder als Zusammensetzung mehrerer Unterteile betrachtet. Daraus ergibt sich eine Teilehierarchie mit mehreren Abstraktionsebenen. Eine solche Hierarchie erscheint vorteilhaft, da sie einige Parallelen zu dem Sehsystem von Säugetieren aufweist, dessen Flexibilität und Leistungsfähigkeit von technischen Systemen bisher nicht erreicht wird. Die Betrachtung mehrerer Abstraktionsebenen erlaubt zudem die nahtlose Integration mehrerer Objektansichten in das Modell. Für das Training einer solchen Hierarchie liegen bisher jedoch nur vereinzelt Erfahrungen vor. Eine Lösungsmöglichkeit wäre die Anwendung eines allgemeingültigen Optimierungsverfahrens aus dem Bereich des Maschinenlernens. Diese Verfahren erlauben aber nur eine geringe Einsicht in den gefundenen Lösungsweg. Da diese Verfahren auch kein Vorwissen über die Bildverarbeitung beinhalten, werden während des Trainings viele erfolglose Optimierungswege überprüft. Aus diesem Grund wird in dieser Arbeit experimentell nach Einflußfaktoren gesucht, die für ein gutes Modell entscheidend sind. Auf diese Weise werden wertvolle Erfahrungen für das Training Teile-basierter Modelle gesammelt.

Die Leistungsfähigkeit des Verfahrens wird anhand praktischer Experimente nachgewiesen. Da diese Arbeit insbesondere auf die Erkennung verformbarer Objekte abzielt, werden Cartoon-Vorlagen [Dis] als Testmaterial für die Objekterkennung gewählt. Die starken Verformungen der Cartoonfiguren unterscheiden die gewählte Stichprobe von vielen anderen Datenbanken, die hauptsächlich auf die Erkennung starrer Objekte ausgerichtet sind, darunter z.B. Haushaltsgegenstände und Spielzeug [SANM96], Autos [SK05] und Legosteine [CAC04].

Die Abbildungen 1.2 und 1.3 zeigen Beispiele aus der Cartoon-Datenbank. Als Positivbeispiele enthält die Datenbank Bilder des Kopfes der Disney-Figur *Donald*, welche aus der Buchvorlage [Dis] gescannt wurden. Als Negativbeispiele wurden zufällige, quadratische Bildausschnitte außerhalb von Donald-Köpfen hinzugenommen (Abb. 1.3e). Die Größe der Negativbeispiele spiegelt die Verteilung der Größe der Donald-Köpfe wieder.

Die hervorstechendsten Merkmale der Cartoon-Stichprobe sind die starken geometrischen Verformungen (Abb. 1.2a) des dargestellten Objekts sowie die Vielfalt an Objektposen und Perspektiven (Abb. 1.2b). Die Verformungen resultieren aus Objektbewegungen, Änderungen der Mimik oder Kontakt mit anderen Objekten. Aufgrund satirischer Überzeichnung sind die Verformungen viel stärker als in der Natur. Der einzige Eingriff in die durch die Buchvorlage gegebene Szenerie besteht in der Auswahl von Donald-Köpfen, deren Gesicht zu einem Großteil sichtbar ist. Bildstörungen durch das Aufnahmegerät sind gegenüber der Variationen in der Vorlage vernachlässigbar. Tiefenschärfe, Interlacing und Kompressionsartefakte spielen keine Rolle. Es gibt keine Hinweise, daß in der verwendeten Vorlage für bestimmte häufige Posen Schablonen eingesetzt werden.

Wie bei Aufnahmen natürlicher Objekte treten Verdeckungen (Abb. 1.2c), Beleuchtungseffekte, Bewegungsunschärfen und seltener Spiegelungen auf. Die Beleuchtungsart kann durch Schraffuren oder schwarze Flächen angedeutet sein (Abb. 1.2d). Schnelle Bewegungen werden durch quer über das Objekt verlau-

fende Striche, teilweise auch die Überlagerung mehrerer Objektposen dargestellt (Abb. 1.2e).

Gegenüber Verformung, Perspektive und Bewegung sind die Objektoberflächen vereinfacht einfarbig und matt dargestellt. Vereinzelt werden Glanzpunkte eingezeichnet. Diese Vereinfachungen werden allerdings durch die Zeichen- und Reproduktionstechnik teilweise kompensiert. Wie Abbildung 6.3 zeigt, muß die Farbe einer Fläche über einen gewissen Bereich des Halbtonrasters gemittelt werden. Abbildung 1.3a zeigt, daß die Farbfüllungen der Flächen auch nicht immer mit den schwarzen Strichen übereinstimmen. Die Farbe wird gelegentlich variiert, um eher symbolische Nebenbedeutungen zu illustrieren (Abb. 1.3b).

Es wurde ebenfalls überlegt, einen Disney-Spielfilm zu analysieren, um kontinuierliche Objektbewegungen zu extrahieren. Aufgrund der starken Bewegungen, der bei Cartoons niedrigen Framerate und der großen homogenen Flächen hat sich dieser Ansatz aber als wenig erfolgversprechend herausgestellt.

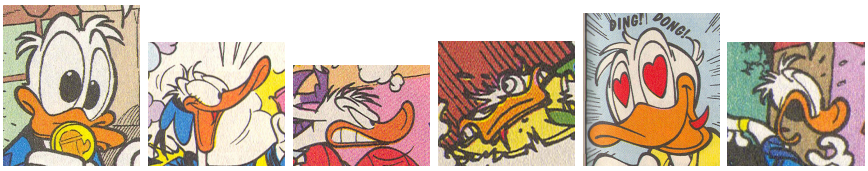
1.3 Verfahrensübersicht und Vorschau auf die folgenden Kapitel

Die zu erkennenden Objekte werden in dieser Arbeit als Menge von Teilen in bestimmten geometrischen Anordnungen beschrieben. Teile dürfen ihrerseits wieder aus Teilen zusammengesetzt sein, wodurch sich ein Modell mit mehreren Hierarchieebenen ergibt. Um die geometrischen Anordnungen zu modellieren, wird einem Objekt eine Referenzkoordinate in der Objektmitte zugeordnet. Die Positionen der Teile werden relativ zu dieser Position modelliert. Bei zusammengesetzten Teilen werden die Positionen der Unterteile relativ zur Position des Oberteils angegeben. Um Rauschen und geometrische Verzerrungen berücksichtigen zu können, wird für jedes Teil eine Ortstoleranz eingeführt, um die seine Position von der idealen Position abweichen darf.

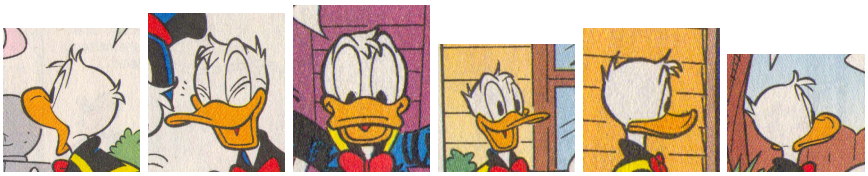
Die auf einer Hierarchieebene modellierten Teile können als Elemente eines *visuellen Alphabets* betrachtet werden, die sich modular zu komplexeren Objekten kombinieren lassen. Nach Hierarchieebene geordnet repräsentieren die visuellen Alphabete

- Klassen von Objekten,
- Mehrere Ansichten von Objekten gleicher Klasse,
- Teile von Objektansichten
- und einfache Basismerkmale.

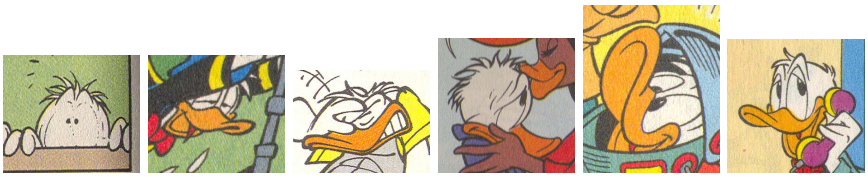
Die Objekterkennung geschieht mit Hilfe von Mehrheitsentscheiden. Ausgehend von den auf der niedrigsten Hierarchieebene erkannten Teilen werden die Positionen der Teile auf den nächsthöheren Ebenen geschätzt. Ein Teil gilt als erkannt, wenn eine ausreichende Anzahl seiner Unterteile erkannt wurden. Der dazu nötige Schwellwert wird für jedes Teil individuell gespeichert. Die Robustheit des Verfahrens wird durch bestimmte weitere Bedingungen gesteigert.



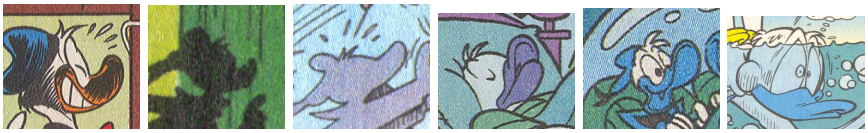
a) Objektverformungen treten als Folge von Bewegungen der Figur auf. Oft drücken sie jedoch auch besondere Emotionen oder äußere Kräfte aus.



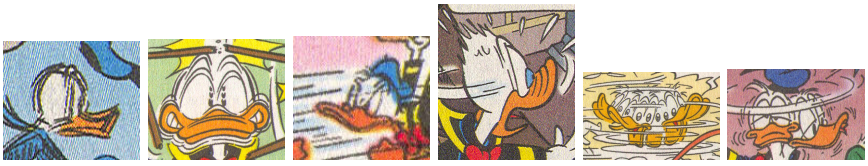
b) Rotation der Figur um die vertikale Achse. Rotationen um andere Achsen treten ebenfalls auf.



c) Teile der Figur sind durch Objekte im Vordergrund verdeckt.

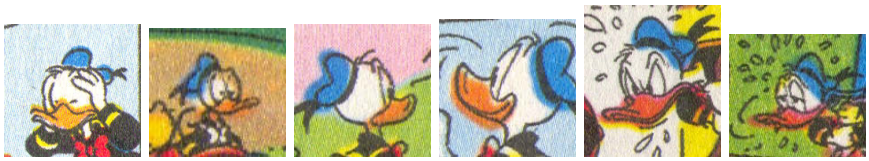


d) Von der üblichen Auflichtbeleuchtung abweichende Beleuchtungsarten. Von links nach rechts: Seitliche Beleuchtung mit Schattenwurf. Durchlicht. Durchlicht bei schwachem Kontrast. Schwache diffuse Beleuchtung. Schwach beleuchtete Aufnahme durch spiegelndes Glas. Unterwasseraufnahme.



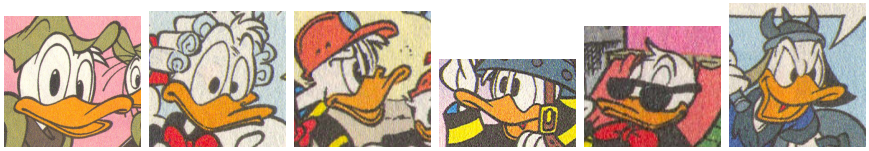
e) Bewegungsunschärfe: 2× Vibration. 1× Translation. 3× Drehbewegung.

Abbildung 1.2: **Beispiele aus der Cartoon-Datenbank.** Die Bilder sind aus Darstellungsgründen auf die gleiche Breite skaliert.

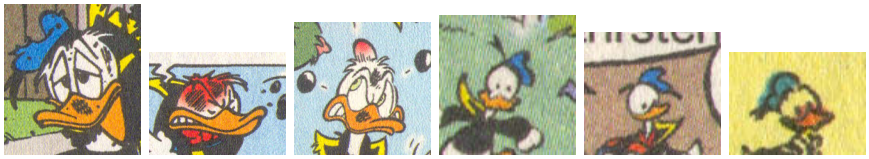


a) Ungenaue Einfärbung von Flächen beim Zeichnen und Toleranzen bei der Ausrichtung der Farbauszüge im Druck.

b) Außergewöhnlicher Farbeinsatz zum Ausdruck bestimmter Situationen und Charaktereigenschaften.



c) Verkleidungen wurden zugelassen, sofern wesentliche Teile des Gesichts sichtbar blieben.



d) Sonstige Effekte: Besondere Oberflächeneigenschaften werden durch Schraffur und Farbe gekennzeichnet (Bilder 1–3 dieser Zeile). Kleine Bilder sind oft wenig detailliert (Bilder 4–6).

e) Beispiele für Ausschnitte aus dem Hintergrund, die während des Trainings als Gegenbeispiele genutzt wurden.

Abbildung 1.3: **Weitere Beispiele aus der Cartoon-Datenbank.** Die Bilder sind aus Darstellungsgründen auf die gleiche Breite skaliert.

Das Training geschieht separat für alle Hierarchieebenen des Modells und beginnt auf der niedrigsten Ebene. Als Basismerkmale dienen Kanten, Ecken und Punkte auf Skelettlinien. Während des Trainings werden günstige Quantisierungsstufen für die Basismerkmale bestimmt. Hier spielen die Güte der Objekterkennung und die Modellgröße eine Rolle.

Auf der nächsten Ebene werden räumlich benachbarte Basismerkmale zu komplexen Teilen zusammengefaßt. Die Geometrieparameter werden mit Hilfe linearer Regeln abhängig von der Teilegröße und der Modellgüte eingestellt. Aus den vielen möglichen Teilen, die sich aus der alleinigen Berücksichtigung der räumlichen Benachbarung ergeben, wird mit Hilfe eines Clusterungsschrittes ein kompaktes Alphabet von Teilen zusammengestellt.

Die Elemente des Teilealphabets werden anschließend so kombiniert, daß sie prototypischen Objektansichten modellieren. Um die prototypischen Ansichten zu finden, wird eine Stichprobe mit Beispielansichten geclustert. Es werden die Teile zu einem abstrakteren Teil kombiniert, die in allen Beispielen einer prototypischen Ansicht detektierbar sind. Die Parametrisierung des neuen Teils geschieht abhängig von bestimmten Merkmalen der Clusterstruktur der Beispielansichten. Auf diese Weise entsteht ein Alphabet von Teilen, welche Objektansichten darstellen.

Auf der höchsten Hierarchieebene werden mehrere Ansichten von Objekten gleicher Klasse kombiniert. Gegenseitige Überlappungen bei der Stichprobenabdeckung durch einzelne Ansichtsmodelle werden hier zur Rauschunterdrückung genutzt. Durch verschiedene Strategien bei der Kombination lassen sich unterschiedliche Modelleigenschaften einstellen.

Der Rest der Arbeit gliedert sich wie folgt:

In Kapitel 2 wird ein entgegengesetzter, speziell auf die Erkennung von Cartoonfiguren ausgelegter Klassifikator vorgestellt. Dieser dient der Gegenüberstellung des anwendungsspezifischen und des allgemeingültigen Ansatzes. Die starken Einschränkungen des anwendungsspezifischen Ansatzes motivieren die Ausrichtung der vorliegenden Arbeit auf den allgemeingültigen Ansatz.

Kapitel 3 faßt den Stand der Forschung zu diesem Ansatz zusammen. Hier werden sowohl neurophysiologische Erkenntnisse als auch technische Lösungsansätze dargestellt. Dabei werden derzeit offene Fragen identifiziert und aussichtsreiche Lösungsmöglichkeiten herausgearbeitet.

In Kapitel 4 geht es dann um die experimentelle Umsetzung der Lösungsansätze. Zunächst wird ein hierarchisches, auf Teilen basierendes Modell vorgeschlagen. Aufgrund seiner flexiblen Graphstruktur ist das Modell ausdrucksstark genug, um verschiedene vor allem die Objektgeometrie betreffende Fragestellungen zu analysieren.

In Kapitel 5 wird dann eine Methode vorgestellt, mit der das Modell zur Objekterkennung eingesetzt werden kann. Diese erlaubt die gleichzeitige Klassifikation und Lokalisation der in einem Bild dargestellten Objekte.

Kapitel 6 behandelt anschließend das Training des Modells. Dieses beginnt mit der Untersuchung von lokalen Merkmalen auf der niedrigsten Abstraktionsebene des Modells. Hier geht es vor allem um die Detektion von Kanten und Flächen, sowie das Rauschen der Merkmalsausprägungen. Als nächstes wird

analysiert, aufgrund welcher Kriterien Objekte unterteilt werden sollen. Die Ergebnisse motivieren die Modellierung einer Menge allgemeingültiger Muster in Form eines visuellen Alphabets. Die Elemente des visuellen Alphabets werden auf den höheren Hierarchieebenen zunächst zu Modellen für verschiedene Objektansichten und dann zu Kategorien bzw. Klassen zusammengesetzt.

Kapitel 7 faßt alle wichtigen Ergebnisse zusammen und nennt Ansatzpunkte für weitere Forschungen.

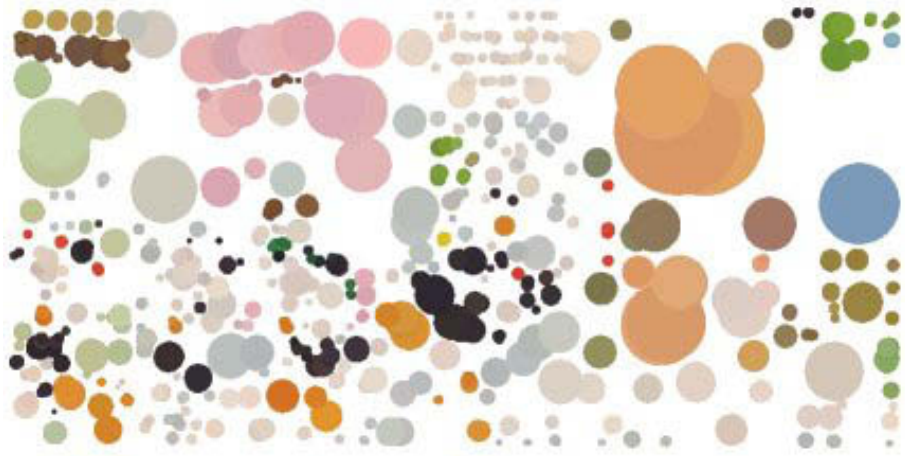
Kapitel 2

Anwendungsspezifische Objekterkennung

Zunächst wird ein anwendungsspezifischer Ansatz untersucht. Bei diesem Ansatz wird nach Merkmalen gesucht, die für eine vorliegende Stichprobe besonders hohe Erkennungsraten liefern. Wie sich herausstellt, unterliegt der dazu vorgeschlagene Farbklassifikator jedoch starken Einschränkungen. Insbesondere ist nicht zu erwarten, daß andere Anwendungsgebiete von der hier erzielten hohen Erkennungsrate profitieren. Die Ergebnisse begründen so den allgemeingültigeren Ansatz, der den Hauptteil dieser Arbeit darstellt. Der in diesem Kapitel vorgestellte Farbklassifikator ist daher keine Komponente des späteren allgemeingültigen Ansatzes, sondern dient der Gegenüberstellung beider Ansätze.

Es wird nun geprüft, wie gut sich die Cartoon-Datenbank mittels eines speziell auf die Erkennung von Donald-Köpfen ausgelegten Verfahrens klassifizieren läßt. Als hervorstechendes Merkmal der Cartoon-Figuren wird der orange Schnabel ausgewählt, von dem man zudem weiß, daß er sich meistens unter einem schwarz-weißen Kopf befindet. Diese Information ist als anwendungsspezifisches Wissen einzuschätzen, das im allgemeinen Fall nicht zur Verfügung steht und durch den Rechner schlecht automatisch hergeleitet werden kann. Es ist daher zu erwarten, daß die Auswertung dieser besonderen Information verglichen mit dem allgemeingültigen Verfahren zu einer höheren Erkennungsrate führt. Die Bedeutung der Farbe für die Cartoon-Bilder illustriert Abbildung 2.1. Hier sind die Farbflächen eines Cartoon-Bildes schematisch dargestellt. Die Köpfe der Enten zeigen sich in orangen Kreisen mit weißen Kreisen darüber. Zuweilen zeigen sich sogar die Augen in Form von schwarzen Punkten. Anhand dieser Informationen lassen sich die Köpfe der Enten im unteren Bild erkennen, selbst wenn keine Objektkonturen zur Verfügung stehen. Die Spezialisierung der ausgewählten Merkmale auf das vorliegende Bildmaterial hat Nachteile für die Übertragung auf weitere Anwendungsbereiche. Da nur wenige Objekte aus orangen, schwarzen und weißen Flecken bestehen, ist selbst die Erkennung von

Originalbild



Farbflächen

Abbildung 2.1: **Bedeutung der Farbe in Cartoons.** Das untere Bild zeigt die farbigen Flächen des oberen Bildes in Form von Kreisen. Kanteninformationen wurden entfernt. Um die untere Darstellung zu erreichen, wurde das obere Bild skelettiert. An den Kreuzungen von Skelettlinien wurde die Farbe und der Abstand zur nächsten Kante bestimmt. Diese sind in Form von farbigen Kreisen dargestellt.

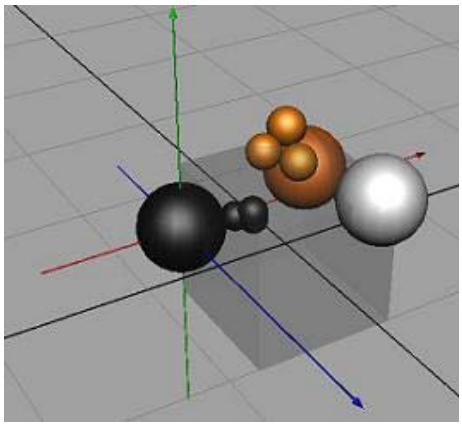


Abbildung 2.2: **Beispiele von Clustern zur Farbsegmentierung.** Die Achsen zeigen den RGB-Farbraum. Die Bereiche der Farben Orange, Weiß und Schwarz wurden durch Kugeln angenähert (Abbildung von Yuan, Xi und Lou).

älteren Cartoon-Figuren der gleichen Serie ist nicht möglich, da diese früher nur in schwarz-weiß gedruckt wurden.

Um die Farbinformation auszuwerten, werden die Stichprobenbilder nach einer Rauschfilterung durch einen 3×3-Binomialfilter zunächst segmentiert. Dabei wird jeder Bildpunkt einer von vier Farben zugeordnet. Die ersten drei Farben sind Orange, Weiß und Schwarz. Sie repräsentieren die Donald-Köpfe und werden ähnlich Abbildung 2.2 in Form von Kugeln im RGB-Farbraum definiert. Die Kugelparameter gibt Tabelle 2.1 an. Die restlichen Farben werden in der vierten Farbkategorie zusammengefaßt.

Während der Trainingsphase wird aus den segmentierten Bildern der Trainingsstichprobe ein Modell erzeugt. Dieses gibt die Wahrscheinlichkeiten für bestimmte Farben an bestimmten Bildkoordinaten an. Während der Testphase werden die Bilder der Teststichprobe mit dem Modell verglichen. Für jedes Bild wird entschieden, ob es sich um einen Donald-Kopf oder ein Hintergrundbild handelt. Hierzu wird ein zweistufiger Klassifikator eingesetzt. Auf der ersten Stufe wird für jeden Bildpunkt entschieden, ob dieser zu einem Donald-Kopf oder zum Hintergrund gehört. Auf der zweiten Stufe wird anhand der klassifizierten Bildpunkte das komplette Stichprobenbild bewertet.

Wenn für eine bestimmte an einer Bildkoordinate x, y auftretenden *Farbe* $\in$ {Orange, Weiß, Schwarz, Rest} die Bedingung

$$p(\text{"Objekt"}|\mathit{Farbe}, x, y)\varsigma_{\mathit{Farbe}} \geq p(\text{"Hintergrund"}|\mathit{Farbe}, x, y) \qquad (2.1)$$

gilt, dann wird der Punkt als Teil des Objekts betrachtet. Der Toleranzwert $\varsigma_{\mathit{Farbe}}$ wird dabei aufgabengerecht eingestellt. Andernfalls wird der Punkt dem Hintergrund zugeordnet. Die Wahrscheinlichkeit $p(\text{"Vordergrund"}|\mathit{Farbe}, x, y)$ wird mit dem Satz von Bayes auf Wahrscheinlichkeitswerte zurückgeführt, die

Farbe	**Rot**	**Grün**	**Blau**	**Radius**
Orange	255	157	70	50
Orange	255	180	0	30
Orange	255	177	70	30
Weiß	250	247	225	50
Schwarz	75	75	75	55
Schwarz	80	80	80	15
Schwarz	95	95	95	15
Schwarz	100	100	100	15
Schwarz	105	105	105	15

Tabelle 2.1: Farbcluster zur Segmentierung. Die Farben Orange, Weiß und Schwarz werden durch Kugeln im RGB-Farbraum repräsentiert. Die angegebenen Werte für Rot, Grün und Blau definieren das jeweilige Kugelzentrum. Die letzte Spalte gibt den passenden Kugelradius an.

während der Trainingsphase berechnet werden:

$$p(\text{"Objekt"}|\mathit{Farbe}, x, y) = \frac{p(\mathit{Farbe}, x, y|\text{"Objekt"})p(\text{"Objekt"})}{p(\mathit{Farbe}, x, y)}$$

Dabei ist $p(Farbe, x, y|\text{"Objekt"})$ die Wahrscheinlichkeit, daß an der Koordinate x, y eines Vordergrundbildes eine bestimmte Farbe auftritt, $p(\text{"Objekt"})$ ist die generelle Wahrscheinlichkeit, daß ein Punkt zu einem Vordergrundbild gehört, und $p(\mathit{Farbe}, x, y)$ ist die Wahrscheinlichkeit daß eine bestimmte Farbe an einer Koordinate (Vorder- oder Hintergrund) auftritt. Der Vergleich in Formel 2.1 kann auch über das Verhältnis der beiden Terme geschehen. Für Entenbilder ist

$$\frac{p(\text{"Objekt"}|\mathit{Farbe}, x, y)}{p(\text{"Hintergrund"}|\mathit{Farbe}, x, y)} \geq \frac{1}{\varsigma_{\mathit{Farbe}}}.$$

Ansonsten handelt es sich um einen Punkt des Hintergrunds. Mit dem Satz von Bayes muß also während der Klassifikation der folgende Term ausgerechnet werden:

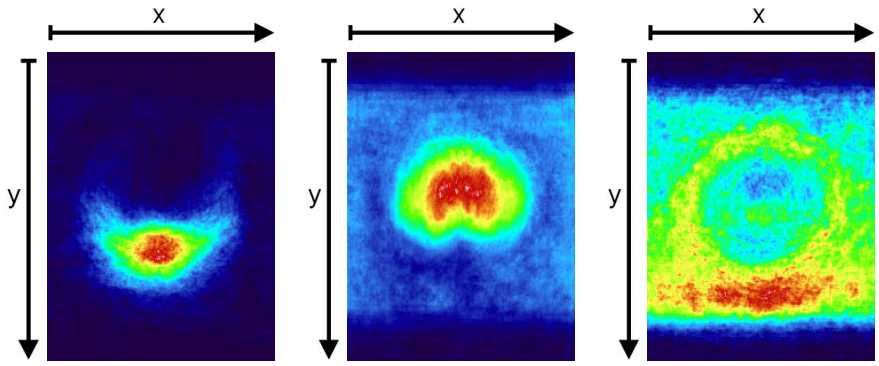
$$\frac{p(\text{"Objekt"}|\mathit{Farbe}, x, y)\varsigma_{\mathit{Farbe}}}{p(\text{"Hintergrund"}|\mathit{Farbe}, x, y)} = \frac{p(\mathit{Farbe}, x, y|\text{"Objekt"})\varsigma_{\mathit{Farbe}}p(\text{"Objekt"})}{p(\mathit{Farbe}, x, y|\text{"Hintergrund"})p(\text{"Hintergrund"})} \tag{2.2}$$

Für Hintergrundbilder muß davon ausgegangen werden, daß die Farben an allen Bildkoordinaten gleich häufig auftreten, d.h. daß die Wahrscheinlichkeit einer bestimmten Farbe nicht von der Position im Bild abhängt. Formel 2.2 vereinfacht sich dann zu

$$\frac{p(\text{"Objekt"}|\mathit{Farbe}, x, y)}{p(\text{"Hintergrund"}|\mathit{Farbe})} = \frac{p(\mathit{Farbe}, x, y|\text{"Objekt"})\varsigma_{\mathit{Farbe}}p(\text{"Objekt"})}{p(\mathit{Farbe}|\text{"Hintergrund"})p(\text{"Hintergrund"})}. \tag{2.3}$$

In der Trainingsphase werden die Wahrscheinlichkeiten für die zweite Hälfte von Gleichung 2.3 bestimmt. Für die Stichprobenbilder werden die Werte

$$p(\text{"Objekt"}) = 0{,}21$$



$p(\text{"Orange"}, x, y)$ $p(\text{"Weiß"}, x, y)$ $p(\text{"Schwarz"}, x, y)$

Abbildung 2.3: **Farbverteilungen von Donald-Köpfen.** Die Diagramme geben jeweils normiert auf das Maximum die Verteilung der Farben über der Bildebene an. Die Verteilungen beziehen sich auf die Positivbeispiele der Trainingsstichprobe.

und

$$p(\text{"Hintergrund"}) = 0{,}79$$

durch Auszählen der als Orange, Weiß und Schwarz segmentierten Punkte in den Vordergrundbildern ermittelt. Das Auszählen der verschiedenen Farbpunkte der Hintergrundbilder liefert folgende Werte:

$$\begin{aligned} p(\text{"Orange"}|\text{"Hintergrund"}) &= 0{,}01 \\ p(\text{"Weiß"}|\text{"Hintergrund"}) &= 0{,}31 \\ p(\text{"Schwarz"}|\text{"Hintergrund"}) &= 0{,}14 \end{aligned} \tag{2.4}$$

Die Berechnung der Terme $p(\textit{Farbe}, x, y|\text{"Vordergrund"})$ geschieht ebenfalls über das Auszählen der verschiedenfarbigen Punkte. Da hier die Bildkoordinate in die Berechnung eingeht, spielt die Bildgröße eine Rolle. Um ein genaueres Modell zu erzeugen werden daher alle Bilder auf eine gemeinsame Größe skaliert. Dies erhöht die Anzahl der korrekten Klassifikationen. Die robustere Unterscheidung zwischen Vorder- und Hintergrundausschnitten geht allerdings zu Lasten der Lokalisation. Da das Modell nun von der Größe des Bildausschnitts unabhängig ist, wird es schwieriger, in einem großen Bild konkrete Bildausschnitte zu wählen, die auf die Anwesenheit einer Cartoon-Figur getestet werden können. Die Methode ist daher nur bezüglich der reinen Klassifikationsleistung mit dem in den folgenden Kapiteln vorgestellten allgemeingültigen Verfahren vergleichbar. Nach der Auszählung der Farbpunkte der umskalierten Stichprobenbilder ergeben sich die in Abbildung 2.3 gezeigten Verteilungen.

Aufgrund der hohen Auftrittshäufigkeiten von Weiß und Schwarz in den Hintergrundbildern, ergeben sich in Gleichung 2.3 so kleine Werte, daß weiße und

schwarze Pixel im Gegensatz zu orangen Pixeln praktisch nie dem Vordergrund zugeordnet werden. Dies ist für die zweite Stufe des Klassifikators nachteilig, da diese die relativen Häufigkeiten und Positionen der Farbpunkte auswertet. Der Effekt wird daher mit Hilfe der Toleranzwerte ς_{Farbe} kompensiert. Diese beeinflussen die Häufigkeit, mit der Punkte als Vordergrund klassifiziert werden. Experimentell haben sich die Werte $\varsigma_{\text{Orange}} = 1, \varsigma_{\text{Weiß}} = 4$ und $\varsigma_{\text{Schwarz}} = 3$ als optimal erwiesen. Da bei dieser Einstellung die in Abbildung 2.3 dargestellten Farbverteilungen die Positionen der als Vordergrund klassifizierten Punkte beeinflussen, gehen nur die Bildpunkte in die nächste Stufe des Klassifikators ein, die mit der allgemeinen Objektgeometrie übereinstimmen. Auf der zweiten Stufe werden nun die relativen Häufigkeiten der Farben verglichen, in denen Bildpunkte auftreten.

Ein Bild wird dann als Donald-Kopf erkannt, wenn die Gesamtzahl der Vordergrundpunkte über einer Schwelle $\vartheta_{\text{Objekt}}$ liegt und einzelne Farben das Ergebnis nicht zu stark dominieren. Um letzteres zu prüfen werden die Schwellen $\vartheta_{\text{Orange}}, \vartheta_{\text{Weiß}}$ und $\vartheta_{\text{Schwarz}}$ eingeführt, mit denen der Anteil der Punkte einer bestimmten Farbe an allen als Vordergrund klassifizierten Punkten bewertet wird. Neben den Einzelhäufigkeiten der verschiedenfarbigen Punkte ist möglicherweise auch das Verhältnis der einzelnen Farben zueinander relevant. Dazu werden die Schwellen $\vartheta_{\text{Orange+Weiß}}, \vartheta_{\text{Weiß+Schwarz}}$ und $\vartheta_{\text{Orange+Schwarz}}$ definiert, die sich jeweils auf die Summen der Anteile von Punkten zweier verschiedener Farben an allen Vordergrundpunkten beziehen. Ein Stichprobenelement wird nur dann als Objekt klassifiziert, wenn die jeweiligen Farbanteile nicht überschritten werden. Die Schwellwerte definieren einen 7-dimensionalen Parameterraum. Eine unabhängige Optimierung der Schwellen auf eine maximale Anzahl richtig erkannter positiver und negativer Beispiele ergibt die Schwellen:

$$\begin{aligned}
\vartheta_{\text{Objekt}} &= 0{,}15 \\
\vartheta_{\text{Orange}} &= 0{,}57 \\
\vartheta_{\text{Weiß}} &= 0{,}81 \\
\vartheta_{\text{Schwarz}} &= 0{,}81 \\
\vartheta_{\text{Orange+Weiß}} &= 0{,}93 \\
\vartheta_{\text{Weiß+Schwarz}} &= 0{,}97 \\
\vartheta_{\text{Orange+Schwarz}} &= 0{,}92
\end{aligned} \tag{2.5}$$

Mit dieser Einstellung werden bereits 86% aller Bilder der Teststichprobe korrekt klassifiziert. Anschließend werden die Schwellen durch ein einfaches Diffusionsverfahren in Kombination anhand der Trainingsstichprobe optimiert. Nach einer ausreichenden Anzahl von Optimierungsschritten ergibt sich folgendes Er-

gebnis:

$$\begin{aligned}
\vartheta_{\text{Objekt}} &= 0{,}09 \\
\vartheta_{\text{Orange}} &= 0{,}62 \\
\vartheta_{\text{Weiß}} &= 0{,}79 \\
\vartheta_{\text{Schwarz}} &= 1{,}00 \\
\vartheta_{\text{Orange+Weiß}} &= 1{,}00 \\
\vartheta_{\text{Weiß+Schwarz}} &= 1{,}00 \\
\vartheta_{\text{Orange+Schwarz}} &= 1{,}00
\end{aligned} \tag{2.6}$$

Die Schwellwerte von 100% zeigen, daß die Schwellwerte für die Farbkombinationen nicht benötigt werden. Mit den optimierten Parametern ergibt sich auf der Trainingsstichprobe eine Rate von 94% für korrekt erkannte Vorder- und Hintergrundbilder. Für die Teststichprobe ergibt sich eine etwas geringere Rate von 92%.

Die Ergebnisse des speziell auf die Cartoon-Bilddaten zugeschnittenen Erkennungssystems lassen sich wie folgt zusammenfassen:

- Der Klassifikator erreicht eine sehr hohe Erkennungsrate.
- Die Ausnutzung von Expertenwissen bei der Auslegung des Klassifikators auf das vorliegende Bildmaterial schränkt die Variabilität ein. Schwarz-Weiß-Bilder lassen sich hier beispielsweise nicht mehr erkennen. Dadurch lassen sich nicht einmal ältere Darstellungen des trainierten Objekts erkennen.
- Die Größennormierung in der Trainingsphase erschwert die Lokalisation.

Trotz der hohen Erkennungsrate ist das Verfahren sehr eingeschränkt, da die Übertragung des Verfahrens auf anderes Bildmaterial sehr problematisch ist. Als nächstes geht es daher darum, wie ohne den Einsatz von anwendungsspezifischem Vorwissen Objekte erkannt werden können.

Kapitel 3

Stand der Forschung

Marr [Mar82] berichtet, daß die ersten Ideen zum Aufbau künstlicher Objekterkennungssysteme stark aus der Psychologie und der Neurophysiologie motiviert waren. Das Wissen darüber, wie Menschen Objekte wahrnehmen und erkennen, beruhte bis in die fünfziger Jahre vor allem auf psychologischen Experimenten und theoretischen Überlegungen, und war dementsprechend abstrakt. Überraschende neue Erkenntnisse lieferten jedoch Messungen der Aktivierung von Gehirnzellen bei Tieren. Insbesondere wurde gezeigt, daß einzelne Zellen ganze Objekte erkennen und komplexe Verhaltensweisen hervorrufen können. Zudem sind solche Zellen teilweise bereits in der Netzhaut zu finden, was auf weniger abstrakte Vorgänge hindeutet als bis dahin angenommen.

Die Versuche, das Verhalten von Gehirnzellen im Rechner zu simulieren und zur Mustererkennung zu nutzen, führten zur Entwicklung der künstlichen neuronalen Netze, die in der verbreiteten rückkopplungsfreien Form auf Rosenblatts [Ros58] Perzeptron-Netzwerk zurückgehen [MP88]. Die kritische Analyse der Leistungsfähigkeit konnektionistischer Methoden und insbesondere des Perzeptrons durch Minsky und Papert 1969 [MP69] verhinderte jedoch bis zur Einführung des Backpropagation-Verfahrens 1986 [RHW86] jedes breitere Interesse an diesen Ansätzen.

Stattdessen kommt ein bodenständigerer Ansatz in Mode. Anregungen durch neurophysiologische Erkenntnisse zusammen mit einer genaueren Abschätzung der rechentechnischen Möglichkeiten führen Marr [Mar82] zu einem strukturellen Objekterkennungsmodell, das bis in die Neunziger Jahre für viele Wissenschaftler richtungsweisend ist und wie folgt aussieht: Ausgehend von den Intensitätswerten eines Bildes wird zunächst eine erste Skizze (*Primal Sketch*) erzeugt. Diese zeigt Kanten und Farbflächen sowie ihren geometrischen Zusammenhang in der Bildebene an. Aus der ersten Skizze wird eine $2\frac{1}{2}$-dimensionale Darstellung berechnet. Diese beschreibt die räumliche Anordnung von sichtbaren Flächen relativ zum Betrachter. Als letztes wird die $2\frac{1}{2}$-dimensionale Darstellung in eine Menge volumetrischer Primitive überführt. Es ergibt sich eine dreidimensionale Szenenbeschreibung, die objektzentriert ist und auch die Rückseiten von Objekten angibt.

Wie Murase und Nayar [NMN96] zusammenfassen, ist eine Objekterkennung aufgrund von 3D-Informationen jedoch nur unter sehr kontrollierten Bedingungen möglich. Zum einen wird die Erzeugung von geometrischen Modellen aus Linien durch die Rauschempfindlichkeit lokaler Operatoren erschwert, zum anderen ist das manuelle Erstellen von 3D-Modellen extrem aufwendig. Murase und Nayar schlagen stattdessen einen vollständig erscheinungsbasierten Ansatz vor, der die 3D-Modellierung umgeht. Ein Objekt wird bei diesem Ansatz durch eine Sammlung von Bildern beschrieben, die das Objekt aus verschiedenen Perspektiven und in verschiedenen Beleuchtungen zeigen. Ein unbekanntes Objekt kann dann dadurch erkannt werden, daß es der ähnlichsten gespeicherten Ansicht zugeordnet wird, sofern eine gewisse Mindestähnlichkeit besteht. Diese Methode hat den Vorteil, daß vor der Entscheidung, um welches Objekt es sich handelt, keine fehleranfälligen Entscheidungen auf einer niedrigeren Verarbeitungsebene gefällt werden müssen.

Aktuelle Arbeiten behandeln daher verstärkt erscheinungsbasierte anstelle von strukturellen Methoden, wofür auch neuere neurophysiologische Erkenntnisse sprechen. In letzter Zeit ist dabei eine verstärkte Tendenz zu teilebasierten Ansätzen zu erkennen, die robuster gegen die Verdeckung von Objekten durch andere Objekte ist. Die Modellierung von Objekten als Graphen oder Netze von Teilen trägt dabei zum Teil deutliche konnektionistische Züge [SWP05a].

Dieses Kapitel behandelt daher zwei Themen intensiver: Da die Natur aktuellen Objekterkennungssystemen im Rechner weit überlegen ist und Ideen für viele aktuelle Systeme liefert [SWP05a, Sel01, MN95, OSB06], werden zunächst neuere Erkenntnisse über das Sehsystem von Menschen und Tieren dargestellt. Anschließend werden die derzeitigen Entwicklungen bei der erscheinungsbasierten Objekterkennung im Rechner geschildert. Es werden aktuelle Fragestellungen erörtert und die Zielsetzung der vorliegenden Arbeit konkretisiert.

3.1 Objekterkennung bei Menschen und Affen

Es folgt eine Beschreibung des Sehsystems von der Netzhaut über den Sehnerv bis in den inferioren Temporallappen, den Teil des Gehirns, der nach derzeitigem Wissen für die Objekterkennung verantwortlich ist. Die Ausführungen beziehen sich meistens auf das Sehsystem von Affen, hin und wieder aber auch auf das von Menschen und Katzen.

3.1.1 Visuelle Verarbeitungsströme im Gehirn

Die Verarbeitung optischer Reize beginnt bereits in der Netzhaut mit einer starken Reduktion der visuellen Eindrücke auf die wichtigsten Informationen. Diese werden vom Sehnerv an das Corpus geniculatum laterale geschickt, das die visuellen Informationen in Bezug auf die Gesamtsituation des Organismus bewertet und entsprechend an das Großhirn weiterleitet, wo die Information zunehmend bewußter verarbeitet wird. Abbildung 3.1 zeigt eine einigermaßen gängige Unterteilung eines Affengehirns in verschiedene visuelle Bereiche. Im

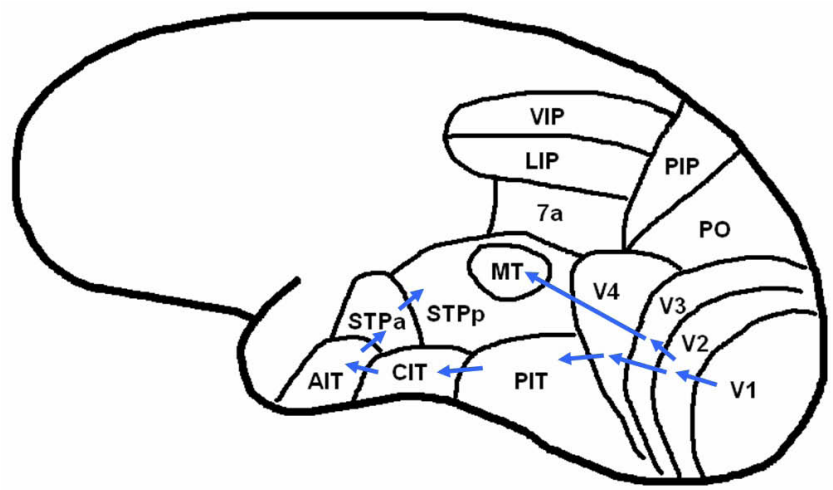


Abbildung 3.1: **Visuelle Gehirnareale des Affen.** Nach dem primären visuellen Kortex V1 teilt sich der Verarbeitungsstrom auf. Der Pfad V2–V4–PIT–CIT–AIT wird *ventraler Strom* genannt und ist für die Objektkategorisierung zuständig. Der *dorsale Strom* V1–V2–V3–MT ist dagegen eher für die räumliche Verarbeitung zuständig.

folgenden wird nur auf den *ventralen Verarbeitungsstrom* eingegangen, der vom primären visuellen Kortex V1 über die Bereiche V2 und V4 in den inferioren Temporallappen (PIT, CIT, AIT) führt. Dies ist die letzte rein visuelle Verarbeitungsstation und nach derzeitigem Erkenntnisstand findet dort – zumindest für einige Objektklassen – eine Zuordnung der visuellen Information an einzelne Objekte statt.

3.1.2 Netzhaut und Sehnerv[1]

Das durch die Pupille in das Auge einfallende Licht wird in der Netzhaut in Nervensignale umgewandelt. Die Netzhaut ist etwa einen Viertel Millimeter dick und besteht aus drei Schichten von Nervenzellen. Die Zellen der dritten Schicht sind lichtempfindlich und werden in Stäbchen und Zapfen unterteilt. Auf die Schicht aus Stäbchen und Zapfen folgt eine lichtundurchlässige Schicht aus schwarz pigmentierten Zellen. Abbildung 3.2 zeigt den Aufbau der Netzhaut schematisch.

Stäbchen sind weitaus zahlreicher als Zapfen. Sie enthalten das Pigment Rhodopsin, das einen weiten Spektralbereich absorbiert, wobei die maximale Lichtempfindlichkeit bei einer Wellenlänge von 510 nm liegt. Das entspricht der Farbe Grün. Die Signale der Stäbchen werden allerdings nicht mit denen von

[1] Die Informationen dieses Abschnitts stammen weitgehend aus [Hub88].

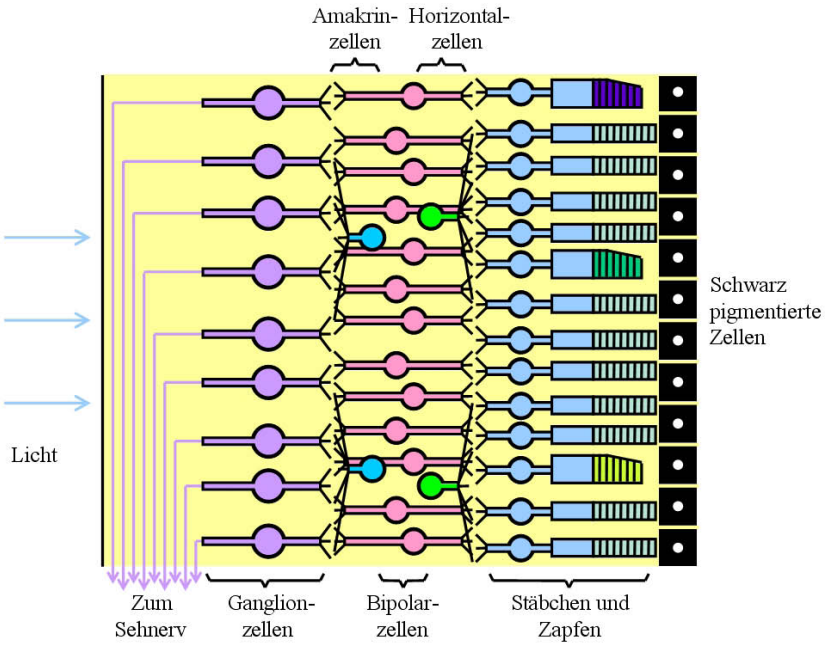


Abbildung 3.2: **Aufbau der Netzhaut.** Das in das Auge einfallende Licht durchdringt die oberen drei Nervenschichten der Netzhaut und wird von den pigmentierten Stäbchen und Zapfen absorbiert. Diese wandeln die Lichtsignale in Nervenpotentiale um, welche über die Bipolarzellen in die Ganglionzellen weitergeleitet werden. Die Nervenenden der Ganglionzellen bündeln sich zum Sehnerv.

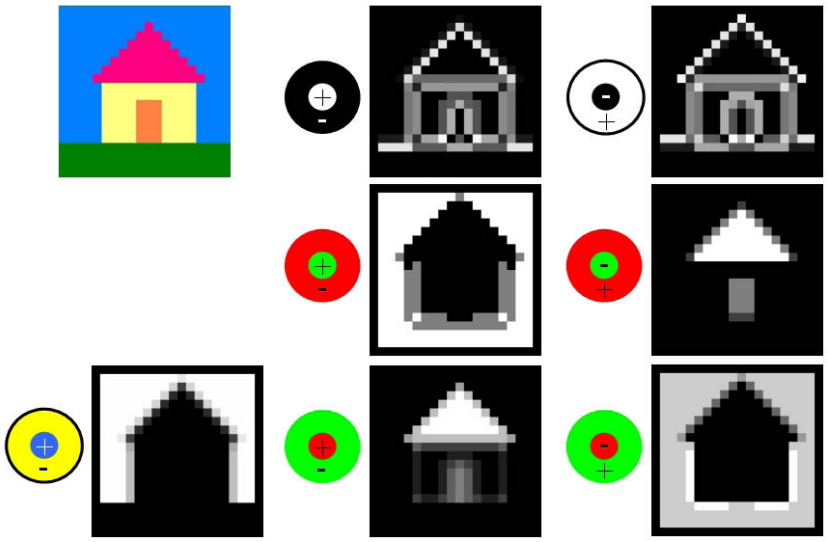


Abbildung 3.3: **Simulation der Ganglionzellen durch einen Laplacefilter.** Links oben ist das Originalbild. Die übrigen Bilder zeigen die simulierte Ausgabe verschiedener Ganglionzellen. Die Kreise neben den Bildern zeigen an, um welche Zellen es sich handelt: Zellen mit einem Plus in der Mitte sind On-Center-Zellen, Zellen mit einem Minus sind Off-Center-Zellen. In der oberen Reihe ist das Ergebnis für nicht-farbsensible Zellen abgebildet. Bei den übrigen Bildern gibt die Farbe der Kreise an, welche Art von Zapfen hemmend oder verstärkend wirkt. Die Simulation der On-Center-Zellen geschieht durch Faltung des Originalbildes mit der Maske $[-1,\ 2,\ -1] + [-1,\ 2,\ -1]^\top$. Für Off-Center-Zellen wird analog die Maske $[1,\ -2,\ 1] + [1,\ -2,\ 1]^\top$ eingesetzt. Bei den farbsensiblen Zellen wirken die Gewichte nur auf einzelne Farbkanäle. Es ist zu beachten, daß sich der blaue Hintergrund im Originalbild nur im Blaukanal vom grünen Untergrund unterscheidet. Man kann erkennen, daß die farbsensiblen Zellen auf Flächen bestimmter Farben ansprechen sowie auf Schwarz-Weiß-Kanten, während die nicht farbsensiblen Zellen nur auf Kanten ansprechen.

für andere Wellenlängen empfindlichen Zellen verglichen. Daher tragen sie nicht zum Farbeindruck, sondern nur zum Helligkeitseindruck bei. Da Stäbchen schon bei schwachem Licht ansprechen, liefern sie die den Hauptbeitrag für das Sehen bei Nacht.

Das Farbensehen beruht auf den Signalen der Zapfen. Diese liegen in drei Varianten vor, die sich in ihrer spektralen Empfindlichkeit unterscheiden. Die maximalen Lichtempfindlichkeiten liegen bei den drei Typen von Zapfen bei 430 nm (Violett), 530 nm (Blaugrün) und 560 nm (Gelbgrün). Die Spektralbereiche, in denen die Zapfen lichtempfindlich sind, sind allerdings breitbandig und überschneiden sich stark – sowohl untereinander, als auch mit dem der Stäbchen. Entsprechend ihrer Spezialisierung auf kurze (short), mittlere (middle) oder lange (long) Wellenlängen, werden die Zapfen *S-Cone, M-Cone* und *L-Cone* genannt.

Bevor das Licht auf Stäbchen oder Zapfen trifft, durchquert es die Schichten von Nervenzellen. Dies verursacht eine leichte Unschärfe des Bildes. Im Zentrum der Netzhaut, der Fovea, ist die Nervenschicht deutlich dünner, sodaß die Auswirkungen dort geringer sind. In diesem etwa 0,5 mm großen Bereich befinden sich ausschließlich Zapfen.

Die Nervenschicht über den Stäbchen und Zapfen enthält die *Bipolarzellen*. In der Fovea mündet jeder Zapfen in eine einzelne Bipolarzelle. In den äußeren Bereichen der Netzhaut münden dagegen mehrere Zapfen und Stäbchen in eine Bipolarzelle. Die Bipolarzellen wiederum münden in die *Ganglionzellen* der obersten Nervenschicht. Die Nervenenden (*Axone*) der Ganglionzellen überqueren die Netzhaut, bündeln sich an einer Stelle und verlassen als Sehnerv das Auge. Im Bereich der Fovea ist eine Bipolarzelle einer Ganglionzelle zugeordnet. Außerhalb der Fovea führen mehrere Bipolarzellen in einer Ganglionzelle. Neben Bipolarzellen enthält die Mittlere Nervenschicht auch *Horizontalzellen* und *Amakrinzellen*. Horizontalzellen empfangen über einen weiten Bereich Signale von Zapfen und Stäbchen und leiten diese an Bipolarzellen weiter. Amakrinzellen fassen die Signale größerer Bereiche von Bipolarzellen zusammen und leiten sie an Ganglionzellen. Inzwischen werden mindestens 10 Typen von Bipolarzellen, 20 Typen von Ganglionzellen und 30–40 Amakrinzellen unterschieden, welche unterschiedliche und oft noch unklare Aufgaben haben. Möglicherweise haben einige Arten von Amakrinzellen etwas mit Bewegungserkennung zu tun. Es wird auch vermutet, daß Amakrinzellen zu einer Schärfung des Bildes beitragen [Kol03]. Außerhalb der Fovea belegen die Zapfen und Stäbchen, die in eine Ganglionzelle münden, eine Fläche von etwa einem Quadratmillimeter. Der geschilderte Aufbau der Nervenschichten bildet die etwa 125 Millionen Stäbchen und Zapfen auf nur 1 Millionen Axone im Sehnerv ab.

Der Bereich der Netzhaut, der Signale an eine Nervenzelle schickt und dort Änderungen hervorruft, wird *rezeptives Feld* genannt. Ganglionzellen treten überwiegend in zwei Varianten auf, die etwa gleich häufig sind und sich in ihrem Ansprechverhalten auf die Beleuchtung ihrer rezeptiven Felder unterscheiden. Das unterschiedliche Ansprechverhalten beruht darauf, daß Synapsen, die Verbindungen zweier Nervenzellen, das Ausgangssignal einer Zelle entweder verstärken oder hemmen können. Die beiden Arten von Ganglionzellen werden

On-Center- und *Off-Center-Zellen* genannt. Sie machen etwa 90 Prozent aller Ganglionzellen aus. Die rezeptiven Felder sind bei beiden Typen kreisförmig und überlappen sich für benachbarte Zellen stark. On-Center-Zellen liefern die stärksten Ausgangssignale, falls ein kleiner, kreisförmiger bis elliptischer Bereich in der Mitte des rezeptiven Feldes beleuchtet wird, während der äußere Bereich dunkel ist. Off-Center-Zellen reagieren dagegen nur auf einen dunklen Punkt in einer hellen Umgebung. Werden die rezeptiven Felder der Ganglionzellen gleichmäßig beleuchtet, sprechen weder On-Center- noch Off-Center-Zellen an. Bei On-Center-Zellen ist der mittlere Bereich des rezeptiven Feldes offenbar hauptsächlich über verstärkende Synapsen verschaltet, während der Randbereich überwiegend über hemmende Synapsen verbunden ist. Bei Off-Center-Zellen ist die Verbindung genau entgegengesetzt. Bei einer gleichmäßigen Beleuchtung heben sich die hemmenden und verstärkenden Signale gegenseitig auf. Da die Feldzentren nicht vollständig rund sind, sind die Ganglionzellen in gewissem Maß richtungsselektiv [PTRL03]. Aufgrund ihrer großen Ausdehnung geht man davon aus, daß Horizontalzellen für den Rand der rezeptiven Felder verantwortlich sind. Horizontalzellen empfangen auch Signale beachbarter Horizontalzellen sowie Neurotransmitter aus dem Bereich zwischen Ganglionzellen und Bipolarzellen und regulieren dadurch die allgemeine Helligkeits- und Farbempfindlichkeit [Kol03].

Abbildung 3.3 gibt einen groben Eindruck von den Ausgabesignalen der Ganglienzellen für farbiges und nicht-farbiges Sehen, wobei das verwendete Modell allerdings sehr einfach ist. Die nicht farbsensiblen Ganglionzellen geben nur den Kontrast an Kanten im Originalbild wieder. Hubel [Hub88] vermutet als Zweck dieser Arbeitsweise, das Sehsystem unabhängig von absoluten Helligkeiten zu machen, da diese stark schwanken können, ohne daß ein Mensch das Aussehen von Objekten entscheidend anders empfindet. Simoncelli und Olshausen erklären dieses Verhalten mit einer Anpassung an die lokale Bildstatistik [SO01], die der Dekorrelation der Signale benachbarter Zellen dient. Sie verweisen dazu auf quantitative Studien zu den Augen von Fliegen.

Der Detailgrad des Sehens hängt wesentlich von der Größe der Zentren der rezeptiven Felder der Ganglionzellen ab. Die kleinsten beim Affen gemessenen Zentren haben eine Größe von 10 μm bzw. 2 Winkelminuten. Der Abstand zweier Zapfen in der Fovea liegt bei 2,5 μm. Die Größe der Zentren nimmt zum Rand der Netzhaut hin zu. Dort können die Zentren Tausende von Zapfen und Stäbchen umfassen und einen Winkelausschnitt im Minutenbereich einnehmen.

An den Ganglionzellen teilt sich das Sehsystem in verschiedene Pfade auf, die mindestens bis in die Region V1 des visuellen Kortex getrennt verlaufen: Die Midget-Ganglionzellen sind dabei für das Farbensehen zuständig. Im Bereich der Fovea treten On-Center- und Off-Center-Zellen sowohl für rote Zentren in grüner Umgebung als auch grüne Zentren in roter Umgebung auf. Eine Minderheit von 2-3 Prozent der Zellen gehört zur Kategorie Off-Center für blaue Zentren in gelber Umgebung. Vom Rand der Fovea bis in die Peripheriebereiche der Netzhaut hin nimmt die Farbsensibilität ab, sodaß die Ganglionzellen nur noch ein farbunspezifisches On- oder Off-Center-Verhalten zeigen. Dacey [Dac00] erklärt dies als Effekt einer zufälligen Verteilung von L- und M-Zapfen

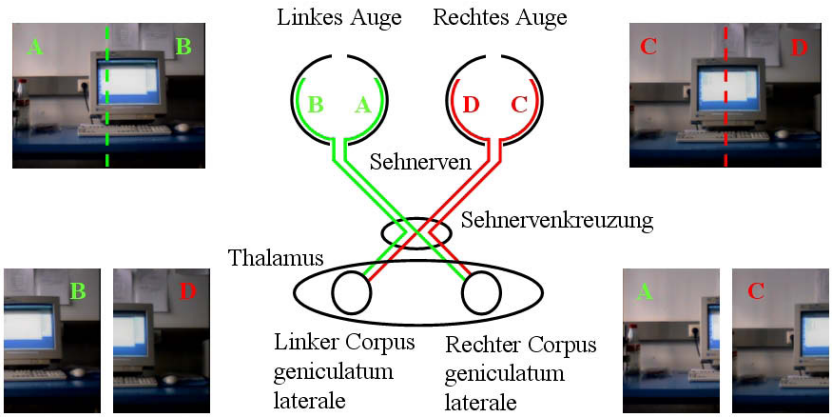


Abbildung 3.4: **Trennung des linken und rechten Gesichtsfeldes in der Sehnervenkreuzung.** Das linke Gesichtsfeld des linken und rechten Auges wird zum rechten Corpus geniculatum laterale geführt, das rechte Gesichtsfeld des linken und rechten Auges zum linken Corpus geniculatum laterale.

in Kombination mit der zunehmenden Größe der rezeptiven Felder zum Netzhautrand hin. In der Fovea bestehen die Zentren der rezeptiven Felder nur aus einem Zapfen. Je nach Zapfentyp ist das Zentrum also ausschließlich für Rot oder Grün empfindlich. In der Netzhautperipherie bestehen die Zentren der rezeptiven Felder dagegen aus mehreren Zapfen. Ein Farbeindruck ergibt sich nur dort, wo ein Zapfentyp überwiegt. Dieser Aufbau könnte darin begründet sein, daß die Unterscheidung zwischen Rot und Grün genetisch jung ist im Vergleich zu Blau, was sich auch in einem ähnlichen molekularen Aufbau der Pigmente für Rot und Grün zeigt [Dac00].

Neben den Midget-Zellen treten auch sog. *Parasol*-Ganglionzellen auf. Diese übertragen die farbunspezifischen Signale der Stäbchen in Form von On- oder Off-Center-Zellen. Ein dritter Typ von Ganglionzellen sind die *Small Bistratified Cells*, die ausschließlich in der On-Center-Variante mit blauem Zentrum auftreten.

Die Funktionsweise der Netzhaut ergibt sich nicht nur aus den elektrischen Signalen, die über Synapsen zwischen den Zellen ausgetauscht werden, sondern auch aus der Diffusion von Neurotransmittern über größere Strecken. Darüberhinaus wurden kürzlich Ganglionzellen entdeckt, die selbst lichtempfindlich sind. Sie wurden mit Pupillenbewegungen in Verbindung gebracht [GMP$^+$07] und rufen anscheinend eine Art Helligkeitseindruck hervor [ZHP$^+$07], auch wenn Menschen, denen aufgrund einer Krankheit oder genetisch bedingt Zapfen und Stäbchen fehlen, ansonsten blind sind.

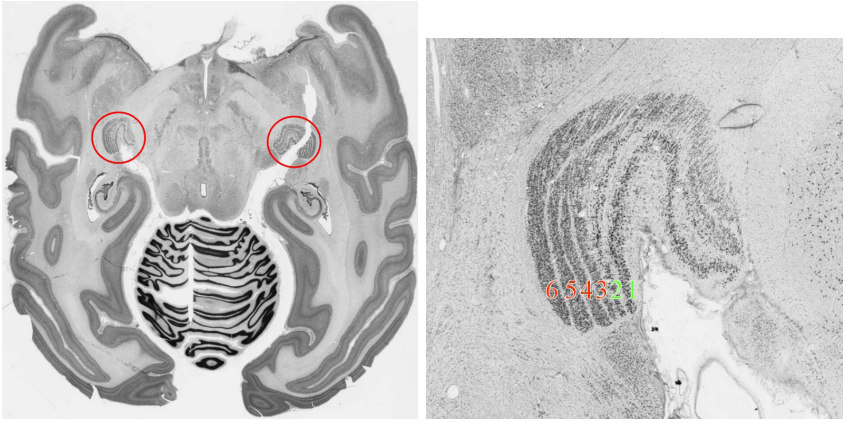


Abbildung 3.5: **Das Corpus geniculatum laterale.** Links: Lage im Gehirn des Affen. Rechts: Vergrößerung des linken Teils des Corpus geniculatum laterale. Am linken Rand sind die sechs Ebenen zu erkennen. Dem linken Auge sind die Ebenen 5, 3 und 2 zugeordnet, das rechte Auge den Ebenen 6, 4 und 1. Die magnozellulären Ebenen tragen die Nummern 1 und 2 [MSBJ07].

3.1.3 Die Sehnervenkreuzung

Hinter den Augen trennen sich die Axone des Sehnervs in Bündel für das jeweils linke und rechte Gesichtsfeld auf. Das linke Gesichtsfeld ist der Teil der Umgebung, der links von der optischen Achse sichtbar ist. Das rechte Gesichtsfeld ist der rechte Teil. Die Aufteilung geschieht in der Sehnervenkreuzung (lat. *Chiasma opticum*). Abbildung 3.4 veranschaulicht dies. Etwa 90 Prozent der nun *Sehstrang* (lat. *Tractus opticus*) genannten Axone verläuft von der Sehnervenkreuzung in das linke bzw. rechte Corpus geniculatum laterale (dt. Bez. *seitlicher Kniehöcker*). Diese sind Teil des Thalamus. Die Axone des linken Gesichtsfeldes führen zu dem Corpus geniculatum laterale in der rechten Gehirnhälfte, die Axone des rechten Gesichtsfeldes laufen in das Corpus geniculatum laterale der linken Gehirnhälfte. Die übrigen Axone führen in Bereiche des Gehirns, die beispielsweise für den Pupillenreflex oder Augenbewegungen zuständig sind [Hub88, SSS06].

3.1.4 Das Corpus geniculatum laterale

Der Thalamus enthält neben den für das Sehen zuständigen Bereichen auch Bereiche für das Hören, Fühlen und Schmerzempfinden. Der Thalamus stellt über diese Bereiche die Verbindung der Sinnesorgane mit dem Großhirn her. Daneben existieren Bereiche, die für Schlafrhythmus, Angstreaktion, Appetit, Sexualtrieb und Motorik zuständig sind. Eine Hauptaufgabe des Thalamus besteht in der situationsabhängigen Filterung der Sinneseindrücke auf dem Weg

ins Großhirn. Der Thalamus entscheidet daher, welche Sinneseindrücke bewußt wahrgenommen werden.

Das Corpus geniculatum laterale besteht aus sechs übereinander geschichteten Ebenen (siehe Abb. 3.5). Dabei bildet jede Ebene das vollständige Gesichtsfeld ab, für das das Corpus geniculatum laterale einer Gehirnhälfte zuständig ist. Eine Ebene ist ungefähr vier bis zehn oder mehr Zellen dick. Die Ebenen sind abwechselnd dem Gesichtsfeld des linken und des rechten Auges zugeordnet. Die unteren beiden Ebenen enthalten dickere Zellen und werden magnozelluläre Schichten (M-Schichten) genannt. Die oberen Ebenen werden entsprechend parvozelluläre Schichten (P-Schichten) genannt. Hinter jeder M- oder P-Schicht befindet sich eine Zwischenschicht, die sog. *koniozelluläre* Schicht. Die Zellen dieser Schichten sind die *K-Zellen*. Die koniozellulären Schichten werden wie die M- und P-Schichten von 1 bis 6 durchnumeriert.

Beinahe alle 1,5 Millionen Zellen des Corpus geniculatum laterale erhalten Signale aus den Axonen des Sehstrangs und beinahe alle sind über ihre Axone weiter mit der primären Sehrinde V1 im Großhirn verbunden. Die Zellen erhalten allerdings wohl vorwiegend Signale aus anderen Bereichen, z.B. aus dem Stammhirn oder der Sehrinde. Viele sind mit ihren Nachbarn verknüpft. In Wachphasen werden die vom Auge kommenden Signale praktisch unverändert an V1 weitergeleitet. In Schlafphasen erzeugt der Thalamus selbst Signale [PMS+05, Hub88].

Die Zellen aus den parvozellulären Ebenen 3–6 erhalten Signale von den Midget-Ganglionzellen, vorwiegend aus der Fovea. Sie sind hochsensibel für Farbe aber relativ gutmütig gegenüber leichten Helligkeitsunterschieden. Sie besitzen die kleinsten Zentren rezeptiver Felder und liefern daher eine hohe Detailschärfe, z.B. für die Erkennung von Texturen. Auf Signale reagieren sie relativ langsam. Die Zellen aus den magnozellulären Ebenen 1 und 2 gelten dagegen als farbenblind, da sie Signale von den Parasol-Ganglionzellen empfangen. Sie ermöglichen auch nur eine geringe Schärfe, da die Zentren der rezeptiven Felder groß sind. Dafür sprechen sie schnell an und sind hochsensibel gegenüber kleinen Helligkeitsunterschieden sowie Bewegungen [Wal94, Bau98].

Obwohl die K-Zellen überwiegend in den dünnen Zwischenschichten hinter den P- oder M-Schichten auftreten, kommen einige dieser Zellen auch in den M- und P-Schichten vor, insbesondere in den Schichten 3 und 4. Sie bilden das komplette Gesichtsfeld ab, am detailreichsten jedoch den Bereich der Fovea. K-Zellen sind immer je einem Auge zugeordnet, demselben wie die darüberliegende P- oder M-Schicht. Eine Ausnahme stellt die koniozelluläre Schicht 2 dar, die selbst in zwei Schichten unterteilt ist, eine für jedes Auge. K-Zellen sind farbselektiv, haben aber homogene rezeptive Felder ohne Gegensätze zwischen den Zentren und den Rändern. In den Schichten 3 und 4 wurden Zellen gefunden, die auf Blau reagieren und von Gelb gehemmt werden. Sie erhalten Signale von Ganglionzellen des Typs *Small Bistratified*. Die K-Zellen in den parvozellulären Schichten 3 und 4 dagegen reagieren gegensätzlich auf Rot und Grün. Insgesamt reagieren K-Zellen weniger einheitlich als die Zellen in den M- oder P-Schichten, was die bevorzugte Ortsfrequenz oder den Kontrast angeht. Die verschiedenen K-Schichten sind auch unterschiedlich verschaltet. Die meisten K-Zellen lassen

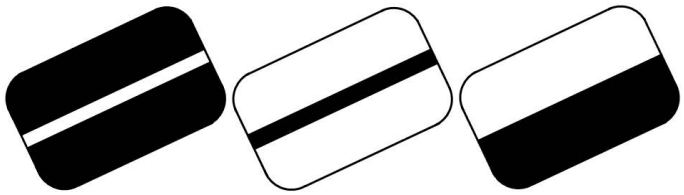


Abbildung 3.6: **Beispiele für Stimuli einfacher Zellen in V1 (Abb. ähnlich [Hub88]).** Einfache Zellen reagieren auf dünne Linien oder Kanten einer bestimmten Orientierung.

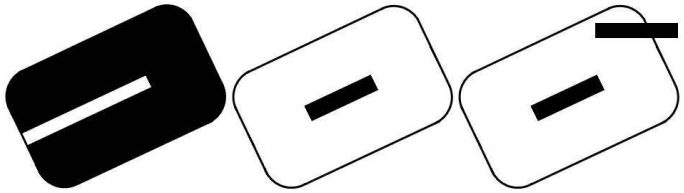


Abbildung 3.7: **Stimuli für Zellen mit End-Stopping (nach [Hub88]).** Eine Zelle, die auf das mittlere Signal reagiert, spricht auch auf das Signal rechts an, da der zusätzliche Balken anders orientiert ist als das Zentrum.

sich neben visuellen Reizen auch durch akustische oder taktile Reize anregen. Die hinteren 2 koniozellularen Schichten erhalten Signale von den sogenannten *P-Giant*-Zellen, die besonders große rezeptive Felder mit einem Durchmesser von 15 Grad besitzen.

Die Zellen aus den M-, P- und K-Schichten senden hauptsächlich Signale an den primären visuellen Kortex V1. Die Zielgebiete innerhalb des V1 unterscheiden sich allerdings je nach Zelltyp. Einige K-Zellen senden auch Signale an die weiter entfernt liegenden Großhirnbereiche V2, V4 und den inferioren Temporallappen (IT) [HR00].

3.1.5 Die primäre Sehrinde

Zellen in der primären Sehrinde V1[2] reagieren auf Kanten und Linien, sowie sich bewegende Kanten und Linien. Hubel [Hub88] unterscheidet *einfache* und *komplexe Zellen*. Einfache Zellen reagieren auf unbewegte Kanten oder schmale Linien einer bestimmten Orientierung (s. Abb. 3.6). Die Durchmesser der rezeptiven Felder variiert von 1/4° für der Fovea zugeordnete Zellen bis 1° am Rand der Netzhaut. Die Dicke der Linien der ersten beiden Zellen aus Abb. 3.6 liegt bei wenigen Bogenminuten und entspricht damit der Größe der Feldzentren der

[2]Gleichbedeutende Bezeichnungen für die primäre Sehrinde: Area Striata, Striate Cortex (engl.), Brodmann Area 17, Primärer Visueller Kortex

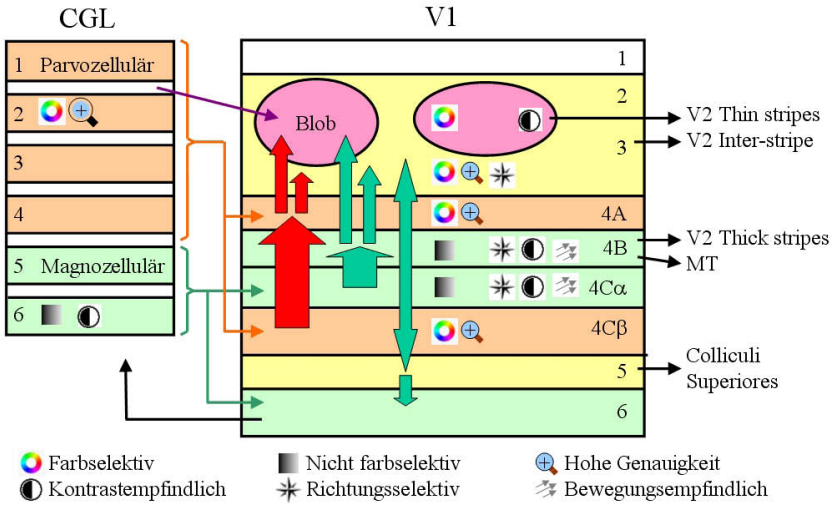


Abbildung 3.8: **Verbindungen zwischen dem Corpus geniculatum laterale (CGL) und dem primären visuellen Kortex.** Verschiedene Schichten des Corpus geniculatum laterale führen zu unterschiedlichen Schichten im V1. Dort bilden sich über mehrere Schichten hinweg Verarbeitungsströme mit unterschiedlichen Funktionen aus, die zu weiteren Regionen des Gehirns führen. Pfeile zeigen die wichtigsten Verbindungen zwischen den Schichten. Die kleinen Symbole geben die Eigenschaften der Zellen in den verschiedenen Schichten an.

Ganglionzellen. Die Stimuli für komplexe Zellen entsprechen im Prinzip denen der einfachen Zellen, allerdings reagieren diese Zellen nur, wenn sich die Kante oder Linie über einen gewissen Bereich in eine bestimmte Richtung bewegt. Im Gegensatz zur Bewegungsrichtung ist die Bewegungsgeschwindigkeit eher unbedeutend. Komplexe Zellen machen geschätzte 75 Prozent der Zellen in V1 aus.

Normalerweise reagieren Zellen umso stärker, je länger die Linie oder Kante in ihrem Zentrum ist. Eine Ausnahme stellen Zellen mit *End-Stopping* [Hub88] dar. Diese Zellen reagieren auf Kanten und Linien einer bestimmter Länge (s. Abb. 3.7), d.h. auf Kanten- und Linienendstücke.

Was die Farbselektivität angeht, lassen sich mehrere Arten von komplexen Zellen unterscheiden. Einige Zellen reagieren unabhängig von der Wellenlänge des Lichts nur auf Helligkeitsunterschiede. Andere Zellen reagieren auf Farbe wie die Zellen im Corpus geniculatum laterale mit einem Zentrum und einem Rand, die für verschiedene Wellenlängen empfindlich sind und auf diese hemmend oder verstärkend reagieren. Eine dritte Zellart wird von Hubel *Double-Opponent* genannt. Diese Zellen kommen als rot-grün- oder blau-gelb-Variante vor. Double-Opponent-Zellen reagieren im Zentrum des rezeptiven Feldes auf eine Farbe hemmend und auf die andere verstärkend. Im Randbereich reagieren die Zellen genau entgegengesetzt: Die Farbe, die im Zentrum hemmend ist, ist am Rand verstärkend, und umgekehrt. Diese Zellen sprechen daher nicht auf homogene Farbflächen an, sondern nur auf die Kanten zwischen zwei verschiedenfarbigen Flächen.

Beim Affen nimmt V1 eine Fläche von etwa 1200 mm^2 bei einer Dicke von 2 mm ein. Die Oberfläche des V1 spiegelt dabei die Topologie der Netzhaut wieder. Zellen auf einer Linie senkrecht zur Oberfläche entsprechen derselben Position auf der Netzhaut. Innerhalb dieser 2 mm lassen sich 6 Schichten unterscheiden, die sich teilweise noch weiter unterteilen lassen. Die Zellen innerhalb einer Schicht sind über eine Strecke von bis zu 2 mm lokal mit Nachbarzellen der gleichen Schicht verbunden. Darüberhinaus existieren Verbindungen, die senkrecht zur Oberfläche verlaufen und die Zellen verschiedener Schichten miteinander verbinden. Die Verbindungen in V1 bilden unterschiedliche Informations- und Verarbeitungsströme, die zu verschiedenen Bereichen des Großhirns führen.

Abbildung 3.8 zeigt die wichtigsten Verbindungen zwischen den Schichten. Die oberste Schicht 1 enthält wenige Zellen und besteht stattdessen hauptsächlich aus Nervenverbindungen. Es wurden Zellen aus den Zwischenschichten 5 und 6, in geringerem Maße auch 1 und 2 des Corpus geniculatum laterale gefunden, die in Schicht 1 senden [HR00].

Zellen in Schicht 4C erhalten Signale aus dem Corpus geniculatum laterale. Sie sind jeweils nur einem Auge zugeordnet. Die Augenzugehörigkeit ist nicht zufällig, sondern bildet ein Muster aus gekrümmten und hin und wieder verzweigten Streifen einer Breite von etwa 0,5 mm entlang der Kortexoberfläche. Innerhalb eines Streifens sind alle Zellen dem selben Auge zugeordnet. Diese Zuordnung wird auch für die anderen Schichten im V1 beibehalten, ist dort aber weniger stark ausgeprägt. Zellen reagieren dann auf Signale von beiden Augen, meistens jedoch unterschiedlich stark.

Die Zellen des V1 sind auch hinsichtlich der Stimulusorientierung regelmäßig angeordnet. Senkrecht zur Kortexoberfläche reagieren alle Zellen auf die gleiche Orientierung. Parallel zur Kortexoberfläche ändert sich die eine Reaktion auslösende Stimulusorientierung über Bereiche von etwa $1\,\mathrm{mm}^2$ kontinuierlich, wobei alle Winkel auftreten. Zwischen diesen Bereichen treten abrupte Änderungen auf. In den Schichten 4C und 4B treten häufig einfache Zellen auf, während in den anderen Schichten komplexe Zellen überwiegen.

Die Schicht 4Cα erhält hauptsächlich Signale aus den magnozellulären Schichten des Corpus geniculatum laterale. Die Zellen in dieser Schicht reagieren daher auf feine Kontrastunterschiede und Bewegungen, sind aber nicht selektiv für Farbe. Der Anteil an orientierungsselektiven Zellen ist im unteren Teil von 4Cα gering, nimmt aber in der Nähe der Schicht 4B stark zu. Von Schicht 4Cα führt der Hauptverarbeitungsweg über die Schicht 4B in die Bereiche V2 und MT des visuellen Kortex. Zellen in Schicht 4B sind ebenso wie die in 4Cα empfindlich für geringe Kontrastunterschiede und Bewegungen, sowie unspezifisch gegenüber Farbe. Sie reagieren sowohl auf die Orientierung von Kanten und Linien als auch auf deren Bewegungsrichtung unterschiedlich. Von 4B führen weitere Verbindungen in die Blob- und Zwischenblobbereiche der Schichten 2 und 3.

Die parvozellulären Schichten des Corpus geniculatum laterale senden Signale an die Schichten 4A und 4Cβ des V1. Diese unterscheiden daher verschiedene Farben und besitzen aufgrund der kleinen rezeptiven Felder eine hohe Detailschärfe. Von 4Cβ führt ein Verarbeitungsweg über 4A in die Blob- und Zwischenblobbereiche der Schichten 2 und 3.

Die Schichten 2 und 3 unterteilen sich in *Blobs* und die Bereiche zwischen den Blobs. Die Blobs bilden ein regelmäßiges Punktmuster entlang der Kortexoberfläche und haben einen Durchmesser von etwa 1 mm.

Zellen in den Bereichen zwischen den Blobs sind orientierungsselektiv und reagieren auf Helligkeits- und Farbkanten. Allerdings bevorzugen die meisten Zellen keine bestimmte Farbe und reagieren auf Kanten zwischen beliebigen Farbflächen. Sie erkennen relativ feine Strukturen. Diese Zellen sind also hauptsächlich empfindlich für Formmerkmale, nicht aber für Farbe.

Blob-Zellen dagegen sind überwiegend farbselektiv ohne eine bestimmte Stimulusorientierung zu bevorzugen. Sie empfangen Signale aus den Zwischenschichten des Corpus geniculatum laterale und aus der Schicht 4C. Ihre rezeptiven Felder sind um ein mehrfaches Größer als die der Zellen aus dem Corpus geniculatum laterale. Etwa 50 Prozent der Blob-Zellen sind Double-Opponent, d.h. sie reagieren nur auf Kanten zwischen Flächen bestimmter Farbe. Sie sind gemischt mit kontrastempfindlichen Zellen ohne klare Farbpräferenz und farbselektiven Zellen ähnlich denen im Corpus geniculatum laterale. Da unterschiedliche Blobs von Zellpopulationen in verschiedenen Schichten des Corpus geniculatum laterale angeregt werden, sind sie bezüglich ihrer jeweiligen Farbpräferenz einheitlich [HR00].

Die meisten Zellen in den Schichten 2 und 3 reagieren umso stärker, je länger eine Linie im rezeptiven Feld ist. Etwa 20 Prozent der Zellen reagieren jedoch nur auf Linienstücke und Kanten einer bestimmten Länge (End-Stopping). Die

Schichten 2 und 3 sind mit dem Großhirnbereich V2 verbunden, allerdings verzweigen die Blob-Zellen und die Zellen zwischen den Blobs zu verschiedenen Unterstrukturen. Außerdem gibt es Verbindungen in die richtungsselektive Schicht 5, die Signale in tiefere Hirnregionen sowie in Schicht 6 überträgt.

Die Zellen in Schicht 5 reagieren auf kleine Linienstücke genauso gut wie auf große. Ihre rezeptiven Felder sind viel größer als die der Zellen in den Schichten 2 und 3.

Neben den Signalen aus Schicht 5 erhält Schicht 6 auch Signale aus den magnozellulären Schichten des Corpus geniculatum laterale und schickt selbst auch wieder Signale in das Corpus geniculatum laterale zurück. In Schicht 6 befinden sich Zellen mit großen, sehr langen und schmalen rezeptiven Feldern, die wieder umso stärker reagieren, je länger die Linie im rezeptiven Feld ist [Bau98, Hub88].

Der primäre visuelle Kortex stellt die erste Stufe dar, auf der Stereoinformationen ausgewertet werden. Dies geschieht durch den Vergleich der Stimuluspositionen in den dem linken und dem rechten Auge zugeordneten rezeptiven Feldern von Zellen, die von beiden Augen Signale bekommen. Cumming [CD01] gibt in einem Review Modelle für einfache und komplexe Zellen an. Dazu werden die rezeptiven Felder der einfachen Zellen als mit einer Gaußkurve überlagerten Sinuskurve modelliert. Die Aktivierung einer einfachen Zelle ließe sich gut durch eine Addition von sinusförmigen Einzelaktivierungen durch das linke und das rechte Auge und ein anschließendes Verwerfen negativer Werte annähern. Die Selektivität für bestimmte Disparitäten ergibt sich aus dem Phasenunterschied der Sinuskurven. Im Gegensatz zu einfachen Zellen sprechen komplexe Zellen nicht nur auf Stimuli an einer bestimmten Position des rezeptiven Feldes an, sondern ortsunabhängig auf alle Stimuli der bevorzugten Ortsfrequenz. Dabei sind komplexe Zellen selektiv für die Phasenverschiebung oder Disparität zwischen den Augen. Das Verhalten komplexer Zellen wird durch ein *Energiemodell* beschrieben, das die Signale mehrerer einfacher Zellen quadriert und aufsummiert. Dabei sind die Einzelzellen für die gleiche Disparität aber unterschiedliche Positionen im rezeptiven Feld selektiv. Im Endeffekt beschreibt das Modell eine Art Korrelation zwischen den Stimuli des linken und des rechten Auges.

Mehrere Tatsachen sprechen für weitere, höhere Hierarchiestufen der Stereoverarbeitung. Zum einen ergeben sich aus den relativen Disparitäten an verschiedenen Bereichen des Sichtfeldes stabilere Tiefeneindrücke als aus den absoluten Dispariäten, für die Zellen in V1 empfindlich sind. Zum anderen sollte man nach dem Energiemodell für antikorrelierte Stimuli (z.B. in Form von antikorrelierten Random-Dot-Stereogrammen) einen vollständig umgekehrten Tiefeneindruck erwarten. Tatsächlich wird aber überhaupt keine Tiefe wahrgenommen. Darüberhinaus signalisieren Zellen in V1 auch falsche Stereokorrespondenzen, die nicht bewußt wahrgenommen werden [CD01].

Die von V1 ausgehenden Verbindungen zielen hauptsächlich auf das Areal V2, wenige Verbindungen führen auch in das Areal MT (mittlerer temporaler Gyrus).

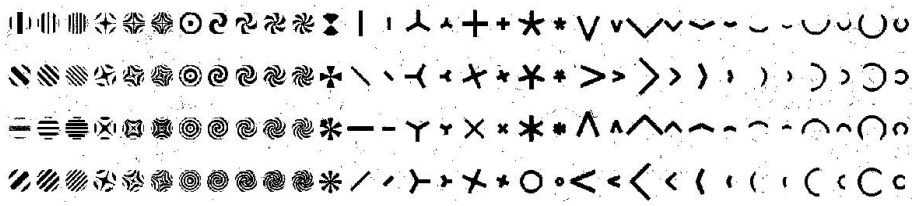


Abbildung 3.9: **Formmerkmale von Hegde und Van Essen [HE00].** Verschiedene Zellen im V2 reagieren auf verschiedene Untergruppen der dargestellten Muster.

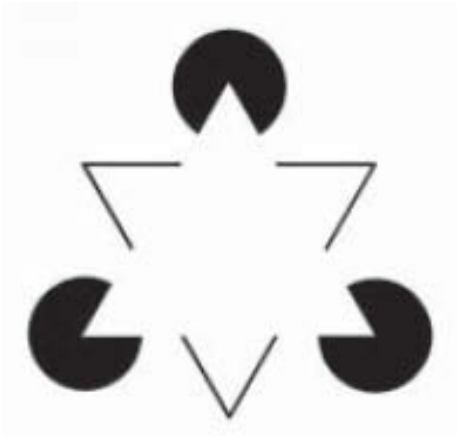


Abbildung 3.10: **Kanizsa-Dreieck mit Scheinkonturen.**

3.1.6 Areal V2[3]

Wird der Bereich V2 mit Zytochromoxidase markiert, ergibt sich ein Muster aus ca. 10–12 dicken und dünnen Streifen, die durch dünne Zwischenstreifen getrennt sind [OE97, RT95]. Die Streifen bilden bezüglich ihrer rezeptiven Felder Kreise um die Fovea. Die verschiedenen Streifenarten erhalten Signale aus unterschiedlichen Bereichen des V1 und bilden offenbar getrennte Funktionseinheiten. Wie Shipp und Zeki nachwiesen [SZ95] verläuft die Informationsverarbeitung auch nach V2 auf getrennten Wegen, da verschiedene Streifenarten zu unterschiedlichen und nicht miteinander verbundenen Gehirnarealen (V4 und MT) führen.

Die rezeptiven Felder der Zellen innerhalb eines Streifenzyklus überlappen sich stark. Eine Ausnahme bilden die Zwischenstreifen, die auch *blasse Streifen* (engl. *Pale Stripes*) genannt werden. Die rezeptive Felder der beiden Zwischenstreifen eines Zyklus überlappen sich zwar mit denen der dicken und dünnen

[3]Oder auch Areal 18 nach Brodmann

Streifen, nicht jedoch untereinander. Die rezeptiven Felder von Zellen verschiedener Zyklen überlappen sich generell nicht [RT95].

Nach Row und Ts'o [RT95] enthalten die dicken Streifen zu 76% disparitätsselektive Zellen, die oft auf sehr lange, schmale und oft vertikale Linienstücke reagieren. Sie sind meistens nur beidäugig stimulierbar. Die disparitätsselektiven Zellen treten in Gruppen von entweder nur farbselektiven oder nur nicht farbselektiven Zellen auf. Einen Anteil von 14% machen Gruppen von nicht farbselektiven, nicht orientierungsselektiven Zellen aus. Die rezeptiven Felder benachbarter Gruppen überlappen sich oft stark, was auf eine weitere funktionale Unterteilung innerhalb der Streifen hindeutet.

Die Zwischenstreifen bestehen zu 74% aus orientierungsselektiven und nicht farbselektiven Zellen mit kleinen rezeptiven Feldern. Bei vielen Zellen tritt *End Stopping* auf.

Dünne Streifen enthalten überwiegend farbselektive Zellen. Sie fanden zu 50% unorientierte, farbselektive Zellen mit größeren rezeptiven Feldern. Orientierungsselektive, farbselektive Zellen mit kleineren rezeptiven Feldern machen 21% aus. Seltener treten farbselektive *Spot Cells* auf, die auf kleine Punkte in großen rezeptiven Feldern (2-3 Grad) reagieren.

Nach Xiao et al. [XWF03] verlaufen die Farben, für die die Zellen der dünnen Streifen selektiv sind, abschnittsweise kontinuierlich. Der Farbverlauf dieser Bereiche ähnelt offenbar der CIE-Farbtafel. Die Bereiche sind rund bis streifenförmig mit einem Durchmesser von etwa 0,5 bis 1 mm.

Die dicken Streifen erhalten Signale aus der Schicht 4B des primären visuellen Kortex und senden Signale weiter an den Bereich MT. Die Zwischenschichten erhalten Signale aus dem Zwischen-Blob-Bereich der V1-Schichten 2 und 3. Sie senden weiter an V4 und möglicherweise an V3. Die dünnen Streifen erhalten Signale aus den farbselektiven Blobs der Schichten 2 und 3 des V1. Sie senden weiter an V4 [LH87].

Die Untersuchungen von Hegde und Van Essen [HE03] zeigen, daß die effektiven Stimuli von V2-Zellen möglicherweise komplexer sind als bisher angenommen. Sie testen V2-Zellen nicht nur auf Linienstücke, sondern auch auf Gittermuster, Linienkreuzungen und Bögen (s. Abb. 3.9), für die bereits selektive Zellen in V4 gefunden wurden. Laut Hegde und Van Essen sprechen die meisten V2-Zellen stärker auf die komplexeren Gittermuster (die linken 12 Spalten von Abb. 3.9) als auf Konturmerkmale an. Auch rufen größere Konturmerkmale stärkere Reaktionen hervor als kleinere. Für Konturmerkmale ergibt sich eine stärkere Selektivität. Per Hauptkomponentenanalyse stellen sie fest, daß die ersten zwei Hauptkomponenten 69% und die ersten acht Hauptkomponenten 90% der Ergebnisse ausmachen. Sie folgern, daß der V2 offenbar einen niedrigdimensionalen Merkmalsraum kodiert. Eine Clusterung der Stimuli nach der Ähnlichkeit des Ansprechverhaltens der Zellen ergibt eine klar erkennbare Gruppe mit komplexen Gittermustern, eine Gruppe mit großen Bögen und Winkeln aus langen Linienstücken und eine Gruppe mit den restlichen Stimuli. Teilweise sind vage Untergruppen (große Rundbögen, kleine Sterne) zu erkennen.

Etwa 30–40 Prozent der Zellen im V2 reagieren auf Illusionskanten wie im Kanizsa-Dreieck (Abb. 3.10)[vdHPB84, Nie02]. Im V1 dagegen reagieren umstrittenerweise nur 2% der Zellen auf diesen Effekt, was möglicherweise auf Verbindungen beruht, die aus dem V2 zurückführen. Scheinkanten erkennen nicht nur Menschen, sondern auch Affen, Eulen, Bienen und Libellen, was auf deutliche evolutionäre Vorteile durch diese Fähigkeit hindeutet. Die Tatsache, daß Scheinkonturen bereits wenige synaptische Verbindungen nach der Reizaufnahme erkannt werden, spricht gegen ältere kognitive Erklärungsansätze, welche die Auflösung mehrdeutiger Wahrnehmungen durch den Vergleich mit erlernten Objektprototypen erklären.

Als nächste Station im ventralen Verarbeitungsstrom wird das Areal V4 dargestellt. Da das Areal V3 im allgemeinen dem dorsalen Verarbeitungsstrom zugerechnet wird, wird es hier übersprungen.

3.1.7 Areal V4

Der Bereich V4 enthält Teile, deren Eingaben aus V2-Zwischenstreifen stammen, sowie Teile, deren Eingangssignale sowohl aus den V2-Zwischenstreifen als auch aus den dünnen Streifen des V2 stammen [SZ95]. Die genaue Funktion dieses Bereichs sowie die Entsprechung im menschlichen Gehirn ist einigermaßen unklar [TH01]. Erste Vermutungen, daß V4 nur der Farberkennung dient, haben sich nicht bestätigt [SD90]: Zum einen sind viele V4-Zellen nicht nur farb-, sondern auch formselektiv, zum anderen liegt V4 mitten in der ventralen Verarbeitungskette, was die Objekterkennung im inferioren Temporallappen weitgehend auf die reine Farbinformationen beschränken würde.

Schein und Desimone [SD90] messen, daß viele der 332 untersuchten Zellen im Bereich der zentralen 5° des Sichtfeldes für mehrere Farben empfindlich sind. 21% der Zellen reagieren schmalbandig auf Licht einer bestimmten Wellenlänge. Die Farbpräferenzen der untersuchten Zellen decken dabei das komplette Farbspektrum ab. 28% der Zellen reagieren gleichermaßen auf zwei verschiedene schmale Frequenzbänder. Auf drei verschiedene Farben reagieren jedoch nur 1% der Zellen. Die meisten farbselektiven Zellen reagieren in schwächerer Form auch auf weißes Licht. Im Gegensatz zu den Zellen der vorhergehenden Verarbeitungsstufen trat bei den untersuchten Zellen keine hemmende Wirkung bestimmter Farben auf. Von den getesteten Zellen sind viele orientierungsselektiv. Die orientierungsselektiven und die farbselektiven Zellen bilden dabei keine separaten Gruppen; eine Zelle kann sowohl farb- als auch orientierungsselektiv sein. Eine Unterscheidung zwischen Farb- und Formmerkmalen findet also offenbar nicht statt.

Ghose und Ts'O [GT97] finden Hinweise auf eine mögliche funktionale Unterteilung in V4, da der foveale Bereich bis zu einer Winkelabweichung (Exzentrizität) von 3° von der optischen Achse anscheinend einen Sonderstatus innehat. Zum einen finden sie nur dort eine systematische Zellanornung gemäß ihrer Orientierungsselektivität, zum anderen finden sich auch Unterschiede, was die Verbindungen zu anderen Gehirnarealen angeht. Der foveale Bereich erhält offenbar als einziger Signale direkt aus V1. Außerdem sendet der Bereich unter

20° Exzentrizität ausschließlich Signale in den inferioren Temporallappen. Die strenge retinotopische Zellanordnung, wie sie in den früheren Stadien des Sehsystems vorliegt, läßt dagegen nach, da kleine visuelle Stimuli Erregungen über weite Bereiche in V4 hervorrufen können. Im Bereich von 5-7° Exzentrizität treten Bereiche mit einheitlicher Farbpräferenz auf.

Die Größe der rezeptiven Felder liegt im fovealen Bereich bei ca. 0,66°, was einigen Tausend Ganglionzellen entspricht. Obwohl nur Stimuli in diesen Bereichen V4-Zellen anregen, können auch weit entfernte Stimuli den Zustand einer Zelle beeinflussen. Schein und Desimone [DS87, SD90] finden heraus, daß Stimuli innerhalb eines sehr großen (bis 16°) Bereichs um das eigentliche rezeptive Feld die Erregung einer Zelle unterdrücken können. Die Umgebung reagiert dabei selektiv auf die gleiche oder eine ähnliche Farbe wie das eigentliche rezeptive Feld. Aufgrund der Größe der unterdrückenden Umgebung vermuten sie, daß V4 einerseits eine Rolle bei der Unterscheidung zwischen Vorder- und Hintergrund spielt, und andererseits dafür verantwortlich ist, daß die Farbe eines Objekts im Wesentlichen relativ zu umgebenden Objekten wahrgenommen wird und damit unabhängig von der tatsächlichen Beleuchtungsfarbe ist. Die zweite These wird durch Versuche gestützt, die darauf beruhen, daß in der Mitte des Sichtfeldes die unterdrückende Umgebung Signale aus beiden Gehirnhälften erhalten muß. Bei Menschen, bei denen die Nervenverbindung zwischen den beiden Gehirnhälften, das Corpus Callosum, durchtrennt ist, findet daher kein Vergleich zwischen einem rezeptiven Feld aus der einen Hälfte des Gesichtsfeldes und dem Teil der unterdrückenden Umgebung aus der anderen Hälfte statt. Land [Lan83] zeigte, daß bei bei Menschen mit einer solchen Durchtrennung tatsächlich die Farbkonstanz bei unterschiedlicher Beleuchtung stark beeinträchtigt ist. Schein und Desimone räumen ein, daß auch andere Gehirnareale für diesen Effekt verantwortlich sein können. Ghose und Ts'O [GT97] geben an, daß Zellen mit unterdrückender Umgebung für 5–7° Exzentrizität auftreten und sich in kleinen Zellbereichen ballen.

Hegde und van Essen [HE07] testen V4-Neuronen auf die in Abb. 3.9 gezeigten Stimuli und finden graduelle Unterschiede im Vergleich zur Reaktion von Neuronen in V2, beispielsweise eine stärkere Reaktion auf Gittermuster in V4. Qualitative Unterschiede, die aus der Hierarchie verschiedener Kortexregionen resultiert, finden sie jedoch nicht. Sie schließen, daß die Hauptunterschiede zwischen den Regionen eher in der Behandlung von Disparität, Größenskalierung, Scheinkonturen oder beim Einfluß von Aufmerksamkeit liegt.

Die Erkennung von Texturkanten und Illusionskanten untersuchen beispielsweise Kastner et al. [KWU00]. Sie vergleichen die Reaktion von Zellen auf Bilder mit Strichtexturen sowohl mit als auch ohne texturbasierte Kanten. Texturierte Bilder rufen demnach in den Bereichen V1, V2/VP, V4, TEO und V3A Reaktionen hervor. Auf Unterschiede zwischen Bildern mit Texturkanten und ohne reagieren jedoch nur die höheren Bereiche V4 und TEO, was die Autoren vor allem auf die größeren rezeptiven Felder zurückführen.

Abbildung 3.11: **Mooney-Bilder zum Test der Erkennung von Gesichtern (aus [Sch05]).**

3.1.8 Der inferiore Temporallappen

Abbildung 3.1 zeigt eine Unterteilung des inferioren Temporallappens (IT) in einen anterioren (AIT, Alternativbezeichnung: TE), einen posterioren (PIT, Alternativbezeichnung: TEO) und den zentralen (CIT) Teil. Die Unterteilung beruht auf Unterschieden in den Verbindungen zu anderen Gehirnregionen. In der visuellen Verarbeitungskette werden die Bereiche V4–PIT–CIT–AIT sequentiell durchlaufen. Am Ende dieser Kette stehen Zellen, die relativ unabhängig von der Perspektive auf bestimmte Objekte oder Teile von Objekten reagieren [Tan96, TT01, PDM$^+$05, KBV03]. In einer Literaturdurchsicht schließt Miyashita [Miy93], daß diese im IT in Form von Prototypen gespeichert sind. Dafür spricht, daß Läsionen im IT die Größentoleranz und die Unterscheidung von Stimuli in verschiedenen Größen, Orientierungen und Schattierungen beeinträchtigen.

Miyashita [Miy93] und Lee [Lee00] weisen auch darauf hin, daß der IT nicht nur für die visuelle Erkennung von Objekten zuständig ist, sondern auch mit dem Gedächtnis zu tun hat. Beim Menschen speichert der linkte Temporallappen vor allem sprachliche Erinnerungen, der rechte dagegen visuelle. Eine Entfernung des rechten anterioren Temporallappens (zur Epilepsiebehandlung) führt beim Menschen zu einer leichten Beeinträchtigung der Objekterkennung, wenn nur wenige visuelle Schlüssel vorliegen wie im Mooney-Closure-Face-Test (Abb. 3.11), sowie einer starken Behinderung bei der Erkennung erlernter, sprachlich schwer beschreibbarer Bilder. Sakai, Miyashita und Naya [SM91, MSHM91, NSM96] zeigen anhand von Versuchen an Affen, daß einige Zellen im AIT nicht nur auf gerade betrachtete Bilder, sondern auch auf erinnerte Bilder reagieren. Dazu trainieren sie Affen auf das sequentielle Erkennen von Paaren von zufälligen Fraktalmustern (s. Abb. 3.12). Sie finden Zellen, die selektiv auf beide Bilder eines Paares reagieren, obwohl diese keine geometrischen Gemeinsamkeiten besitzen. An der Reaktion auf die das Expe-

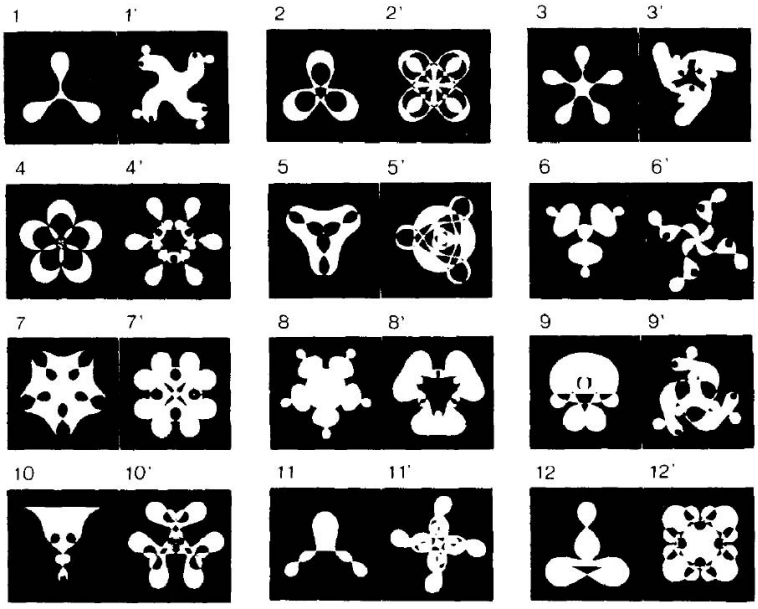


Abbildung 3.12: **Fraktalpaare von Miyashita et al. [NSM96, Miy93].** Zellen des IT sind anscheinend in der Lage, diese Muster zu erlernen.

riment begleitenden visuellen Signale können sie erkennen, daß tatsächlich ein Gedächtnisabruf stattgefunden haben muß. Miyashita [Miy93] schlägt vor, daß die Fähigkeit zur Assoziation zeitlich zusammenhängender aber geometrisch unterschiedlicher Muster eingesetzt wird, um Objektdarstellungen zu erlernen, die unabhängig von der Betrachtungsperspektive sind. Möglicherweise besteht hier auch ein Unterschied zwischen Affen und Menschen. Nielsen et al. [NLR08] zeigen, daß Menschen zur Erkennung von gedrehten Objekten rotationsinvariante Merkmale nutzen, wohingegen Affen für jede Rotation einen separaten Satz Merkmale benutzen. Die Ergebnisse von Tarr und Pinker [TP89] deuten an, daß sowohl eine mentale Rotation neuer Bilder auf bekannte Objekte in einer bestimmten Lage stattfindet, als auch neue Objektorientierungen durch ausreichendes Training erlernt werden können, wodurch eine mentale Rotation entfällt. Diese Ergebnisse beziehen sich allerdings nicht zwangsläufig auf den IT.

Der in den anderen Bereichen bereits aufgetretene Effekt, daß die rezeptiven Felder entlang der Verarbeitungskette immer größer werden, setzt sich im IT fort. Die rezeptiven Felder von IT-Zellen haben einen Durchmesser von 50° bis 70° [TRS01, AR05] und schließen die Fovea mit ein. Es stellt sich daher die Frage, wie das Gehirn bei solch großen rezeptvien Feldern mit Szenen umgeht, die sich aus mehreren kleinen Objekten zusammensetzen oder in denen das gleiche Objekt mehrfach auftritt. Aggelopoulos und Rolls [AR05] zeigen, daß die rezeptiven Felder nur dann so groß sind, wenn dem Versuchstier ein Objekt vor

einem leeren Hintergrund gezeigt wird. Dies stellt die vorherrschende experimentelle Anordnung dar. Bei der Betrachtung natürlicher Szenen schrumpfen die rezeptiven Felder jedoch auf Bereiche mit einem Durchmesser von etwa 12° in oder nahe der Fovea. Dabei sprechen die Neuronen am stärksten auf Objekte im Bereich der Fovea an. Positionen ca. 10° außerhalb des Zentrums der Fovea bewirken ebenfalls Reaktionen, aber deutlich schwächere. Das Auftreten mehrerer Objekte bewirkt zudem Asymmetrien in den rezeptiven Feldern. Neuronen im IT geben in natürlichen Szenen also Aufschluß darüber, welches Objekt genau im fovealen Bereich liegt, und welche Objekte in einem kleinen Bereich um die Fovea liegen, letzteres jedoch ohne das zentrale Objekt zu überdecken. Die Asymmetrien in den rezeptiven Feldern liefern darüberhinaus noch gewisse geometrische Informationen.

Suzuki et al. [SMT06] gehen der Frage nach, ob die Reaktion von Zellen im IT möglicherweise von einer gerade ausgeführten Aufgabe abhängt. Insbesondere sind Menschen in der Lage, Objekte je nach Notwendigkeit entweder nur grob einzuordnen oder aber individuell zu identifizieren. Da von Zellen im V4 bekannt ist, daß die Stärke der Reaktion von der Aufmerksamkeit auf den Ort eines Reizes abhängt, prüfen sie, ob etwas ähnliches im IT geschieht. Sie finden jedoch heraus, daß die IT-Zellen immer gleich selektiv auf die Stimuli reagieren, unabhängig davon, wie genau die Versuchstiere die gezeigten Objekte einordnen sollen.

Janssen et al. [JVO00] untersuchen anhand von Random-Dot-Stereogrammen die Reaktion von AIT-Zellen auf dreidimensionale Stimuli. Sie finden Zellen, die für senkrechte Flächen (0. Ordnung), geneigte Flächen (1. Ordnung) und gekrümmte Flächen (2. Ordnung) selektiv sind. Anhand ihrer Selektivität lassen sich die Zellen jedoch nicht in separate Klassen einteilen, sondern bilden anscheinend ein Kontinuum. Die Selektivität für bestimmte Muster nimmt mit steigender Ordnung zu. Bezüglich des Abstands des Betrachters vom Stimulus, der Position senkrecht zur optischen Achse und der Objektgröße besteht eine gewisse Toleranz. Im Vergleich zu reinen 2D-Objekten ist diese jedoch niedriger. Die Positionstoleranz beträgt nur wenige Grad. Die Größentoleranz liegt bei etwa einem Faktor zwei, wohingegen Janssen et al. andere Autoren zitieren, die einen Faktor von bis zu 32 für 2D-Objekte nennen. Als Ursache für diese Unterschiede wird vermutet, daß zur Bestimmung der Disparitäten eine gewisse Minimalgröße der rezeptiven Felder erforderlich ist.

Tanaka et al. [Tan96, Tan00, Tan93, AG90, TT01] versuchen, genaue 2D-Merkmale zu ermitteln, die Zellreaktionen im AIT hervorrufen. Ein erster dazu ins Auge gefaßte Ansatz stammt von Gross et al. und beruht auf der Konstruktion von Stimuli aus Fourierdeskriptoren für eine Zelle, die auf symmetrische, sternförmige Muster reagiert [AG90]. Der konstruktive Ansatz scheitert aber daran, daß die Zelle nicht linear auf die zusammengesetzten Stimuli reagiert. Offenbar funktioniert dieser Ansatz nicht, wenn die Basisfunktionen für die Stimuli nicht genau passen, bzw. unbekannt sind. Tanaka arbeitet stattdessen mit einem alternativen reduktiven Ansatz von Gross. Zunächst werden für AIT-Zellen anhand einer großen Bildersammlung effektive Stimuli ermittelt. Anschließend werden die effektivsten Stimuli einer Zelle unter Überwachung der Zellreaktion

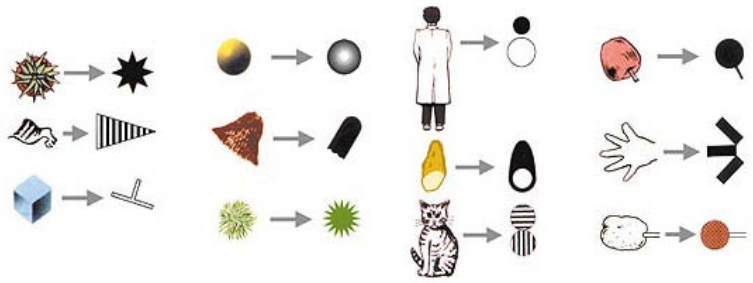


Abbildung 3.13: **Vereinfachung komplexer stimuli auf ihre kritischen Merkmale (Abb. ähnlich [Tan00])** Die Bildpaare zeigen jeweils links das komplexe Merkmal und rechts die Vereinfachung.

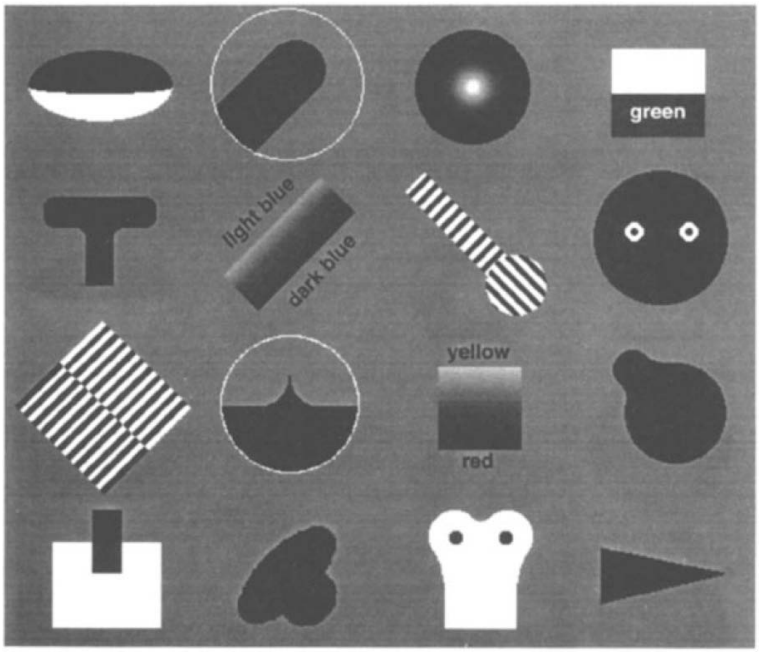


Abbildung 3.14: **Beispiele moderat komplexer Merkmale [Tan96]**

schrittweise vereinfacht, bis jede weitere Reduktion des Stimulus die Zellreaktion stoppt. Abbildung 3.13 zeigt einige komplexe Stimuli und die jeweilige vereinfachte Version. Tanaka beschreibt die resultierenden Stimuli als 'moderat komplexe' Merkmale. Meistens handelt es sich um mäßig komplexe Konturbilder oder Figuren mit einer bestimmten Farbe oder Textur. Die Abbildung 3.14 zeigt einige Beispiele. AIT-Zellen sind meistens selektiv für die Polarität des Kontrasts der Konturbilder und die Art des Füllmusters. Die Merkmale sind offenbar nicht angeboren, sondern können auch durch intensives Training erlernt werden.

Zellen, die auf ähnliche Stimuli reagieren, sind systematisch in "Säulen" angeordnet [Tan93, Tan96]: Senkrecht zur Kortexoberfläche sprechen praktisch alle Zellen auf dasselbe oder sehr ähnliche Merkmale an, beispielsweise Sterne mit 4 und 8 Spitzen. Entlang der Kortexoberfläche reagieren Bereiche mit einem mittleren Durchmesser von 400 μm auf gleiche oder ähnliche Stimuli. Eine Säule enthält somit etwa 10 000 Neuronen, der gesamte IT etwa 2000 Säulen. Gruppen benachbarter Säulen auf einer Fläche von etwa 1 mm^2 bilden bestimmte Merkmalsvariationen kontinuierlich ab, beispielsweise verschiedene Rotationen. Säulen unterschiedlicher,aber benachbarter Gruppen reagieren auf völlig unterschiedliche Stimuli. Tanaka [Tan93] sieht die Bedeutung der Säulen entweder in der Bildung eines visuellen Alphabets zur Bildbeschreibung, das von kleinen Merkmalsvariationen abstrahiert, oder entgegengesetzt in der Betonung der feinen Unterschiede ähnlicher Objekte, beispielsweise zur Unterscheidung von Sternen mit unterschiedlich vielen Spitzen. Andererseits sprechen auf einen bestimmten Stimulus häufig unterschiedlich selektive Zellen aus mehreren Säulen an. Tanaka [Tan00] deutet an, daß die Reaktionen größerer Zellbereiche auf Objektvariationen möglicherweise präzisere Informationen liefert als das Ansprechen einzelner Zellen.

Diesen Gedanken führen Rolls et al. [RTT97, FRAJ07] fort. Sie überprüfen, ob Objekte im IT lokal in Zellen gespeichert sind (1 Zelle = 1 Objekt) oder global (n Zellen = 2^n Objekte, sofern kein Rauschen vorliegt), bzw. inwiefern eine *spärliche* Speicherform gewählt wird, die zwischen den genannten Extremen liegt. Die Vorstellung einer lokalen Kodierung ist die Konsequenz aus dem hierarchischen Aufbau des visuellen Kortex. Zweifel an dieser Idee gab es allerdings schon früh [Hub88]. Bei einer globalen Kodierung kann jede Kombination von Erregungszuständen der Zellen in der Population ein separates Objekt bezeichnen. In dem Fall wächst die Anzahl speicherbarer Objekte exponentiell. Bei spärlichen Kodierungen (engl. *sparse neural codes*) ist dagegen zu jedem Stimulus immer nur eine kleine Gruppe von Neuronen der Gesamtpopulation aktiv. Um die Frage zu lösen, bestimmen Rolls et al. [RTT97] für 14 gesichtserkennende IT-Zellen die gespeicherte Information, indem sie die Reaktion auf 20 verschiedene Stimuli auswerten. Sie stellen fest, daß die in den Zellreaktionen der Population ausgedrückte Information annähernd linear von 0,33 Bits für eine Zelle auf 2,77 Bits für 14 Zellen steigt, was einem exponentiellen Anstieg der unterscheidbaren Stimuli entspricht. Dies spricht gegen eine lokale Kodierung. Die Zellreaktionen auf einen Satz von Stimuli beschreiben Rolls et al. genauer anhand der *Spärlichkeit* (engl. *sparseness*). Für Populationen ist die Spärlichkeit

ein Maß für die Anzahl der im Mittel aktivierten Zellen. Für einzelne Zellen gibt die Spärlichkeit die Selektivität bezüglich der Stimuli dar. Für die gesichtserkennende Zellpopulation ermitteln Rolls et al. einen hohen Wert, der auf eine eher breite Verteilung der Aktivierungen hindeutet [FRAJ07]. Zudem ergibt sich für die Zellpopulation die gleiche Spärlichkeit wie für einzelne Zellen, sodaß die Autoren das Zellverhalten als schwach ergodisch bezeichnen. Rolls et al. schließen, daß die Zellen innerhalb der Population unkorreliert sind.

Den Vorteil spärlicher Kodierungen im Gegensatz zu kompakten sah man anfangs in einem geringeren Stoffwechsel aufgrund der im Mittel geringeren Anzahl aktiver Neuronen. Da jedoch auch inaktive Zellen Energie verbrauchen, gilt diese These als widerlegt. Willmore und Tolhurst [WT01] sehen den Vorteil spärlicher Kodierungen darin, daß sie fehlertoleranter sind, da ähnliche Stimuli im Gegensatz zu dem Fall einer kompakten Kodierung nicht zu völlig unterschiedlichen Aktivierungsmustern führen müssen.

3.2 Erscheinungsbasierte Objekterkennung im Rechner

Die Erkennung von Objekten mit Hilfe von Modellen der dreidimensionalen Objektstruktur wie von Marr [Mar82] vorgeschlagen beginnt mit der Extraktion primitiver Strukturelemente. Dies sind in der Regel Geraden, Ecken, Flächen oder Bögen. Benachbarte Strukturelemente werden anschließend analysiert und gruppiert, bis die dreidimensionale Struktur ermittelt worden ist. Üblicherweise sind solche Modelle beleuchtungsunabhängig und können auch die geometrischen Beziehungen zwischen den Teilen flexibel speichern. Als Vorteil solcher Modelle für die Objekterkennung geben Burge und Burger [BB97] daher eine hohe Toleranz gegen Beleuchtungsschwankungen, Objektdeformationen und teilweise Verdeckung an.

Die Qualität der Objekterkennung mit Strukturmodellen hängt entscheidend davon ab, wie zuverlässig die primitiven Elemente in unbekannten Bildern wiedergefunden werden können. Aufgrund der Anfälligkeit dieses Schritts gegen Bildrauschen und andere störende Einflüsse hat sich dieser Punkt jedoch immer wieder als problematisch herausgestellt [NMN96, Low99, Sel01, BBM96, SK04]. Burge und Burger [BB97] merken darüberhinaus an, daß für eine Gruppierung verstreuter Einzelteile oft nur schwache lokale Hinweise auf die globale Objektstruktur zu finden sind, sodaß diese oft verfehlt wird. Die Methode gilt daher nur für stark kontrollierbare Anwendungsbereiche als praktikabel. Ponce et al. [PLRS04] verweisen insbesondere auf die Modellierung starrer Objekte.

Die erscheinungsbasierte Objekterkennung ist der Versuch, die Probleme des strukturellen Ansatzes zu umgehen. Man verzichtet zumeist darauf, die dreidimensionale Struktur eines Objekts zu modellieren, und speichert stattdessen, wie das Objekt aus verschiedenen Perspektiven aussieht. Schwankungen in der Erscheinung aufgrund der Oberflächeneigenschaften der Objekte in Verbindung mit Beleuchtungsänderungen werden als Bestandteil des Modells gesehen. Er-

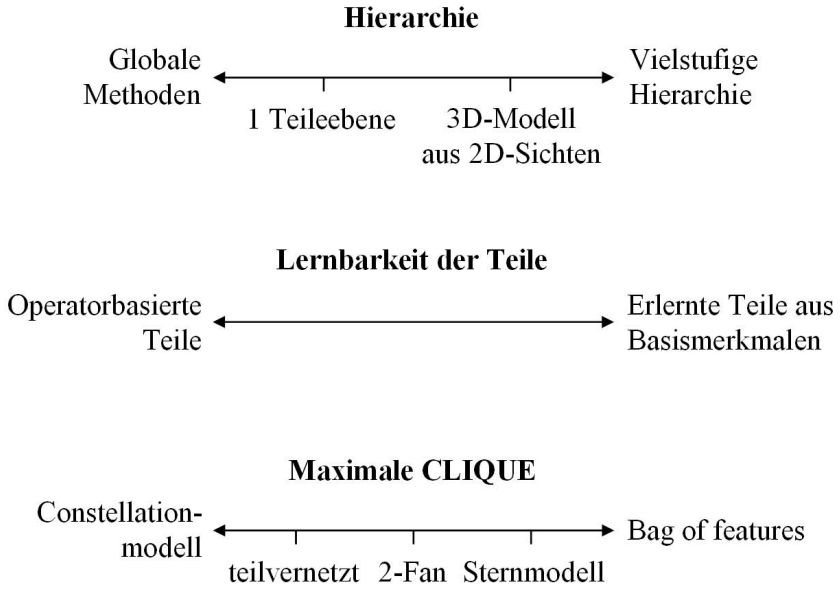


Abbildung 3.15: **Unterschiedliche Auslegungen erscheinungsbasierter Modelle:** Objekte können als Teilehierarchie modelliert werden oder als Ganzes erkannt werden. Oft werden die Teile eines Objekts durch Regionendetektoren ermittelt, sie können sich jedoch auch aus wiederum einfacheren Merkmalen zusammensetzen. Die geometrischen Abhängigkeiten zwischen Teilen können wie beim Constellation-Modell sehr vernetzt modelliert werden oder wie bei dem Bag-of-features-Ansatz vernachlässigt werden.

scheinungsbasierte Ansätze sind auch nicht auf die fehleranfällige Extraktion geometrischer Merkmale angewiesen. Zur Erkennung genügt eine Zuordnung eines unbekannten Objekts zu einer bekannten Erscheinung. Derzeit werden verschiedene Modelle, Trainings- und Erkennungsverfahren diskutiert. Die folgenden Abschnitte geben einen Überblick über aktuelle Fragestellungen und vielversprechende Lösungsansätze.

3.2.1 Objektmodelle

Das Objektmodell hat den Zweck, die typischen Eigenschaften der zu erkennenden Objekte zu speichern (s. Abb.1.1). Zu den derzeit meist diskutierten Entwurfsentscheidungen (s. Abb. 3.15) gehört zunächst die Hierarchie der Objektdarstellung, d.h. das Konzept, daß Objekte sich aus Teilen zusammensetzen und diese möglicherweise selbst wieder aus Teilen bestehen. Dieser Gesichtspunkt ist mit Blick auf das Strukturmodell von Marr [Mar82] zwar nicht neu, muß aber aufgrund der anderen Herangehensweise bei der erscheinungsbasierten Objekterkennung neu untersucht werden. Da keine dreidimensionalen, volume-

trischen Teile von Objekten modelliert werden, stellt sich als nächstes auch die Frage, welche Kriterien ansichtsbasierte Teile charakterisieren und ob eine Hierarchie für die Erkennung vorteilhaft ist. Was die Teile angeht, gibt es derzeit die Vorschläge, diese mit Hilfe lokaler und regionaler Operatoren zu berechnen, oder aber mit Hilfe eines Lernverfahrens aus einfacheren Merkmalen zusammenzustellen.

Die dritte wichtige Frage betrifft die Modellierung der geometrischen Abhängigkeiten zwischen den Teilen. Konkret bedeutet dies, ob die Auftrittswahrscheinlichkeit eines bestimmten Teils eines Objekts davon abhängt, ob und wo ein anderes Teil gefunden wird. Festlegungen zu diesem Punkt des Modells betreffen nicht nur die Erkennungsleistung des Gesamtsystems, sondern in erster Linie den Rechenaufwand.

Im folgenden werden die genannten Punkte anhand aktueller erscheinungsbasierte Objektmodelle genauer beschrieben.

Hierarchie der Modelle

Die hierarchische Organisation aktueller Objektmodelle reicht von globalen Ansätzen, bei denen Objekte nur als ganzes betrachtet werden, bis zu einer mehrfach verschachtelten Beschreibung von Teilen als Zusammensetzung aus jeweils weiteren Teilen. Die folgenden Abschnitte beschreiben Modelle mit verschieden vielen Hierarchiestufen und geben die Gründe für deren Einführung an.

Globale Ansätze In der von Turk und Pentland [TP91] eingeführten und von Murase und Nayar [MN93, MN95, NMN96] allgemeiner beschriebenen Methode werden Objekte anhand vollständiger 2D-Ansichten modelliert. Um den Speicheraufwand zur Modellierung aller Ansichten zu verringern, werden nur prototypische Ansichten gespeichert, zwischen denen linear interpoliert wird. Diese werden über eine Hauptkomponentenanalyse ermittelt.

Zunächst führen Nayar et al. [NMN96] den Begriff des *visuellen Arbeitsbereichs* (orig. *Visual Workspace*) ein, der die Menge aller in einem Anwendungsbereich möglichen Erscheinungen eines Objekts beschreibt. Die Erscheinung eines Objekts hängt vor allem von der Objektgeometrie, den Oberflächeneigenschaften, der Beleuchtung und der Betrachterperspektive ab. Die für einen Anwendungsbereich relevanten Größen $\{\theta_1, \theta_2, \dots\}$, beispielsweise der Drehwinkel eines Objekts oder die Position einer Lampe, sind die Parameter des visuellen Arbeitsbereichs. Das Modell wird aus einer Stichprobe des visuellen Arbeitsbereichs berechnet. Aus praktischen Erwägungen skalieren Murase und Nayar die Bilder der Stichprobe, sodaß alle Abbildungen gleich groß sind. Außerdem wird die Gesamthelligkeit der Bilder normiert, was auf einfache Weise einen weiteren freien Parameter entfernt.

Die Stichprobe läßt sich so als eine Menge $\{\boldsymbol{I}_1, \dots, \boldsymbol{I}_{n_I}\}$ von n_I Bildern $\boldsymbol{I}$, beschreiben, wobei ein Bild ein Vektor aus n_x Bildpunkten $\boldsymbol{I} = (\iota_1, \dots, \iota_{n_x})^\top$ ist, deren Helligkeit gemäß der Gleichung $||\boldsymbol{I}|| = 1$ normiert ist. Als nächstes

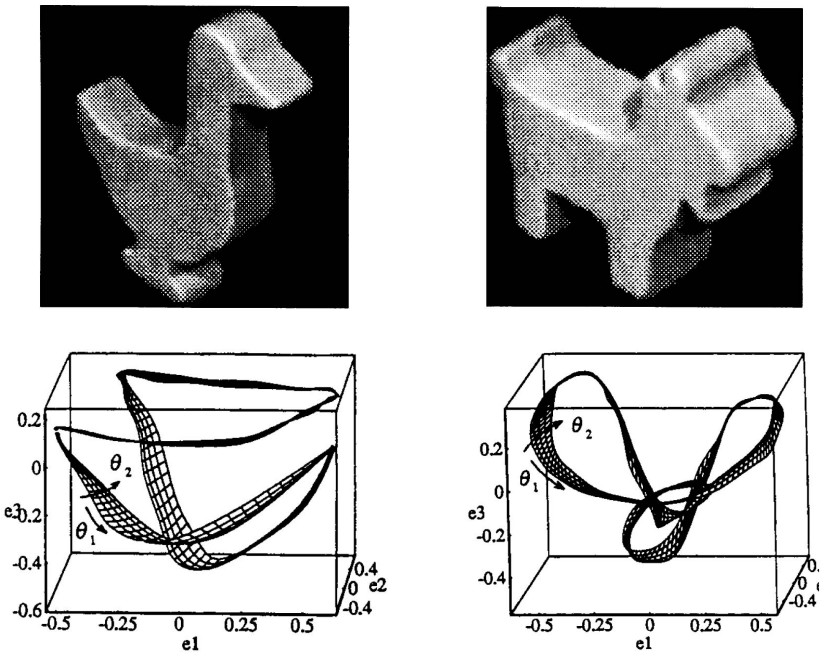


Abbildung 3.16: **Zwei Objekte und ihr jeweiliger parametrischer Eigenraum [NMN96].** Die Achsen $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ sind die Eigenvektoren des universellen Eigenraums. Die Parameter θ_1, θ_2 geben die Rotation des Objekts (θ_1) und die Richtung der Beleuchtung (θ_2) an.

wird eine Kovarianzmatrix Q berechnet. Dazu wird der Mittelwert

$$\boldsymbol{I}_\Sigma = \frac{1}{n_I} \sum_{j=1\ldots n_I} \boldsymbol{I}_j$$

aller Bilder von der Stichprobe subtrahiert, was eine Bildermatrix

$$X = [\boldsymbol{I}_1 - \boldsymbol{I}_\Sigma, \ldots, \boldsymbol{I}_{n_I} - \boldsymbol{I}_\Sigma]$$

mit n_I Spalten und n_x Zeilen ergibt. Die Kovarianzmatrix ergibt sich zu

$$Q = XX^\top.$$

Durch das Auflösen der Gleichung

$$\lambda_k \mathbf{e}_k = Q\mathbf{e}_k$$

ergeben sich n_x Eigenvektoren $\mathbf{e}$. Diese spannen einen neuen Vektorraum auf, von Murase und Nayar [MN95] *universeller Eigenraum* (orig. *universal eigenspace*) genannt, in dem sich die Stichprobe ebenfalls darstellen läßt. Für ein einzelnes Bild ergibt sich die neue Koordinate

$$\boldsymbol{f}_j = (\boldsymbol{e}_1, ..., \boldsymbol{e}_n)^\top (\boldsymbol{I}_j - \boldsymbol{I}_\Sigma).$$

Dabei gibt ein Eigenwert λ_j an, wie wichtig die Achse j zur Darstellung der Stichprobe ist. Da Q eine $n_x \times n_x$-Matrix ist, ist die Berechnung numerisch anspruchsvoll. Normalerweise genügt es, nur die Λ Eigenvektoren mit den größten Eigenwerten zu behalten und die anderen zu verwerfen, um die Elemente der Stichprobe ausreichend gut unterscheiden zu können. Die Bildvektoren der Stichprobe können also gemäß der Gleichung

$$\boldsymbol{f}_j = (\boldsymbol{e}_1, ..., \boldsymbol{e}_\Lambda)^\top (\boldsymbol{I}_j - \boldsymbol{I}_\Sigma). \tag{3.1}$$

auf wenige (bei Murase und Nayar auf 10 bis 20) Dimensionen verkürzt dargestellt werden, was ihre Handhabung stark vereinfacht.

In dem universellen Eigenraum wird nun ein zweiter Raum erstellt, der das eigentliche Objektmodell darstellt. Dazu wird zum einen ausgenutzt, daß daß sich ähnliche Bilder im universellen Eigenraum auf ähnliche Koordinaten abbilden, und zum anderen, daß die Erscheinungen eines Objekts für kleine Änderungen der Parameter des visuellen Arbeitsbereichs meistens sehr ähnlich sind. Werden nun alle Bilder der Stichprobe in den universellen Eigenraum abgebildet, liegt die Stichprobe auf einer Mannigfaltigkeit, die eine kontinuierliche Funktion über den Parametern des visuellen Arbeitsbereichs bildet. Diese Mannigfaltigkeit ist der *parametrische Eigenraum*. Eine auf wenige Dimensionen reduzierte Darstellung der parametrischen Eigenräume von zwei Objekten zeigt die Abbildung 3.16. Die Koordinatenachsen sind hier die drei Eigenvektoren mit den höchsten Eigenwerten. Als Parameter des visuellen Arbeitsbereichs wurden der Betrachtungswinkel der Objekte auf einem Drehtisch sowie die Position einer ungerichteten Lichtquelle in Form eines Winkels bezüglich Objekt

und Kamera gewählt. Der parametrische Eigenraum ist daher als Fläche in den universellen Eigenraum eingebettet.

Der parametrische Eigenraum wird zur Objekterkennung genutzt, indem ein unbekanntes Bild gemäß Gleichung 3.1 in den universellen Vektorraum projeziert wird. Falls die Projektion nahe dem parametrischen Eigenraum liegt, wird das nächste Abtastbild des visuellen Arbeitsbereichs bestimmt. Die Parameter dieses Bildes werden für dann dem unbekannten Bild zugeordnet. Um dabei eine erschöpfende Suche über alle Bilder der Stichprobe zu vermeiden, schlagen Murase und Nayar vor, die Stichprobe mit Hilfe eines Entscheidungsbaums zu sortieren oder die Nächster-Nachbar-Zuordnung über ein Radial-Basis-Netzwerk zu erlernen. Die Klassifikationsgüte hängt davon ab, wie gut sich die parametrischen Eigenräume verschiedener Objekte im universellen Eigenraum unterscheiden lassen.

Die Darstellungen verschiedener Eigenräume [NMN96] deuten an, daß die Objektform einen stärkeren Einfluß auf die parametrischen Eigenräume hat als die Beleuchtung. Trotzdem kann die Methode auch auf vorverarbeitete Bilder angewandt werden, um eine höhere Unabhängigkeit von Beleuchtungseffekten zu erreichen. Darüberhinaus zeigen Nayar und Murase [NM94], daß sich für Objekte mit diffuser Oberfläche starke Vereinfachungen für den parametrischen Eigenraum ergeben. In diesem Fall kann der parametrische Eigenraum durch Interpolation aus nur wenigen Stichprobenelementen konstruiert werden. Die Anzahl der nötigen Stichprobenelemente verringert sich zusätzlich, wenn die Objektgeometrie Symmetrien enthält.

Den Vorteilen des Verfahrens, nämlich der Unabhängigkeit von Beleuchtungsmodellen oder der Extraktion geometrischer Merkmale, stehen jedoch auch Nachteile entgegen, die zum einen auf der globalen Arbeitsweise des Verfahrens beruhen. Da einzelne Objekte durch Vektoren aus Bildpunkten beschrieben werden, funktioniert das Verfahren nur, wenn die Objekte das Bild vollständig ausfüllen und der Hintergrund nicht zu stark variiert. Dies ist vor allem bei Lokalisierungsaufgaben nachteilig, d.h. bei der Suche nach kleinen Objekten in größeren Bildern, da passende Bildausschnitte ausgewählt werden müssen, in denen die Objekte gut zentriert sind. Zum anderen kann das Verfahren Überdeckungen nicht behandeln, da diese direkte Auswirkungen auf die Achsen der Eigenräume haben.

Objektmodellierung anhand teilebasierter 2D-Ansichten Die populärste Methode zur Lösung des Verdeckungsproblems besteht darin, die globalen Ansichten in Teile zu zerlegen. Es ergibt sich ein hierarchisches Modell aus zwei Ebenen: Auf der Basisebene werden Teile ansichtsbasiert modelliert, auf der abstrakteren Ebene werden die Beziehungen zwischen den Teilen modelliert. Da schon die Erkennung einer Untermenge aller Teile genügend Hinweise auf das Vorhandensein des vollständigen Objekts liefert, ist dieser Ansatz tolerant gegen teilweise Verdeckungen. Wenn die Beziehungen zwischen den Teilen auch noch eher lose modelliert werden, ergibt sich zusätzlich eine gewisse Toleranz gegen Störungen und Objektverformungen. Ein weiterer Vorteil ist die Möglichkeit,

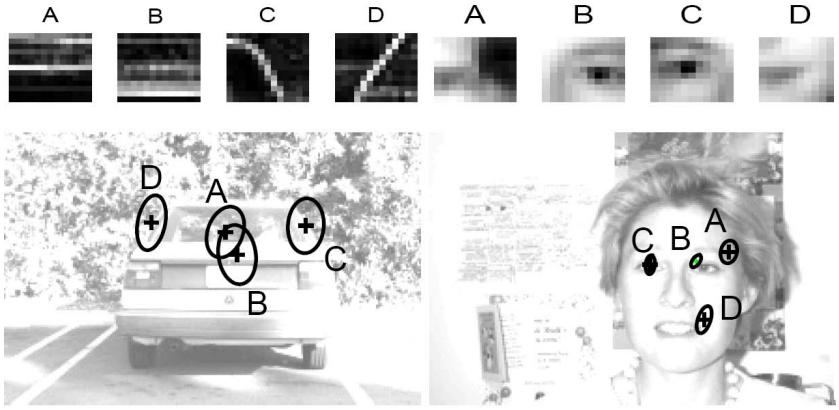


Abbildung 3.17: **Teilebasierte Modellrepräsentation [WWP00]:** Links ein Modell aus vier Teilen zur Erkennung von Autos, rechts für Gesichter. Die kleinen Bilder oben sind die im Modell enthaltenen Teile (Autoteile hochpaßgefiltert). Die Markierungen in den Beispielbildern darunter geben die relativen Positionen der Teile an sowie eine räumliche Toleranz.

Teile nur einmal zu modellieren und zur Modellierung verschiedener Objekte zu benutzen. Dadurch lassen sich Modelle kompakter speichern und effizienter auswerten.

Die Abbildung 3.17 zeigt die Modelle von zwei Objekten nach dem einflußreichen Constellation-Modell von Weber et al. [WWP00]. Die Objekte werden in diesem Beispiel durch jeweils vier Teile dargestellt.

Zu den interessanten Ansätzen gehören auch

- die Weiterentwicklungen des Constellation-Modells [BWP98, WWP00] durch Fergus, Perona und Zisserman (z.B. [FPZ06]).
- das Objekterkennungssystem von Mikolajczyk, Leibe und Schiele [MLS06] und sein Nutzen zum Vergleich verschiedener Merkmale [LMS06b].
- das Objekterkennungssystem von Crandall und Huttenlocher [CH06] zur Untersuchung verschiedener Arten von Teilevernetzungen.
- der kubistische Ansatz von Nelson und Selinger [NS98a]: Teile bestehen aus Gruppen von benachbarten Kantenstücken. Die Eigenschaften eines Teils dienen als Schlüssel für eine Datenbank, über die Informationen zur Klasse und Perspektive von übereinstimmenden Objekten gefunden werden können.
- die Boosting-Methode von Viola und Jones [VJ04]: Mittels AdaBoost (s. [FS99]) wird für jedes Teil ein eigener Klassifikator erzeugt. Diese werden in eine Entscheidungssequenz umgeordnet, die eine Objekterkennung in Echtzeit ermöglicht.

Hierarchische Modelle mit einer Teileebene stehen derzeit im Zentrum des Interesses. Die Aufzählung zeigt, daß die interessanten Fragen vor allem die Modellierung der Teile und ihre gegenseitigen geometrischen und stochastischen Abhängigkeiten betreffen.

Gruppierung von 2D-Ansichten zu 3D-Ansichten Selinger [Sel01] fügt den 2D-Ansichten von Objekten eine weitere, abstraktere Hierarchieebene hinzu. Auf dieser zusätzlichen Ebene sind die 2D-Ansichten topologisch nach der jeweiligen Kameraperspektive bei der Bildaufnahme angeordnet, was einer dreidimensionalen Objektdarstellung nahekommt. Für die Objekterkennung ist diese allerdings nicht zwingend notwendig, da die 2D-Ansichten bereits alle notwendigen Informationen besitzen. Stattdessen erlaubt diese Darstellung virtuelle Rundgänge um ein Objekt, daß aufgrund einer 2D-Ansicht bereits erkannt wurde. Selinger zieht hier eine Querverbindung zu der Diskussion über die mentale Rotation beim Menschen (vgl. tarr89) und vergleicht die Objekterkennung auf der 2D-Ebene ihres Modells mit der Objekterkennung aufgrund der im inferioren Temporallappen gespeicherter Ansichten beim Menschen und Affen. Zusätzlich erlaube die 3D-Ebene die ebenfalls diskutierte mentale Rotation.

Ponce et al. [PLRS04] arbeiten dagegen mit einem echten dreidimensionalen Modell und setzen es auch zur Objekterkennung ein. Dieses basiert auf einer Darstellung von Teilen in Form von elliptischen Bildbereichen und deren affinen Abbildung auf einen Einheitskreis nach Mikolajczyk und Schmid [MS02]. Da die Ellipsen in mehreren Rektifizierungsschritten datenabhängig gewählt werden, ist die Teilebeschreibung invariant gegen affine Transformationen. Ponce et al. nehmen darüberhinaus vereinfachend an, daß diese Invarianz auf für die Betrachterperspektive gilt, wenn die Bildbereiche nur ausreichend klein sind. Um den Schritt von zweidimensionalen Teilebeschreibungen zu einem 3D-Modell zu machen, deuten die Autoren die Abbildungsvorschrift auf den Einheitskreis als "fiktionale" Ansicht eines eigentlich kreisförmigen Bildbereichs. Durch Triangulation über mehrere Ansichten eines Objekts aus verschiedenen Perspektiven rekonstruieren Ponce et al. nun die tatsächliche dreidimensionale Struktur. Die 3D-Informationen der Teile schränken während der Objekterkennung den zu untersuchenden Lösungsraum zusätzlich ein. Ein Nachteil des Verfahrens ist, daß es sich nur für starre Objekte eignet.

Vielschichtige Teilehierarchie Ommer, Sauter und Buhmann [OB06, OSB06] begründen ihre Arbeiten an einem mehrstufigen hierarchischen Objektmodell damit, das Kompositionalität ein allgemeines und natürliches Prinzip sei, insbesondere auch im menschlichen Sehsystem. Die Vorteile dieses Ansatzes sehen sie zum einen in der Wiederverwertung einfacher Teile, vor allem jedoch in der leichteren Erlernbarkeit. Ansichten von Objekten aus unterschiedlichen Perspektiven führten häufig zu einer hohen Varianz innerhalb einer Klasse. Statt solche komplexen Klassen direkt und als Ganzes zu erlernen, halten Ommer et al. es für einfacher, die komplexen Klassen in homogenere Teile zu zerlegen und diese getrennt zu behandeln.

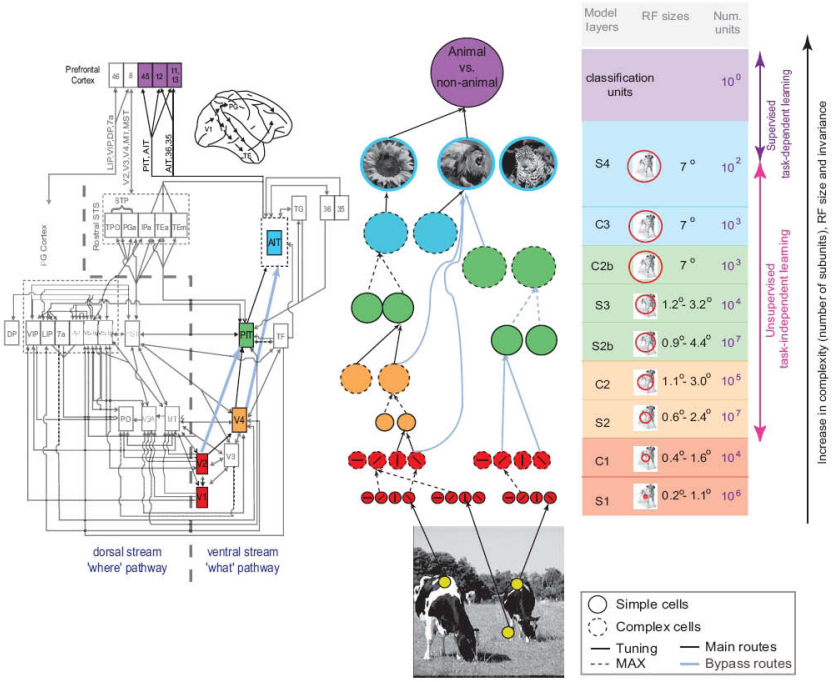


Abbildung 3.18: **Vielschichtige Teilehierarchie (rechts) nach Serre et al. [SOP07].** Der von Serre et al. angestellte Vergleich mit einem Schema des Sehsystem (links) von Affen zeigt starke Einflüsse aus der Neurowissenschaft.

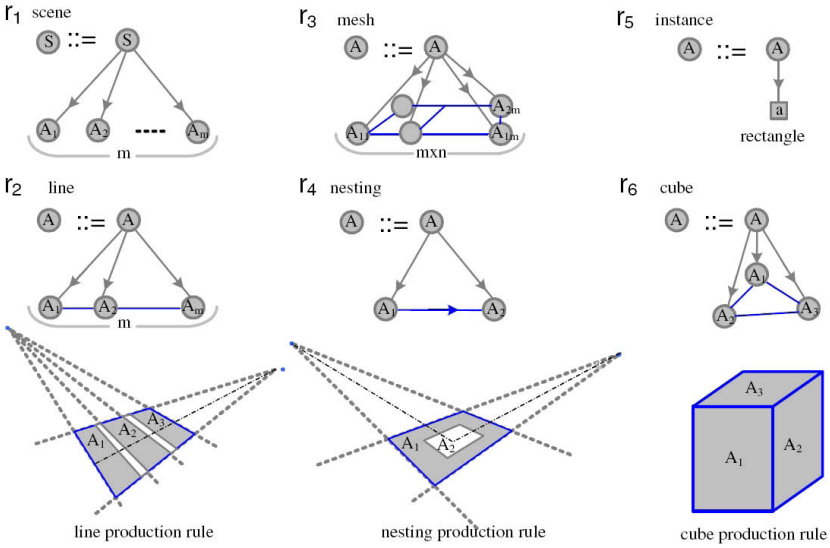


Abbildung 3.19: **Beispiele für Produktionsregeln in der syntaktischen Mustererkennung [HZ05].** Die grauen Pfeile zeigen, welche Terminal- und Nichtterminalzeichen durch die Regelanwendung aus einem Nichtterminalzeichen resultieren. Blaue Pfeile zeigen Beschränkungen auf Attributwerten an. Dies sind hier geometrische Beziehungen, beispielsweise, daß die resultierenden Zeichen entlang einer Linie angeordnet sein sollen.

Auf der untersten Ebene des Modells beschreiben Ommer et al. Bilder durch Kantenpunkte. Diese werden auf der nächsthöheren Ebene zu Geraden und Kurven zusammengesetzt, welche die sog. Basiskompositionen für die weiteren Ebenen des Modells darstellen. Nun werden iterativ immer die zwei Kompositionen zu einer neuen Komposition zusammengefaßt, für die sich die höchste Wahrscheinlichkeit einer bestimmten Klassenzugehörigkeit ergibt. Die so entstehenden abstrakteren Hierarchieebenen haben die Form eines Binärbaums.

Der Nutzen dieses Modells liegt weniger in der direkten Verwendung zur Objekterkennung als in der einfacheren Konstruktion eines Objektmodells mit nur einer Teileebene. Der Binärbaum gibt an, welche Gruppen von Kantenstücken sich besonders gut zur Objekterkennung eignen und identifiziert daher homogene Untergruppen von Objektteilen. Ommer et al. nutzen dies zur Auswahl von guten Teilen für ein einfacheres Objektmodell, das ähnlich zu den in den vorangegangenen Abschnitten erläuterten Modellen mit wenigen Hierarchieebenen ist. Im Vergleich zu einem Modell ohne Hierarchie, bei dem das Ergebnis direkt von den Basiskompositionen abhängt, wird sowohl eine höhere Erkennungsrate erreicht als auch ein höherer Konfidenzwert während der Erkennung ermittelt [OB05].

Serre, Poggio et al. [SWP05a] schlagen dagegen ein vielschichtiges Objektmodell vor, daß direkt der Objekterkennung dient. Von allen hier vorgestellten Modellen ist es am stärksten neurowissenschaftlich beeinflußt.

Es besteht aus einer Menge (in der Größenordnung von 10^7 Elementen) von Einzelklassifikatoren, die eine Vereinfachung der Zellen des visuellen Kortex darstellen und auch ähnlich angeordnet sind (s. Abb. 3.18). Die Klassifikatoren der niedrigsten Ebene werden von Serre et al. "S1-Zellen" genannt und stellen das Gegenstück zu den *einfachen Zellen* (vgl. [Hub88]) des primären visuellen Kortex dar. Sie sind als Gaborfilter (in [RP02]) implementiert und erkennen Kanten in vier verschiedenen Orientierungen. Wie die Zellen des visuellen Kortex sind sie retinotop angeordnet und folgen so den Bildkoordinaten. Die Klassifikatoren der nächsten Ebene, "C2-Zellen" genannt, fassen mit Hilfe einer Maximumoperation die Ergebnisse von Klassifikatoren der darunterliegenden Schicht zusammen. Als Eingabe dienen jeweils Klassifikatoren der gleichen Orientierung eines kleinen Bildausschnitts. C2-Zellen sind als Entsprechung zu den *komplexen Zellen* [Hub88] des visuellen Kortex gedacht und stellen eine gewisse Ortsinvarianz her, ohne die Orientierungsselektivität zu verlieren. Auf der dritten Ebene befinden sich die "S2-Zellen", die als Radial-Basis-Funktion implementiert sind und den euklidischen Abstand zwischen den Zellen am Eingang und einem gespeichertem Vektor berechnen. Der gespeicherte Vektor wird in der Trainingsphase durch zufällige Abtastung von C1-Zellen ermittelt. Als Eingabebilder dienen positive Beispiele einer Trainingsstichprobe. S2-Zellen sind daher nicht mehr nach Bildkoordinaten organisiert. Auf der vierten Ebene findet wieder eine Maximum-Operation über alle S2-Zellen aller Skalierungen und Translationen statt. Diese Ebenen werden mit dem Areal V4 des visuellen Kortex verglichen.

In einer neueren Arbeit [SOP07] werden weitere Ebenen eingeführt, die wie die Ebenen Drei und Vier aufgebaut sind. Der Signalfluß darf Ebenen überspringen, was eine gewissen Skalierungsinvarianz erlaubt. In [JSWP07] wer-

den Merkmale auch über mehrere Bilder einer Sequenz gebildet. Die Ausgabe der höchsten Ebene wird mit dem posterioren inferioren Temporallappen (PIT) verglichen und dient der Erkennung einzelner Ansichten von Objektteilen. Die höchste Schicht vereint diese Ansichten zu vollständigen Objekten, worauf der aufgabenspezifische Teil der Bildverarbeitung folgt. Die Autoren setzen hier auf Support Vector Machines.

Das Verfahren liefert vor allem bezüglich Rotations- und Translationstoleranz gute Erkennungsraten, schlechtere bezüglich Verdeckungen. Für letztere machen die Autoren die fehlenden Rückverbindungen in der Hierarchie verantwortlich. Sie erklären dazu, daß sich das System in gewisser Weise ähnlich verhält wie der visuelle Kortex in den ersten 50ms nach dem Ansetzen eines Stimulus, einer Zeitdauer, die zu kurz für Rückkopplungen aus den höheren Gehirnarealen ist. Rückkopplungen werden auch in anderen aktuellen erscheinungsbasierten Systemen nicht modelliert.

Einen sehr theoretischen Ansatz zur Objekterkennung wählt Fu [Fu82, PF75]. Er beschreibt eine Teilehierarchie als Grammatik, so daß ihm die mathematisch fundierten Erkenntnisse und Werkzeuge aus dem Gebiet der formalen Sprachen zur Verfügung stehen [GPC98, Lee96].

Die Abbildung 3.19 zeigt Beispiele für Produktionsregeln P einer Grammatik

$$G = (V_n, V_t, P, \mathfrak{K})$$

bestehend aus einer Menge Nichtterminalzeichen V_n, einer Menge Terminalzeichen V_t, einer Menge von Produktionsregeln P und einer Menge von Konfigurationen $\mathfrak{K}$ wie von Han [HZ05] zur Objekterkennung vorgeschlagen. Terminal- und Nichtterminalzeichen tragen jeweils vektorförmige Attribute, die hier mit $\mathfrak{A}$ bezeichnet sind. Terminalzeichen entsprechen den Primitiven, aus denen sich eine Szene zusammensetzt, beispielsweise Rechtecken oder Linien. Die Attributwerte geben die Parametrisierung dieser Primitive an, beispielsweise die Lage, Orientierung und Größe. Die Produktionsregeln legen fest, wie ein Nichtterminalzeichen in weitere Nichtterminalzeichen oder Terminalzeichen überführt wird, z.B. legt die Regel $\mathfrak{p}$

$$\mathfrak{p} : \nu \rightarrow (\nu_1, \nu_2)$$

fest, daß das Zeichen ν in die Zeichen ν_1, ν_2 überführt wird. Darüberhinaus legt jede Produktionsregel Beschränkungen zwischen den Attributwerten der beteiligten Zeichen fest. Diese haben für das obige Beispiel die Form

$$g_i(\mathfrak{A}(\nu)) = q_i(\mathfrak{A}(\nu_1), \mathfrak{A}(\nu_2)), i = 1, 2, ..., n_B$$

Dabei sind g und q Funktionen auf den Attributen, n_B ist die Anzahl der Beschränkungen zu der Regel. Beschränkungsgleichungen können auch genutzt werden, um während der Objekterkennung Attributwerte von einem Zeichen zum anderen weiterzureichen. Die Konfigurationen $\mathfrak{K}$ sind Regelableitungen, die durch die Ersetzung eines Startzeichens entstehen können. Sie entsprechen den durch die Grammatik darstellbaren Szenen. Laut Basu et al. [BBB05] liegen die Stärken dieses Ansatzes hauptsächlich in der Modellierung regelmäßiger Strukturen. Tanaka [Tan95] untersucht die syntaktische Mustererkennung bezüglich

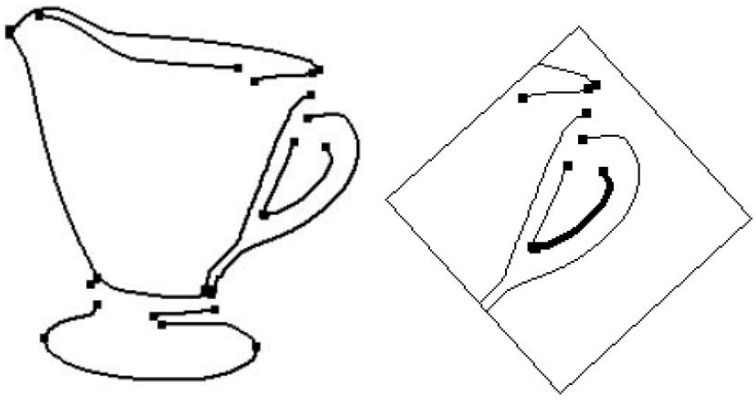


Abbildung 3.20: **Context Patch nach Selinger [Sel01].** Links: Kantenerkennung für ein Objekt. Rechts: Auswahl der Innenkante des Henkels als Schlüsselmerkmal und Normierung eines Ausschnitts um das Schlüsselmerkmal auf dessen Größe und Orientierung.

der Ausdruckskraft der Grammatik, der darstellbaren Szenen, der Trainierbarkeit und der Berechenbarkeit und zieht folgendes Fazit:

- Kontextfreie Grammatiken haben eine für praktische Beispiele zu begrenzte Ausdruckskraft. Kontextsensitive Grammatiken dagegen sind rechentechnisch kaum behandelbar.
- Es gibt Einschränkungen bezüglich der Repräsentation von Prototypen durch Grammatiken.
- Viele Nichtterminale und Produktionsregeln haben keine anschauliche Bedeutung, was das Anpassen des Modells an selbst geringfügig veränderte Muster stark erschwert.
- Es ist keine zwingend konvergente Methode bekannt, um Grammatiken für unendliche Sprachen zu erzeugen. Die Suche nach einer solchen Methode wird durch die Unanschaulichkeit der Regelableitungen erschwert.

Modellierung der Teile

In den vorangegangenen Abschnitten wurde die Teilehierarchie der aktuell verwendeten Modelle beschrieben. Als nächstes wird dargestellt, in welcher Form die Teile modelliert werden. Hier läßt sich zwischen einer operatorbasierten und einer aus einfachen Basismerkmalen erlernten Teilebeschreibung unterscheiden.

Bei der operatorbasierten Teilemodellierung dienen zumeist kleine Bildausschnitte als Teile. Die Position der Teile wird durch einen Filteralgorithmus, den *Detektor*, festgelegt. Die lokale Umgebung um die durch den Detektor ermittelten Punkte wird durch einen Deskriptor beschrieben. Dieser gibt als Vektor z.B.

die Parametrisierung eines bestimmten Merkmals an oder zählt die Bildpunkte direkt auf. Traditionell dienen Linien, Kurven, Ecken oder Kreuzungen als Merkmale, was den Bereichen V1 und V2 des visuellen Kortex entspricht.

Ein Modell, das auf Linien und Kurven basiert, wird von Selinger und Nelson vorgeschlagen [Sel01, SN99, NS98b, NS98a, NS98c, NS00]. Als Teile dienen jeweils Mengen benachbarter Kurven innerhalb eines quadratischen Bildausschnitts, da einzelne Kurvenstücke für einen zuverlässigen Vergleich von Objekten nicht aussagekräftig genug sind. Diese werden von Selinger und Nelson *Context Patch* genannt (s. Abb. 3.20). Die jeweils größte Kontur eines Context Patches wird als Schlüsselkontur ausgewählt. Um die Toleranz gegen perspektivische Transformationen zu steigern, werden die Context Patches bezüglich der Rotation und Größe der Schlüsselkonturen normiert. Die Context Patches werden in einer Datenbank gespeichert zusammen mit Informationen über das entsprechende Ursprungsbild. Letztere betreffen die Perspektive der Aufnahme und die Art des Objekts und erlauben so die Bildung von Hypothesen über ein unbekanntes Objekt, falls ein bestimmter Context Patch gefunden wurde.

Ein Modell, das Kurven mit Flächenmerkmalen kombiniert, wird von Stommel und Kuhnert [SK05, SK06] vorgestellt. Der Detektor wählt sowohl Punkte auf Kanten als auch auf Flächen aus. Der Deskriptor gibt Aufschluß über die Art des Merkmals und speichert für Kanten die Gradientenrichtung und für Flächen die Rot-, Grün- und Blau-Werte. Die Kombination der verschiedenen Merkmale verhindert eine zu starke Spezialisierung auf eine bestimmte Klasse von Bildmaterial und erlaubt die Erkennung von Objekten in realen Bildern.

Baker et al. [BNM98] schlagen als Verbesserung der gängigen Detektoren eine Projektionsmethode ähnlich dem Ansatz von Nayar und Murase [NMN96] vor. Sie fassen Detektoren abstrakt als Funktion auf einem lokalen Bildbereich auf. Für die Intensitätswerte des Bildbereichs, auf denen der Detektor anspricht, bildet die Funktion auf einen passend parametrisierten Deskriptor ab. Die möglichen Deskriptoren liegen demnach auf einer Mannigfaltigkeit mit der Dimensionalität des Deskriptors in einem Vektorraum, der aus den Punkten des lokalen Bildbereichs gebildet wird. Baker et al. argumentieren, daß zur Erzeugung von Detektoren in der üblichen geschlossenen Form Vereinfachungen am Merkmalsmodell vorgenommen werden müssen und das Sensorverhalten vernachlässigt wird. Stattdessen schlagen sie vor, die Mannigfaltigkeit durch Prototypen darzustellen und die Detektion anhand einer Nächster-Nachbar-Suche in einem niederdimensionalen Eigenraum durchzuführen. Der Ansatz scheint jedoch noch keine allzu breite Verbreitung gefunden zu haben. Einer Bemerkung von Lowe [Low99] nach zu urteilen, könnte die Nähe zu einem aufwendigen Template-Matching über alle Bildpositionen ein größeres Interesse verhindern.

Lowe schlägt dagegen eine Kombination aus Detektor und Deskriptor namens SIFT (Scale Invariant Feature Transform) vor, die invariant ist gegen Skalierung, Translation und Rotation sowie tolerant gegenüber Beleuchtungsschwankungen und 3D-Projektionen [Low99]. Als direkte Motivation dient die Arbeit von Edelman, Intrator und Poggio [EIP97] über die Bedeutung der Arbeitsweise komplexer Zellen der primären Sehrinde für die Objekterkennung. Edelman et al. finden anhand eines korrelationsbasierten Klassifikators heraus,

Name	**Wahl der Position**	**Wahl der Skalierung**
Harris-Laplace	Harris-Eckenoperator	Laplacian-of-Gaussian
DoG	Lokale Maxima	Difference-of-Gaussian
Hessian-Laplace	Lokale Maxima der Hesse'schen Determinante	Laplacian-of-Gaussian
Salient regions	Entropiemaximierung durch lokale Histogramme	
MSER	Maximierung der Stabilität einer Bildsegmentierung	

Tabelle 3.1: Von Mikolajczyk et al. [MLS05] getestete Regionendetektoren

daß die Translationstoleranz komplexer Zellen entscheidend die Erkennungsrate steigert. Die von Lowe vorgeschlagene Methode hat allerdings nur von der Idee her Ähnlichkeit mit der von Edelman et al.

Lowe beginnt die Bestimmung der Deskriptorpositionen mit dem Aufbau einer Bildpyramide. Eine neue Pyramidenstufe wird iterativ dadurch gebildet, daß die untere Stufe zweimal mit einem Gaußkern der Standardabweichung $\sigma = \sqrt{2}$ geglättet wird, was jeweils einer Gesamtglättung von $\sigma = 2$ entspricht. Die nächste Pyramidenstufe ergibt sich durch Unterabtastung mit einem Verhältnis von 1 : 1,5. Zwischen den beiden geglätteten Bildern einer Pyramidenstufe wird jeweils die Differenz berechnet. Die lokalen Minima und Maxima der Differenzbilder werden mit den entsprechenden Punkte auf den benachbarten Pyramidenebenen verglichen. Wenn sie auch diesen Test als Minima oder Maxima passieren, werden sie als Position für einen Deskriptor ausgewählt. Die Methode spricht nur auf strukturierte Bildbereiche an, sodaß eine gewisse Mindestinformation für die Objekterkennung vorhanden ist. Zur Berechnung des Deskriptors wird eine Kantenerkennung in Form einer einfachen Pixelsubtraktion auf den Pyramidenebenen durchgeführt. Punkte ab einem Gradientenbetrag größer als 10 Prozent des Maximalen Gradientenbetrags gelten als Kantenpunkte. Zu jedem Kantenpunkt wird außerdem die Gradientenrichtung berechnet. Aus dem Maximum des Histogramms über eine gewisse Umgebung um einen gefundenen Punkt wird die Hauptorientierung der Umgebung bestimmt. Der Deskriptor besteht nun aus 20 Histogrammen, die im Bild auf zwei benachbarten Pyramidenstufen zu einem regelmäßigen Gitter angeordnet sind. Jedes Histogramm gibt die Häufigkeiten der acht Hauptrichtungen von Kantenpunkten in der Umgebung an. Der SIFT-Deskriptor enthält damit 160 Werte. Um den Deskriptor rotationsinvariant zu machen, werden alle Gradientenrichtungen durch Subtraktion auf die Hauptorientierung normiert.

SIFT-Deskriptoren und deren Abwandlungen sind zur Zeit sehr verbreitet [BZM08, FPZ03, FPZ07, OB06] und werden aufgrund ihrer Toleranz gegenüber dreidimensionalen Abbildungen auch bei der Navigation mobiler Roboter erprobt [SBO$^+$07, Tom06, WZ06]. Mikolajczyk et al. [MLS05] testen daher verschiedene Kombinationen von Detektoren und Deskriptoren auf ihre Eignung für die Objekterkennung. Zum Einsatz kommen die in Tab.3.1 genannten fünf ska-

Name	**Bemerkung**	**Größe**
SIFT	Histogramm über Gradientenposition und -orientierung, hier allerdings nur auf einer Skalenebene	128
GLOH	Ähnlich SIFT, aber feiner	128
PCA-SIFT	Dimensionsreduktion durch Hauptkomponentenanalyse	36
Invariante Momente	Berechnet auf lokalen Bildableitungen	20
Kreuzkorrelation	Glättung, Unterabtastung und Normierung einer Bildregion	

Tabelle 3.2: Von Mikolajczyk et al. [MLS05] getestete Deskriptoren

lierungsinvarianten Regionendetektoren und die fünf in Tabelle 3.2 aufgeführten Deskriptoren. Als optimale Kombination für die Klassifikation der Bilder mehrerer Datenbanken ergeben sich Hesse-Laplace-Regionen in Kombination mit einem erweiterten SIFT-Deskriptor (GLOH). Letzterer stellt feiner aufgelöste Gradientenhistogramme dar als der originale SIFT-Operator und wird daher per Hauptkomponentenanalyse auf 128 Merkmalsdimensionen reduziert wird. Stark und Schiele [SS07] stellen in dem Zusammenhang auch fest, daß die Wahl des Merkmalsorts wichtiger ist als die Art der Beschreibung des lokalen Bildbereichs.

Leibe und Schiele [LS03] gehen der Frage nach, wie leistungsfähig erscheinungsabhängige Merkmale im Vergleich zu den weniger beleuchtungsabhängigen Konturmerkmalen sind. Dazu wenden sie 4 verschiedene Methoden auf eine Datenbank mit über 3000 Bildern von 8 verschiedenen Objektklassen an. Zum Einsatz kommt ein χ^2-Test zum Vergleich von Farbmerkmalen, ein χ^2-Test zum Vergleich von Texturhistogrammen, eine Hauptkomponentenanalyse auf der globalen Objektkontur und eine Methode basierend auf lokalen Konturen. Die besten Ergebnisse ergeben sich für die konturbasierten Methoden mit einem geringfügigen Vorsprung für die lokalen Konturen, die schlechtesten Ergebnisse liefert die Farbklassifikation. Da sich die Klassifikatoren bezüglich verschiedener Objektarten unterschiedlich verhalten, kombinieren die Autoren alle Klassifikatoren mit Hilfe eines Entscheidungsbaums und erhalten einen neuen Klassifikator, der bessere Ergebnisse liefert als jedes einzelne Verfahren. Tests verschiedener Kombinationen zeigen eine besondere Bedeutung der Objektkontur für die Klassifikation.

Stark und Schiele [SS07] berichten darüberhinaus, daß erscheinungsbasierte Merkmale bessere Ergebnisse liefern, wenn die Klassifikation auf reinen Auftrittswahrscheinlichkeiten von Merkmalen beruht. Mit zunehmender Ortsinformation steigt jedoch die Leistungsfähigkeit von konturbasierten Merkmalen bis diese schließlich bessere Ergebnisse liefern.

Jedoch steht nicht bei allen Ansätzen einzig die Fehlerrate bei der Objekterkennung im Vordergrund. Viola und Jones [VJ04] stellen ein System zur Erken-

nung von Gesichtern vor, das vor allem aufgrund seiner Echtzeitfähigkeit Aufmerksamkeit erregt (z.B. bei [ASG07]). Durch die Ausrichtung auf eine schnelle Erkennung sind Detektor und Deskriptor so einfach wie möglich entworfen: Die Teile eines Objekts setzen sich jeweils aus zwei bis vier aneinandergrenzenden Rechtecken zusammensetzen, denen jeweils ein Vorzeichen zugeordnet wird. Dieses bezieht sich auf die Bildpunkte innerhalb des Rechtecks. Der Wert eines Merkmals berechnet sich aus der Summe bzw. Differenz der Bildpunkte innerhalb der Rechtecke. Da eine große Zahl dieser Merkmale möglich und erforderlich ist, entwickeln Viola und Jones mit den *Integral Images* eine für die Deskriptorberechnung besonders effiziente Bilddarstellung. Diese geben für jede Bildposition die Summe aller Pixelintensitäten links oberhalb der Position an.

Zu den Verfahren mit einer erlernten Teilezusammensetzung aus einfacheren Merkmalen gehört die bereits erwähnte biologisch inspirierte Methode von Serre, Wolf und Poggio [SWP05b] oder aber auch der Ansatz von Ommer, Sauter und Buhmann [OB06, OSB06]. Letztere modellieren Teile als Zusammensetzung von SIFT-ähnlichen Deskriptoren. Dazu werden benachbarte Deskriptoren in mehreren Schritten zu den sog. *Kompositionen* zusammengefaßt. Es ergibt sich eine Hierarchie von Kompositionen, mit eng benachbarten Deskriptoren an der Basis und Kompositionen aus vielen Deskriptoren über größere Bildbereiche an der Spitze. Die zu Objekterkennung an besten geeigneten Kompositionen der Hierarchie werden ausgewählt. Sie stellen dann die Teile des Objektmodells dar.

Crandall und Huttenlocher [CH06, CH07] optimieren das Modell schon bei der Auswahl der Teile auf die Klassifikationsgüte. Sie beginnen die Auswahl der Objektteile mit einem einfachen Vorverarbeitungsschritt, der jedem Pixel bestimmte grob quantisierte Attributwerte zuordnet. Diese können die Gradientenrichtung oder die Farbe sein. Als nächstes werden aus den positiven Elementen der Stichprobe zufällig kleine quadratische Bildausschnitte in fest eingestellten Größen ausgewählt. Diese werden mit einem iterativen Verfahren so optimiert, daß sie mit möglichst vielen positiven Stichprobenelementen korrelieren. Anschließend werden Ausschnitte mit einer geringen Trefferwahrscheinlichkeit verworfen. Für die verbleibenden Ausschnitte werden nun die Wahrscheinlichkeiten für alle Attributwerte über der Position innerhalb des Ausschnitts berechnet. Für jeden Ausschnitt ergeben sich dadurch so viele separate Wahrscheinlichkeitsverteilungen wie es unterschiedliche Attributwerte gibt. Die Wahrscheinlichkeiten ergeben sich durch die in den Stichprobenbildern auftretenden Attributwerte.

Gegenseitige Abhängigkeiten der Teile

Die Modellierung der Beziehungen zwischen den Teilen hat nicht nur einen großen Einfluß auf die Erkennungsraten, sondern auch in die Geschwindigkeit eines Objekterkennungssystems. Der Grund dafür ist die Komplexität der Berechnungen von Verbundwahrscheinlichkeiten zwischen Eigenschaften von Teilen. Dies tritt in den strenger wahrscheinlichkeitstheoretisch formulierten Ansätzen von Fergus et al. [FPZ06, FPZ07] und Crandall et al. [CFH05, CH06, CH07] besonders deutlich hervor, die jeweils Variationen des

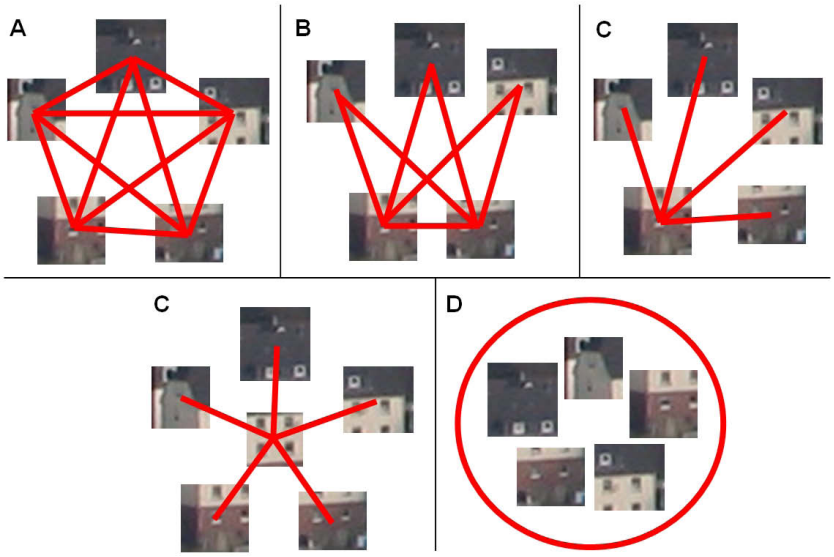


Abbildung 3.21: **Modellierung geometrischer Abhängigkeiten zwischen Teilen.** A: Vollvernetzt, B: 2-Fächer, C: 1-Fächer/Sternmodell, D: Bag of features

Constellation-Modells [BWP98, WWP00] sind. Bei diesen Ansätzen läßt sich die den Teilen übergeordnete Struktur durch einen ungerichteten Graphen beschreiben. Dieser besteht aus den m Teilen $\boldsymbol{T} = (t_1, t_2, ..., t_m)$ mit einer Parametrisierung $\boldsymbol{A}_M = (a_1, a_2, ..., a_m)$, durch die einem Teil t_i ein Attribut a_i zugeordnet wird. Attribute können beispielsweise die lokale Bildstruktur beschreiben. Zwischen verschiedenen Teilen t_i, t_j des Modells können stochastische Abhängigkeiten $B = \{b_{ij}, i \neq j\}$ bezüglich bestimmter abhängiger Attribute l_i einer zusätzlichen Teileparametrisierung $\boldsymbol{L}_M = (l_1, l_2, ..., l_m)$ modelliert werden. Meistens wird die Teileposition als abhängiges Attribut modelliert, damit geometrische Beziehungen zwischen Teilen ausgedrückt werden können (s. Abb. 3.21). Die Attribute a_i werden dagegen als stochastisch unabhängig betrachtet. Ein Modell $\boldsymbol{M}$ besteht also aus $(\boldsymbol{T}, B, \boldsymbol{L}_M, \boldsymbol{A}_M)$.

Werden nun in einem Testbild nun $\mathfrak{n}$ interessante Kandidaten für Teile mit den entsprechenden Attributen $\boldsymbol{A}_I = (a_1, a_2, ..., a_\mathfrak{n})$ und $\boldsymbol{L}_I = (l_1, l_2, ..., l_\mathfrak{n})$ gefunden, läßt sich eine Entscheidung über das Vorhandensein eines bestimmten Objekts über die Gleichung

$$\rho = \frac{p(Objekt|\boldsymbol{L}_I, \boldsymbol{A}_I)}{p(Hintergrund|\boldsymbol{L}_I, \boldsymbol{A}_I)} \tag{3.2}$$

treffen, welche nach dem Satz von Bayes die praktikable Form

$$\rho = \frac{p(\boldsymbol{L}_I, \boldsymbol{A}_I|Objekt)p(Objekt)}{p(\boldsymbol{L}_I, \boldsymbol{A}_I|Hintergrund)p(Hintergrund)}$$

annimmt. Da die wahre Objekt- und Hintergrundstatistik nicht zugänglich ist, wird die Objekterkennung anhand eines Objektmodells $\boldsymbol{M}$ und eines Hintergrundmodells $\boldsymbol{M}_H$ durchgeführt:

$$\rho = \frac{p(\boldsymbol{L}_I, \boldsymbol{A}_I|\boldsymbol{M})p(Objekt)}{p(\boldsymbol{L}_I, \boldsymbol{A}_I|\boldsymbol{M}_H)p(Hintergrund)} \tag{3.3}$$

Die a-priori-Wahrscheinlichkeiten $p(Objekt)$ und $p(Hintergrund)$ können direkt aus der Stichprobe geschätzt werden. Da die Anzahl der Kandidaten üblicherweise nicht mit der Anzahl von Teilen im Modell übereinstimmt, müssen alle möglichen Zuordnungen von Kandidaten zu Teilen überprüft werden. Eine solche Zuordnung wird Hypothese genannt [WWP00] und ist ein Vektor $\boldsymbol{h}$ mit m Einträgen, der jedem Teil einen Kandidaten zuordnet. Doppelzuordnungen einzelner Kandidaten sind unzulässig, fehlende Teile werden durch den Wert Null markiert. Zur Berechnung von $p(\boldsymbol{L}, \boldsymbol{A}|\boldsymbol{M})$ müssen alle Hypothesen $\mathfrak{H}$ überprüft werden. Es ergibt sich die Gleichung

$$p(\boldsymbol{L}_I, \boldsymbol{A}_I|\boldsymbol{M}) = \sum_{\boldsymbol{h}\in\mathfrak{H}} p(\boldsymbol{L}_I, \boldsymbol{A}_I, \boldsymbol{h}|\boldsymbol{M}),$$

welche sich nach

$$\sum_{\boldsymbol{h}\in\mathfrak{H}} p(\boldsymbol{L}_I, \boldsymbol{A}_I, \boldsymbol{h}|\boldsymbol{M}) = \sum_{\boldsymbol{h}\in\mathfrak{H}} p(\boldsymbol{A}_I|\boldsymbol{L}_I, \boldsymbol{h}, \boldsymbol{M})p(\boldsymbol{L}_I|\boldsymbol{h}, \boldsymbol{M})p(\boldsymbol{h}|\boldsymbol{M})$$

umformen läßt. Üblicherweise wird angenommen, daß $\boldsymbol{A}$ nicht von $\boldsymbol{L}$ abhängt. Da die Attribute in $\boldsymbol{A}$ voneinander unabhängig sind, vereinfacht sich der erste Faktor zu einem Produkt aus Einzelwahrscheinlichkeiten

$$p(\boldsymbol{A}_I|\boldsymbol{L}_I, \boldsymbol{h}, \boldsymbol{M}) = p(\boldsymbol{A}_I|\boldsymbol{h}, \boldsymbol{M}) = \prod_{i=1}^{m} p(a_{\boldsymbol{h}(i)}|\boldsymbol{h}, \boldsymbol{M}),$$

welches mit Hilfe der Stichprobe schnell bestimmt werden kann. Die Wahrscheinlichkeit $p(\boldsymbol{L}_I|\boldsymbol{h}, \boldsymbol{M})$ ist aufgrund der gegenseitigen Abhängigkeiten der Attribute unter Umständen aufwendiger zu bestimmen. Crandall und Huttenlocher [CFH05] systematisieren die Modellüberlegungen an dieser Stelle, indem sie den Begriff der Clique in die Diskussion bringen. In der Graphentheorie bezeichnet eine Clique eine vollvernetzte Menge von Knoten. Er läßt sich hier einsetzen, um Teileabhängigkeiten zu identifizieren und die Komplexität des Modells zu beschreiben.

Im einfachsten Fall haben alle Cliquen des Modells die Länge 1. In diesem Fall liegen keine gegenseitigen Teileabhängigkeiten vor, was als *Bag of features* bezeichnet wird. In diesem Fall ist

$$p(\boldsymbol{L}_I|\boldsymbol{h}, \boldsymbol{M}) = \prod_{t_i\in\boldsymbol{T}} p(l_i|\boldsymbol{h}, \boldsymbol{M}).$$

Crandall und Huttenlocher zeigen, daß Gleichung 3.2 für ein Bag-of-features-Modell in $\mathcal{O}(m\mathfrak{n})$ gelöst werden kann.

Der nächstkomplexere Fall liegt vor, wenn ein Referenzteil t_r festgelegt wird und die übrigen Teile als von diesem abhängig modelliert werden. Die Länge der maximalen Cliquen beträgt hier 2. Dieses Modell wird 1-Fächer (orig. *1-Fan*) [CFH05] oder Sternmodell [FPZ06] genannt. Die Wahrscheinlichkeit einer bestimmten Konstellation von Parametern $\boldsymbol{L}$ wird durch die Gleichung

$$p(\boldsymbol{L}_I|\boldsymbol{h},\boldsymbol{M}) = p(l_r|h,M) \prod_{t_i \neq t_r} p(l_i|l_r,\boldsymbol{h},\boldsymbol{M}).$$

gegeben. Crandall [CFH05] und Fergus [FPZ06] geben für die Objekterkennung mit dem Sternmodell eine Komplexität von $\mathcal{O}(m\mathfrak{n}^2)$ an.

Crandall formuliert dies noch allgemeiner für Teilestrukturen in Form von $\mathfrak{c}$-Fächern. Diese besitzen $\mathfrak{c}$ untereinander vollvernetzte Referenzteile $R \subseteq T = t_{\boldsymbol{h}(1)}, t_{\boldsymbol{h}(2)}, ..., t_{\boldsymbol{h}(m)}$, d.h. sie bilden eine Clique der Länge $\mathfrak{c}$. Die übrigen Teile $\bar{R} = T - R$ besitzen keine gegenseitigen Abhängigkeiten. Allerdings hängt jedes Teil aus $\bar{R}$ von allen Teilen in R ab. Für eine bestimmte Konstellation $\boldsymbol{L}_R$ von Referenzteilen ergibt sich dann die Gleichung

$$p(\boldsymbol{L}_I|\boldsymbol{h},\boldsymbol{M}) = p(\boldsymbol{L}_R|h,M) \prod_{t_i \in \bar{R}} p(l_i|\boldsymbol{L}_R,\boldsymbol{h},\boldsymbol{M}).$$

Dies führt zu einer Komplexität von $\mathcal{O}(m\mathfrak{n}^{\mathfrak{c}+1})$. Für ein vollvernetztes Modell ist $\mathfrak{c} = m - 1$, was eine Komplexität von $\mathcal{O}(m\mathfrak{n}^m)$ ergibt.

Die Komplexität des Modells macht sich sowohl in der Objektdetektion und Lokalisation als auch im Training des Objekts bemerkbar. Ein vollvernetztes Teilemodell wird beispielsweise im Constellation-Modell [BWP98, WWP00] und Nachfolgearbeiten [FPZ03, FPZ07] eingesetzt. Die Erscheinung der Teile wird unabhängig, die Teilepositionen und relativen Skalierungen dagegen als voneinander abhängig modelliert. Die hohe Komplexität des Modells limitiert in der Praxis jedoch die Anzahl der Teile auf sechs und die Anzahl an Teilekandidaten in einem Bild auf 20-30 [FPZ06]. Da die Leistungsfähigkeit des Systems so nur von wenigen Teilen und Bildpositionen abhängt, hängt das Ergebnis stark von dem eingesetzten Merkmalsdetektor und dem Bildmaterial ab. Für ein Sternmodell erreichen Fergus et al. [FPZ06] dagegen in jeder Hinsicht bessere Ergebnisse. Dabei räumen Fergus et al. ein, daß das Sternmodell gegenüber der vollen Vernetzung den Nachteil hat, daß das Referenzteil immer vorhanden sein muß, um die Verbundwahrscheinlichkeiten zu berechnen. Das Sternmodell hat dagegen weniger Parameter und ist daher unempfindlicher gegen Überlernen. Dazu kommt die erhöhte Zahl von 40 Detektionen pro Bild und eine Teileanzahl von 12. Aufgrund der Einfachheit des Modells und der trotzdem hohen Zuverlässigkeit werden sternförmige Teilebeziehungen am häufigsten modelliert [OSB06, MLS06, Sel01, PL00, HZ05].

Crandall und Huttenlocher [CH06, CH07] testen mit ihrem Erkennungssystem verschiedene Cliquengrößen. Als Modell kommen ein Bag-of-features-Modell, ein Sternmodell und ein 2-Fächer zum Einsatz. Da beim Bag-of-features-Modell keine gegenseitigen Abhängigkeiten modelliert werden, hängt die Erkennung nur von der Anzahl der gefundenen Teile ab und nicht wie bei Fergus et

al. von Relativpositionen und -skalierungen. Beim Sternmodell werden geometrische Beziehungen modelliert. Die Teileanordnung ist daher translationsinvariant. Alle diese Anordnungen lassen sich aufgrund der geringen gegenseitigen Vernetzung effizient berechnen.

Die Autoren stellen zunächst fest, daß die automatische Optimierung der Lage der Objektteile bessere Ergebnisse liefert als eine manuelle Festlegung. Die Autoren geben weiterhin eine Verbesserung im Vergleich zu dem Ansatz von Fergus et al. [FPZ03] mit fester Merkmalsextraktion und vollvernetztem Modell an. Der Übergang von Bag-Modell zum Sternmodell bewirkt eine deutlich höhere Erkennungsrate. Eine genauere Modellierung mit zwei Referenzteilen bringt allerdings keine weiteren Vorteile. Die Autoren merken dazu an, daß die Stichprobenelemente stark unterschiedlich sind und möglicherweise keine allzu genaue Modellierung der Geometrie erlauben. Sie kommen damit zu anderen Ergebnissen als Serre et al. [SWP05a], die ein Bag-Modell favorisieren und damit sehr gute Ergebnisse erzielen. Serre et al. argumentieren, daß die Objektgeometrie im visuellen Kortex auch keine Rolle spiele.

Im Gegensatz zu den meisten anderen Autoren modellieren Burge et al. [BB97, BBM96] Teilebeziehungen nicht aufgrund von räumlicher Nachbarschaft, sondern auf der Ähnlichkeit der Teiledeskriptoren. Der Grund dafür liegt in der speziellen Graph-Matching-Methode, welche die Autoren während der Objekterkennung einsetzen. Aufgrund der hohen Komplexität beim Vergleich von vollvernetzten Teilemodellen erzeugen die Autoren nur teilvernetzte Modelle. Die Entscheidung, welche Teile vernetzt werden, wird anhand eines Abstandsmaßes getroffen. Da bei großen Objekten wichtige Teile weit auseinander liegen können, besteht die Gefahr, daß das Modell in mehrere unverbundene Graphen zerfällt oder das Objekt unvollständig modelliert wird, wenn die Entscheidung aufgrund des räumlichen Abstands der Teile getroffen wird. Da das Erkennungssystem auf Pfade angewiesen ist, die alle Teile eines Objekts verbinden, sind Entscheidungen aufgrund von räumlicher Nachbarschaft ungünstig. Burge et al. lösen dieses Problem, indem sie die den Abstand zwischen den Teilen im Merkmalsraum der Teiledeskriptoren berechnen. Die Autoren geben gute Ergebnisse für die Behandlung von Verdeckungen bzw. fehlenden Teilen an. Dies liegt möglicherweise in der stärkeren Vernetzung im Vergleich zum Sternmodell.

3.2.2 Training der Modelle

Idealerweise optimiert ein gutes Modell die Wahrscheinlichkeit, Objekte zu erkennen, wenn sie im Bild vorliegen. Für Modelle ähnlich dem Constellation-Modell ergibt sich aus Gleichung 3.3, daß die optimalen Modellparameter $\boldsymbol{A}^*$ und $\boldsymbol{L}^*$ das Auftreten der Stichprobe $\boldsymbol{A}_I$, $\boldsymbol{L}_I$ am besten erklären:

$$\boldsymbol{L}^*, \boldsymbol{A}^* = \arg \max_{\boldsymbol{L}_M, \boldsymbol{A}_M} p(\boldsymbol{L}_I, \boldsymbol{A}_I | \boldsymbol{L}_M, \boldsymbol{A}_M) \tag{3.4}$$

Der Aufwand für diese Optimierung hängt von der gegenseitigen Abhängigkeit der Teile ab. Beim Sternmodell hängt die Komplexität quadratisch von der Teilezahl ab, bei vollvernetzten Teilen dagegen exponentiell [FPZ07, CFH05].

Um eine erschöpfende Berechnung aller möglichen Modellparametrisierungen zu vermeiden, geschieht das Training für aufwendigere Modelle praktisch immer über heuristische Verfahren.

Optimierung auf Teile oder Teilebeziehungen

Bei der Auswahl von Teilen für ein Objektmodell kann eine *harte Detektion* oder eine *weiche Detektion* unterschieden werden [BWP98, CFH05]. Bei der harten Detektion bestimmt der Detektor eine Reihe von Bildpositionen, die sich aufgrund der lokalen Bildstruktur zuverlässig wiederfinden lassen. Deskriptoren an diesen Positionen stellen Kandidaten für mögliche Teile dar. In der Trainingsphase werden die Kandidaten als Teile ausgewählt, die bei der Modellierung der gegenseitigen geometrischen Abhängigkeiten eine optimale Erkennungsleistung ergeben. Die Optimierung geschieht nur innerhalb der Grenzen, die sich durch die Vorauswahl des Detektors ergeben. Bei einer weichen Detektion dagegen ist die Auswahl von Teilen an allen Bildpositionen möglich. Eine Vorauswahl von Teilepositionen in Form einer Menge von Kandidaten wird nicht getroffen. Da sich allerdings nicht alle Positionen gleichermaßen als Teilekandidat eignen, wird zu jeder Position eine Bewertung angegeben, die angibt, wie zuverlässig das entsprechende Teil gefunden würde. Diese Bewertung fließt in die Auswahl der Teile und die Optimierung der Teilebeziehungen mit ein. Dadurch ergeben sich zusätzliche Freiheiten beim Training des Modells. Das Modell kann nun auf eine gute Teiledetektion oder auf gute Teilebeziehungen optimiert werden. Auch Kompromisse sind möglich. Wie Burl et al. [BWP98] berichten, ist eine optimale Strategie aus Komplexitätsgründen nur für den Fall möglich, daß alle Teile voneinander unabhängig sind. Für das Training von Modellen mit Verbundwahrscheinlichkeiten, sind keine optimalen Verfahren bekannt.

Burl et al. und Crandall et al. [CH06] berichten übereinstimmend, daß eine weiche Detektion mit einem Kompromiß aus guter Detektion und guten Teilebeziehungen deutlich bessere Ergebnisse liefert als eine harte Detektion. Burl et al. berichten darüberhinaus, daß eine einseitige Optimierung auf gute Teilebeziehungen unter Vernachlässigung der Detektion schlechtere Ergebnisse liefert als ein Kompromiß.

Harte Detektionen besitzen gegenüber einer weichen Detektion jedoch den Vorteil einer einfacheren Handhabung und werden daher häufiger eingesetzt. Oft spielt dabei auch ein gewisser Modularitätsgedanke eine Rolle: Indem Teile durch ihre Ähnlichkeit zu einer Menge abstrakterer Prototypen beschrieben werden, ergibt sich eine kompakte Darstellung, die Ähnlichkeiten zu spärlichen Kodierungen nach Rolls et al. [RTT97] aufweisen. Im folgenden werden einige Methoden zur Auswahl guter Teile dargestellt.

Bestimmung guter Teile

Serre, Wolf und Poggio [SWP05a] modellieren die bereits erwähnten S1-Zellen, welche die unterste Ebene ihrer Teilehierarchie (s. Abb. 3.18) darstellen, als Gaborfilter. S1-Zellen sprechen damit innerhalb einer gewissen Umgebung auf

sinusförmige Helligkeitsverläufe einer bestimmten Richtung, Orientierung und Phase an. Die Parameter der S1-Zellen folgen streng dem biologischen Vorbild: Serre et al. erzeugen zunächst eine Menge S1-Zellen, die den Parameterraum vollständig abdecken. Anschließend testen sie die so erzeugten Zellen anhand von Stimuli, wie sie auch bei der Vermessung echter Nervenzellen eingesetzt werden. S1-Zellen, die nicht mit echten Zellen übereinstimmen, werden verworfen. Es bleiben 16 Filter in 4 Orientierungen übrig. Die Gaborfilter sprechen über verschieden große Bildbereiche an und werden daher in 8 Bänder, d.h. Größenstufen, unterteilt. Die C1-Zellen der nächsten Hierarchiestufe geben das Maximum von S1-Zellen innerhalb eines gewissen Umkreises an. Die Selektivität für bestimmte Muster ist ebenfalls so gewählt, daß sich eine große Ähnlichkeit zur Natur ergibt. C1-Zellen erhalten als Eingabe daher nur Signale von S1-Zellen des gleichen Bandes und der gleichen Orientierung. Dagegen findet die Maximum-Operation über alle Positionen innerhalb einer Umgebung und über alle Filterskalen innerhalb eines Bandes statt. Zum Training der S2-Zellen werden Eingabebilder an das System angelegt und die Ausgaben von Gruppen von C2-Zellen in Form von Prototypen abgetastet. Die C2-Zellen innerhalb einer Gruppe korrespondieren mit kleinen zusammenhängenden Bildbereichen, wobei alle Filter-Orientierungen auftreten. Die Prototypen werden in den S2-Zellen gespeichert. Diese vergleichen eingehende Signale mit den gespeicherten Prototypen und geben ein gaußförmig von der Ähnlichkeit abhängiges Signal an die C2-Zellen der nächsten Schicht. Die C2-Zellen bilden für jeden Prototyp das Maximum über die S2-Zellen aller Positionen und Skalen. Die Ausgaben der C2-Zellen dienen als Eingabe eines aufgabenabhängigen Klassifikators [SWP05a] oder als Eingabe weiterer Schichten [SOP07].

Die Trainingsphase, in der Prototypen festgelegt werden, dient dem Aufbau einer Menge von formspezifischen, allgemeingültigen Einzelklassifikatoren. Diese werden zwar mit Hilfe von positiven und negativen Beispielen trainiert, allerdings findet keine ausdrückliche Optimierung auf eine möglichst hohe Trefferwahrscheinlichkeit statt. Stattdessen sind die Einzelklassifikatoren auf eine möglichst große Ähnlichkeit zu realen Nervenzellen optimiert. Da sie einem nachfolgenden, aufgabenspezifischen Klassifikator als Eingabe dienen, stellen sie die direkte Entsprechung zu Tanakas visuellem Alphabet [Tan96] dar. Serre et al. machen vor allem die Größe des Alphabets, das je nach Schicht aus etwa 10^4 bis 10^7 Muster besteht, für die hohe Leistungsfähigkeit ihres Ansatzes verantwortlich. Die Autoren geben auch an, daß ab etwa 5000 Elementen die Leistungsfähigkeit nur noch geringfügig steigt, sodaß das visuelle Alphabet viele Redundanzen enthält. Die eigentliche Einordnung eines Eingabebildes in eine bestimmte Kategorie findet bei Serre et al. in dem nachfolgenden Klassifikator statt.

Auch bei operatorbasierten Teilemodellen werden visuelle Alphabete eingesetzt. Mikolajczyk et al. [MLS06, MLS05, LMS06b] erzeugen dies, indem sie eine Merkmalsextraktion auf der Trainingsstichprobe durchführen und so eine große Menge von mehreren 10 000 SIFT-ähnlichen Deskriptoren erhalten. Diese werden mittels Clusterung [LMS06a] zu einem *hierarchischen Codebook* gruppiert, das ebenfalls ein visuelles Alphabet darstellt. Die Hierarchie bezieht sich hier

übrigens nicht auf möglicherweise zu erkennende Objekte, sondern auf verschiedene Abstufungen in der Ähnlichkeit der Deskriptoren. Als Ähnlichkeitsmaß bei der Clusterung dient der euklidische Abstand zwischen den Deskriptoren. Zusätzlich zu den gespeicherten Mustern wie bei Serre et al. enthält das Codebook jedoch noch Geometrie- und Kategorieinformationen, die während der Objekterkennung eine Rolle spielen.

Ommer, Sauter und Buhmann [OB06, OSB06] arbeiten ebenfalls mit einem Codebook, das 100 Prototypen von SIFT-ähnlichen Deskriptoren enthält. Dieses wird während der Trainingsphase durch Clusterung der Deskriptoren über das K-Means-Verfahren aufgebaut. Das Codebook wird zur Darstellung nicht gelernter Deskriptoren verwendet, indem für einen neuen Deskriptor die Ähnlichkeit zu allen Einträgen des Codebooks angegeben wird. Anders als in dem Ansatz von Mikolajczyk et al. [MLS06] dienen die Einträge des Codebooks noch nicht direkt der Objekterkennung, sondern werden zu der bereits erwähnten Hierarchie von immer komplexeren *Kompositionen* zusammengesetzt. Die Bildung von Kompositionen wird aufgrund von räumlicher Benachbarung getroffen: Zunächst werden in einem Bild 30 Deskriptoren zufällig ausgewählt. Diesen werden weitere Deskriptoren in einem von der Regionengröße abhängigen Umkreis von 60 bis 100 Pixeln zugeordnet. Da auf diese Art viele Kompositionen entstehen, müssen die relevanten ausgewählt werden. Die Autoren verdichten daher häufige Kompositionen mit Hilfe des K-Means-Verfahrens zu 1000 Prototypen. Der nächste Schritt besteht darin, aus diesen Prototypen die für Klassifikation von Objekten geeignetsten herauszusuchen. Dazu trainieren die Autoren Klassifikatoren auf Zwei-Klassen-Teilprobleme des Gesamtproblems. Diese Idee geht auf Roth und Tsuda [RT01] zurück, die zur Modellierung komplexer Klassengrenzen eine Menge von Einzelklassifikatoren vorschlagen, die den Klassenraum jeweils binär unterteilen. Die tatsächlichen Klassengrenzen können durch Mittelung über die Einzelklassifikatoren bestimmt werden, sofern diese nicht alle gleich sind. Ommer et al. erzeugen auf diese Weise Gruppen von Kompositionen, die jeweils auf die Erkennung bestimmter Untermengen aller Kategorien spezialisiert sind. Genauer gesagt, wird die Wahrscheinlichkeit

$$p_\Gamma = p(S = s|\Gamma)$$

der Einordnung einer Komposition Γ in die Klasse s maximiert, wobei S als Zufallsvariable aufgefaßt wird, die das Klassifikationsergebnis repräsentiert. Für die Objekterkennung ist es außerdem nützlich, wenn eine Komposition selektiv auf bestimmte Klassen reagiert. Ommer und Buhmann bewerten diese Eigenschaft anhand der Entropie

$$H(p_\Gamma) = -\sum_s p_\Gamma \log p_\Gamma$$

über die erkannten Klassen s. Dies erlaubt eine Unterteilung der 1000 Prototypen in 250 relevante Kompositionen und 750 irrelevante. Die besten Kompositionen werden ausgewählt und über die Modellierung der gegenseitigen geometrischen Abhängigkeiten zu einem gemeinsamen Modell vereint. In diesem Fall

sind daher die Kompositionen eher als visuelles Alphabet zu verstehen als das Codebook.

Erlernen der Teilebeziehungen

Wenn Kandidaten für Teilepositionen ermittelt wurden, stellt sich als nächstes die Frage, welche Kandidaten in das Modell aufgenommen werden, und wie die Beziehungen zwischen ihnen trainiert werden.

Weber [WWP00] und Fergus et al. [FPZ07] schlagen einen gierigen Algorithmus zur Auswahl von Teilen und Optimierung der Teilebeziehungen vor. Dazu wird ein initiales Modell mit wenigen Teilen und zufälliger Parametrisierung gewählt. Über einen Expectation Maximization Algorithmus [DLR77] bezüglich der Wahrscheinlichkeit in Gleichung 3.4 wird das Modell für die gewählte Anzahl an Teilen optimiert. Dazu wird jeweils ein Teil zufällig variiert und die beste Variation beibehalten. Anschließend wird ein weiteres Teil hinzugenommen und der Vorgang wiederholt bis sich keine Verbesserung mehr ergibt. Um eine frühe Konvergenz gegen ein suboptimales Maximum zu vermeiden, muß der Parameterbereich des initialen Modells sorgfältig ausgewählt werden. Da sich für Teilepositionen, die nicht auf den Objekten, sondern im Hintergrund liegen, geringe Erklärungsbeiträge für die Stichprobe ergeben, ist eine mühsame manuelle Segmentierung der Stichprobe nicht nötig. Die Autoren nennen ihre Methode folglich 'schwach überwacht' (org. *weakly supervised* [FPZ07]). In der beschriebenen Form wird nur ein Modell für eine Objektansicht trainiert. Fergus et al. merken dazu an, daß Objekte oft in einer charakteristischen Pose auftreten. Eine Lösung des Problems sei eine Mischung aus mehreren Modellen.

Abhängig vom jeweiligen Erkennungsverfahren kann es auch sinnvoll sein, die Position eines Teils innerhalb des Gesamtobjekts zu erlernen. Dadurch können bereits während der Erkennung der Teile Hypothesen über das Auftreten und die Position des Gesamtobjekts erzeugt werden. Selinger [Sel01] legt dazu eine Datenbank an, in der zu jedem Teil die Objektklasse und die Kameraperspektive des zugrundeliegenden Stichprobenelements gespeichert wird. Auch Mikolajczyk et al. [MLS06, MLS05, LMS06b] speichern die Objektklasse zusammen mit den Teilen ab. Da die Autoren Teile in einem Codebook speichern, ergibt sich eine zusätzliche Abstraktion. Für jedes Teil wird daher die Verteilung aller Objektklassen über der Position und der Skalierung relativ zu dem betreffenden Teil ausgezählt. Die Geometrieinformation wird dabei in Polarkoordinaten relativ zur Hauptrichtung eines Teils abgelegt und nicht in absoluten Bildkoordinaten. Dadurch wird das Objektmodell auf rechentechnisch günstige Weise rotationsinvariant. Da die Objekterkennung bei diesem Verfahren auf Mehrheitsentscheidungen beruht, müssen keine weiteren geometrischen Beziehungen trainiert werden.

Viola und Jones [VJ04] legen den Schwerpunkt vor allem auf die Geschwindigkeit. Zunächst stellen sie fest, daß die von ihnen eingesetzten Merkmalsextraktion 160 000 verschiedene Kandidaten für Teile eines Objekts ergibt. Aus diesen Teilen wird während der Trainingsphase eine Untermenge ausgewählt, was aufgrund der großen Datenmenge jedoch einen hohen rechentechnischen

Aufwand darstellt. Viola und Jones schlagen zur Lösung dieser Aufgabe eine Boosting-Methode vor. Zunächst wird zu jedem Merkmal eine einfache Klassifikationsfunktion

$$\mathfrak{f}(\boldsymbol{x}, f, \varpi, \vartheta) = \begin{cases} 1 & \text{wenn } \varpi f(\boldsymbol{x}) < \varpi\vartheta \\ 0 & \text{sonst} \end{cases}$$

erzeugt, welche die Werte Eins oder Null annimmt, je nachdem ob ein Merkmal an einer bestimmten Bildposition $\boldsymbol{x}$ erkannt wird oder nicht. Diese Funktion repräsentiert die Teile aus denen sich das Gesamtmodell zusammensetzt. Wie gut ein Teil in einem Stichprobenbild erkannt wird, hängt von der Art des Merkmals f, der Schwelle ϑ und einer Polarität ϖ ab. Zur Auswahl guter Teile setzen Viola und Jones eine Boosting-Methode ein. Dazu werden allen Elementen $\boldsymbol{x}$ der Stichprobe Gewichte w zugeordnet. Diese sind anfangs alle gleich und ergeben in der Summe Eins. Dann wird das Teil mit den Parametern f, ϖ, ϑ ausgewählt, daß den kleinsten Klassifikationsfehler

$$\varepsilon(f, \varpi, \vartheta) = \sum_i w_i |\mathfrak{f}(\boldsymbol{x}_i, f, \varpi, \vartheta) - s_i|$$

über der Stichprobe ergibt. Dabei bezeichnet s die Sollklasse eines Stichprobenbildes. Dem Teil wird ein Gewicht α zugeordnet, daß umso größer ist, je kleiner der Fehler ist. Das ausgewählte Teil mit dem Gewicht α wird in das Gesamtmodell aufgenommen. Als nächstes werden die Gewichte w der Stichprobe angepaßt. Gewichte von korrekt erkannten Stichprobenelementen werden verringert, da sie schon gelernt wurden und die 'schwierigeren' Stichprobenelemente das Lernen dominieren sollen. Die Gewichte werden wieder in ihrer Summe auf Eins normiert und das Verfahren setzt sich mit der Auswahl des nächsten fehlerminimierenden Teils fort. Wenn eine ausreichende Anzahl von Teilen gefunden wurde, bricht das Verfahren ab. Das Gesamtmodell besteht nun aus der Summe der jeweils mit α gewichteten Teile. Der generelle Vorteil von Boosting besteht darin, daß aus einer Menge von schwachen Klassifikatoren ein beliebig starker Klassifikator erzeugt werden kann, wenn nur genügend Terme zusammengestellt werden [FS99]. Viola und Jones geben als Hauptvorteil des Verfahrens die geringe Komplexität beim Training an, wodurch die Bearbeitung einer großen Zahl von Merkmalen möglich wird.

Die über den Trainingsverlauf zunehmende Gewichtung schwieriger Stichprobenelemente ist auch ein entscheidendes Merkmal des evolutionären Lernverfahrens von Stommel und Kuhnert [SK05, SK06]. Dieses erzeugt über eine bestimmte Anzahl von Trainingszyklen verschiedene Teilekonstellationen und bewertet diese bezüglich ihrer Eigenschaften zur Objekterkennung. Um das Problem der Varianz innerhalb einer Klasse zu vereinfachen, wird für jede Klasse eine Menge von Teilekonstellationen erzeugt, die jeweils auf einen Teilaspekt optimiert werden. In jedem Trainingszyklus wird das schlechteste Element durch ein zufälliges neues Element ersetzt, ein zufälliges Element wird durch Addition von normalverteiltem Rauschen mutiert und die Güte der Teilekonstellationen aktualisiert. Die Güte setzt sich aus der Selektivität bezüglich der gewünschten

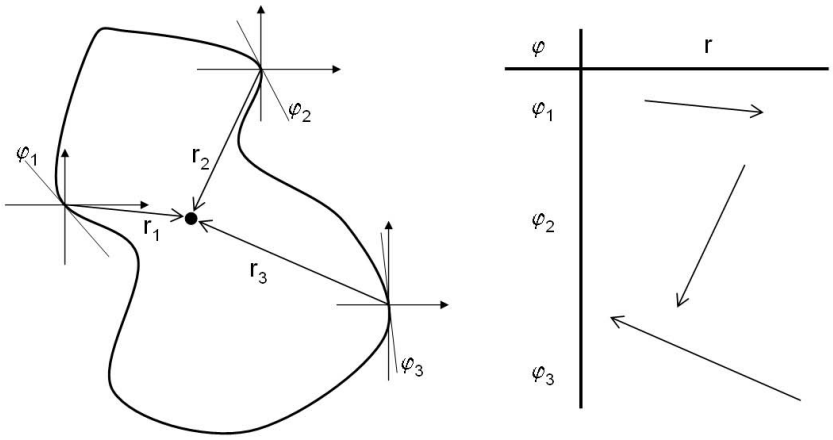


Abbildung 3.22: **Erzeugung der R-Tabelle bei der erweiterten Hough-Transformation [Bal81]:** Zu jeder Kantenrichtung φ wird der Vektor **r** zu dem Bezugspunkt des Objekts gespeichert.

Klasse und der Generalisierbarkeit über die Stichprobenelemente der Klasse zusammen. Um letztere zu berechnen, wird eine Abdeckungsmatrix bestimmt, die für jedes Stichprobenelement die Anzahl der übereinstimmenden Konstellationen angibt. In dieser Hinsicht hat eine Teilekonstellationen praktisch die Bedeutung eines Klassifikators zur Erkennung bestimmter Stichprobenelemente. Um zu vermeiden, daß sich die Menge der Konstellationen einseitig auf einfach zu erkennende Stichprobenelemente konzentriert, werden bei der Güteberechnung die einzelnen Stichprobenelemente quadratisch mit der Anzahl der erkennenden Konstellationen bewertet. Die Erkennung weniger schwieriger Stichprobenelemente ergibt auf diese Art höhere Bewertungen bezüglich der Verallgemeinerbarkeit als die Erkennung vieler einfacher Elemente. Dies erzwingt die nötige Kooperation der Konstellationen, um auch für Klassen mit starker Streuung eine breite Abdeckung zu erzielen.

3.2.3 Ablauf der Objekterkennung

Objekterkennungssysteme mit globalen Modellen können Objekte nur erkennen, wenn sie ein Bild vollständig ausfüllen. Zur Lokalisierung kleiner Objekte in größeren Bildern muß daher ein Fenster über das Bild geschoben werden. Die Bildausschnitte aller Fensterpositionen müssen separat klassifiziert werden. Wenn die Objektgröße nicht bekannt ist, muß zusätzlich entweder noch die Fenstergröße variiert werden oder eine Bildpyramide berechnet werden. Der zusätzliche Aufwand für die Lokalisation kann hoch sein.

Die häufigsten Methoden, um mit teilebasierten Modellen Objekte zu erkennen, sind Mehrheitsentscheidungen und Entscheidungsbäume. Bei Mehrheits-

entscheiden wird ein Objekt erkannt, wenn eine bestimmte Mindestzahl von Teilen erkannt wird. Wenn mehrere Objekte konkurrieren, wird das erkannt, von dem die meisten Teile gefunden wurden. Als zusätzliches Kriterium dient meistens noch die Auftrittswahrscheinlichkeit der Teile. Eine weitere Erkennungsmethode besteht natürlich in der Berechnung des Wahrscheinlichkeitsverhältnisses in Gleichung 3.3. Zur Lokalisation müssen dann noch alle Hypothesen geprüft werden, die zu dem erkannten Objekt gehören.

Mehrheitsentscheide

Die meisten auf Mehrheitsentscheiden basierenden Methoden können als Abwandlung oder Erweiterung der Hough-Transformation angesehen werden, die der klassische Vertreter dieser Kategorie ist. Der Vorteil der Hough-Transformation ist, daß die Lokalisation gleichzeitig mit der Klassifikation geschieht, da sich von der Position detektierter Teile direkt auf die Objektposition schließen läßt.

Ursprünglich ist die Hough-Transformation eine Methode zur Erkennung von Geraden. Sie nutzt aus, daß sich eine Gerade sowohl durch die Punkte auf der Geraden im Bild als auch durch die Parameter einer Geradengleichung beschreiben läßt. Die Erkennung startet zuerst mit einem lokalen Operator, der eine Menge kleiner Geradenstücke extrahiert, von denen einige Teil der gesuchten großen Geraden sind. Zu den kleinen Geradenstücken werden dann jeweils die Geradenparameter berechnet. Über diese wird anschließend ein Histogramm berechnet. Da die kleinen Geradenstücke, die Teil der großen Geraden sind, die gleichen Parameter besitzen wie die großen, ergibt sich für diese eine Häufung im Histogramm. Das Histogramm wird daher auch *Akkumulator* genannt. Die lokalen Maxima des Histogramms parametrisieren also größere Geraden im Bild.

Ballard weitet die Methode auf die Erkennung beliebiger, nichtparametrischer Konturen aus [Bal81, BB82]. Als Parameterraum dienen die Koordinaten möglicher Bezugspunkte der Konturen. Eine zu erkennende Kontur wird zunächst als Liste von Bildpunkten beschrieben, welche die Teile des Objekts darstellen. Von entscheidender Bedeutung für die Qualität des Verfahrens ist, daß zu jedem Punkt die Richtung der Tangente vorliegt, so daß sich verschiedene Konturpunkte in gewissem Maße unterscheiden lassen. In der Trainingsphase wird ein beliebiger Bezugspunkt für die Kontur festgelegt und eine sog. R-Tabelle ('table of edge-orientation reference-point correspondence termed an R-table', [Bal81]) angelegt. Die R-Tabelle ordnet den Gradientenrichtungen der Konturpunkte die Vektoren zum Bezugspunkt zu (s. Abb. 3.22). Da in den meisten Konturen Tangentenrichtungen mehrfach auftreten, können in die R-Tabelle mehrere Vektoren zu einer Richtung eingetragen werden.

Die Erkennung von Konturen in einem unbekannten Bild geschieht über die Rekonstruktion der Bezugspunkte aus Konturstücken. Diese müssen in einem Vorverarbeitungsschritt extrahiert werden. Zu jedem gefundenen Konturstück wird eine Liste die Richtung der Tangente in der R-Tabelle nachgeschlagen. Es ergibt sich zu jedem Konturstück eine Liste mit Relativpositionen möglicher Bezugspunkte. Durch Addition der Relativpositionen auf die Positionen der Kon-

turstücke ergeben sich mögliche Absolutpositionen von Bezugspunkten. Jede solche Position stellt eine Hypothese über das Vorhandensein eines Objektes dar. Sie werden in den Akkumulator eingetragen, was gleichbedeutend mit der Berechnung eines Histogramms ist. Wenn mehrere Konturstücke im richtigen geometrischen Verhältnis zueinander stehen, häufen sich die Koordinaten der zugehörigen Bezugspunkte. Die Objekterkennung kann wieder über die Suche nach Maxima im Akkumulator geschehen.

Für die Erkennung eines Objekts ist es nicht immer erforderlich, daß alle Punkte eines Testbildes in der R-Tabelle nachgeschlagen werden. Oft läßt sich eine Geschwindigkeitssteigerung erzielen, indem eine zufällige Untermenge der Bildpunkte nachgeschlagen wird. Diese Optimierung ist unter dem Begriff *Probabilistic Hough Transform* bekannt. Shaked, Yaron und Kiryati [SYK94] untersuchen, nach welchen Kriterien die Anzahl der auszuwertenden Punkte bestimmt werden kann. Die einfachste Methode ist eine feste Begrenzung der Punktanzahl. Als besser erweist es sich, eine feste Obergrenze für den Wert der maximalen Häufung einzuführen. Von der Qualität her zwischen diesen Methoden liegt die Überprüfung der Rangordnung der lokalen Häufungen. Wenn diese stabil bleibt, müssen keine weiteren Punkte mehr nachgeschlagen werden. Shaked, Yaron und Kiryati bemerken einen theoretischen Zusammenhang zwischen dem Abbruchproblem und der Sequenzanalyse nach Wald [Wal47]. Mit diesem Hintergrund führen sie zwei neue Kriterien ein, die auf dem Vergleich der beiden höchsten Häufungen beruhen und zu einer nochmals niedrigeren Fehlerrate führen.

Bezüglich der Leistungsfähigkeit warnt Ballard, daß die R-Tabelle leicht praxistaugliche Größen übersteigt. Grimson und Huttenlocher [GH88] stellen ebenfalls Überlegungen und Experimente zur Stabilität der erweiterten Hough-Transformation an und kommen zu dem Ergebnis, daß die Wahrscheinlichkeit falscher Treffer während der Objekterkennung sehr hoch sein kann. Ein direkter Grund dafür ist die hohe Anzahl an Zellen im Akkumulator, auf die das in der R-Tabelle gespeicherte Modell abbilden kann. Da für alle Konturstücke des Eingabebildes separat größere Mengen an Hypothesen erzeugt werden, können sich an den Schnittmengen falsche Maxima aufbauen. Da alle erzeugten Hypothesen im Akkumulator aufsummiert werden, hängt die Wahrscheinlichkeit falscher Treffer direkt von dem Verhältnis der erzeugten Hypothesen zur Anzahl der Akkumulatorzellen ab. Die Fehlerwahrscheinlichkeit wird daher erhöht durch

- eine grobe Quantisierung des Parameterraums, da dies den Akkumulator verkleinert. Dieser Effekt wird nicht dadurch aufgewogen, daß mehr Hypothesen in eine Zelle fallen.

- Mehrdeutigkeiten im Modell, z.B. Linien: Ein einzelnes Konturstück bildet komplette Linien in den Akkumulator ab. Befinden sich zusätzlich noch Linien im Bild, wird die komplette Linie aus dem Modell einmal für jedes Teilstück einer Linie im Bild in den Akkumulator übertragen und häuft sich dort.

- Rauschen, da es eine grobe Quantisierung des Parameterraums erzwingt.
- Vorverarbeitungsschritte, da sie Ungenauigkeiten einführen und daher wie Rauschen wirken.
- freie Parameter, z.B. die Rotation von Objekten, da sie die Anzahl möglicher Hypothesen vervielfachen.
- niederdimensionale Akkumulatoren, da sie weniger Zellen enthalten.

Grimson und Huttenlocher untersuchen auch, wie die Hypothesen im Akkumulator verteilt sind, insbesondere inwiefern falsche Hypothesen zu Häufungen tendieren. Unter der optimistischen Annahme, daß die Hypothesen über dem Akkumulatorraum gleichverteilt und statistisch unabhängig sind, beträgt die Wahrscheinlichkeit p, daß eine erzeugte Hypothese in einer bestimmten von n_C Akkumulatorzellen auftritt,

$$p = \frac{1}{n_C}.$$

Die Wahrscheinlichkeit, daß genau k Hypothesen in einer Zelle auftreten, wird durch die Binomialverteilung

$$p_k = \binom{n_{\mathfrak{H}}}{k} p^k (1-p)^{n_C - k}$$

beschrieben, wobei $n_{\mathfrak{H}}$ die Anzahl der Hypothesen ist. Die Wahrscheinlichkeit, daß eine Zelle mehr als l Hypothesen enthält, wird dann gemäß

$$p_{\geq l} = 1 - \sum_{k=0}^{l-1} p_k$$

berechnet. Grimson und Huttenlocher kommen anhand von Beispielrechnungen zu dem Ergebnis, daß unter realistischen Bedingungen schnell viele Akkumulatorzellen mehr Hypothesen enthalten als das Modell Teile. Gute Ergebnisse ergeben sich nur für Eingabebilder mit wenigen Konturstücken und Modelle aus vielen eindeutigen Teilen. Die Autoren schließen, daß die Methode schlecht mit der Komplexität der Eingabebilder skaliert und nur für Situationen geeignet ist, in denen der Anteil korrekter Daten nicht zu klein ist im Vergleich zur Menge inkorrekter Daten.

Selinger [Sel01] führt die Hough-Transformation auf größeren Objektteilen durch. Da diese jeweils aus mehreren benachbarten Kantenzügen bestehen, enthalten sie viel mehr Information als die Kantenstücke der originalen Hough-Transformation, die sich nur auf wenige Bildpunkte beziehen. Selinger erreicht dadurch eine viel höhere Selektivität bei der Erkennung des gesamten Objekts. Außerdem führt sie eine Reihe von eher heuristischen Optimierungen ein, die das Signal/Rausch-Verhältnis verbessern. Zunächst werden ähnlich einem bayesischen Ansatz die Hypothesen mit ihren Auftrittshäufigkeiten gewichtet. Experimentell hat sich als zweites eine systematische Bevorzugung von Hypothesen

von Objekten mit wenigen Komponenten als vorteilhaft herausgestellt. Dies kann als Suche nach einer möglichst einfachen Bildbeschreibung gedeutet werden. Als dritte Strategie wird eine in maßvollem Umfang stärkere Gewichtung größerer Komponenten im Vergleich zu kleineren angegeben. Der Erfolg dieser dritten Strategie wird mit einigen Zweifeln auf eine möglicherweise größere Stabilität großer Teile zurückgeführt. Mit diesen Verbesserungen erreicht Selinger trotz der von Grimson und Huttenlocher genannten Einwände gute Erkennungsraten für eine Datenbank mit starren Objekten, sowohl vor schwarzem als auch vor strukturiertem Hintergrund.

Die Objekterkennung läßt sich noch zielgerichteter durchführen, wenn die genauen Wahrscheinlichkeiten für das Auftreten von Objekten berücksichtigt werden. Mikolajczyk, Leibe und Schiele [MLS06] ermitteln diese während der Erzeugung des Codebooks in Trainingsphase. Die Objekterkennung geschieht dann durch den Vergleich der Merkmale eines Testbildes mit dem Codebook. Um die kombinatorische Vielfalt zu reduzieren, wird bei dem Codebook auf eine Baumstruktur mit Ähnlichkeitsinformationen aus der Trainingsphase zurückgegriffen. Die Merkmale des Testbildes werden auf die gleiche Weise in einer Baumstruktur angeordnet, sodaß anstelle einer erschöpfenden Suche nach korrespondierenden Merkmalen nur die Ähnlichkeitsgraphen in Übereinstimmung gebracht werden müssen. Aus den übereinstimmenden Einträgen des Codebooks werden nun Wahrscheinlichkeitsfunktionen für alle Klassen bestimmt. Dies geschieht, indem die Wahrscheinlichkeitsverteilungen, die im Codebook gespeichert sind, passend zu den Merkmalen aus dem Testbild rotiert und skaliert werden. Die Objekterkennung läuft dann über eine Suche nach lokalen Maxima.

Entscheidungsbäume

Zur Objekterkennung mit teilebasierten Modellen kommen neben Mehrheitsentscheiden auch alle Techniken des Maschinenlernens in Frage. Serre et al. setzen dazu beispielsweise Support Vector Machines [CV95] ein. Aufgrund ihrer Geschwindigkeit, möglicherweise aber auch aufgrund des ähnlichen Aufbaus zu den Modellen, werden häufig Entscheidungsbäume eingesetzt.

Die Objekterkennung verläuft bei Burge und Burger [BB97, BBM96] über das sequentielle Erkennen der Einzelteile. Dies entspricht einem zu einer Kette degenerierten Entscheidungsbaum, wobei an jedem Knoten ein Objekt entweder verworfen oder zur weiteren Untersuchung an den nächsten Knoten weitergereicht wird. Wenn der letzte Knoten erkannt wird, ist das gesamte Objekt erkannt. Der Vorteil des Modell von Burge und Burger besteht darin, daß mehrfache Vernetzungen zwischen den Teilen möglich sind. Daher haben einzelne fehlende Teile nur eine geringe Auswirkung auf die relativen Teilebeziehungen. Damit einzelne fehlende Teile die Entscheidungskette nicht unterbrechen und diesen Vorteil zunichte machen, werden vor der Objekterkennung erschöpfend nicht zyklische Pfade durch das Modell ermittelt. Diese werden zu einem Entscheidungsbaum zusammengesetzt, der eine Liste von Teilen eines unklassifizierten Bildes mit dem erlernten Modell vergleicht.

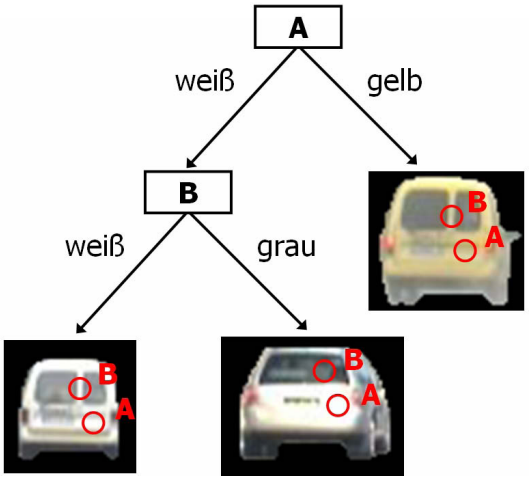


Abbildung 3.23: **Entscheidungsbaum zur Erkennung von Autos nach dem Ansatz von Stommel und Kuhnert [SK06]:** Das Modell besteht aus zwei Merkmalen A und B. An jedem Knoten des Entscheidungsbaum wird ein Merkmal überprüft und entsprechend der Ausprägung des untersuchten Objekts verzweigt. An den Blattknoten des Baums findet eine Zuordnung in eine Klasse statt. Im Beispiel hier würde ein unbekanntes Objekt mit den Merkmalen A = weiß und B = grau dem mittleren Auto zugeordnet.

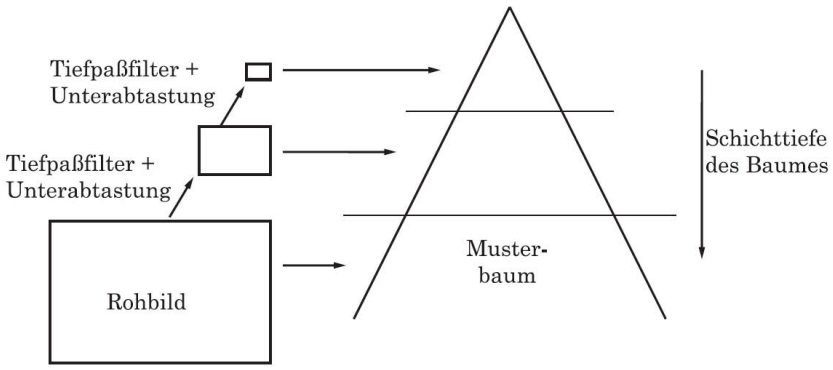


Abbildung 3.24: **Entscheidungsbaumbasierte Objekterkennung auf einer Bildpyramide [Ett06]:** Die ersten Knoten des Entscheidungsbaums beziehen sich auf ein stark unterabgetastetes Testbild. Mit steigender Tiefe im Baum werden detailliertere Mustervergleiche auf höher auflösenden Darstellungen des Testbildes durchgeführt.

Stommel und Kuhnert [SK05, SK06] wandeln ihr Teilemodell vor allem deshalb in einen Entscheidungsbaum um, um die große Anzahl an Merkmalsvergleichen zu verringern, die eine volle Korrelation mit den Testbildern mit sich bringen würde. Dazu werden alle Teile einer Gesamtansicht eines Objekts wie bei Viola und Jones als Entscheidungssequenz aufgefaßt, die einen Pfad von der Wurzel des Entscheidungsbaums bis zu den Blattknoten angibt (s. Abb. 3.23). Durch die Identifizierung gleicher Teile über verschiedene Ansichten von Objekten können die entsprechenden Knoten der Entscheidungssequenzen zusammengefaßt werden, wodurch sich eine Baumstruktur ergibt. Die Autoren orientieren sich hier vage an Quinlans C4.5 Algorithmus [Qui93]. Zwei Teile können zusammengefaßt werden, wenn sie die gleiche Position haben und die gleiche Merkmalsart haben, d.h. eine Kante oder Fläche an der gleichen Position bezeichnen. Die Teile dürfen sich in der Ausprägung des Merkmals unterscheiden, beispielsweise der Helligkeit. Die Merkmalsausprägung entscheidet über die Verzweigung zu den weiteren Knoten des Entscheidungsbaums. Bei der Erzeugung von Entscheidungsknoten und auch bei der späteren Klassifikation von Testbilder tritt jedoch häufig der Fall auf, daß an einem Knoten eine Entscheidung getroffen werden muß, für die keine Informationen vorliegen, da in dem Testbild das entsprechende Merkmal fehlt. In diesem Fall muß nacheinander in alle Kindknoten verzweigt werden. Um den Entscheidungsbaum robust gegen Rauschen zu machen, wird zudem eine gewisse Toleranz um die jeweiligen Merkmalsausprägungen festgelegt, die ebenfalls dazu führen kann, daß zu mehreren Kindknoten verzweigt werden muß. Um die Anzahl falsch Positiver zu reduzieren, werden unselektive Teilekonstellationen nicht in den Entscheidungsbaum eingefügt. Die genannten Maßnahmen verringern die Anzahl der zur Objekterkennung nötigen Teilevergleiche etwa um den Faktor vier. Die Objekterkennung erfolgt nun so, daß alle Positionen eines Testbildes nacheinander mit dem Entscheidungsbaum klassifiziert werden. Aufgrund möglicher Mehrfachverzweigungen bei unbekannten Merkmalen schwankt die zur Erkennung nötige Zeit über die Bildpositionen. Die Dauer der Objekterkennung hängt damit stark vom Bildinhalt ab. Bei den zum Test eingesetzten Straßenszenen haben sich vor allem die Straßen selbst als ungünstig herausgestellt. Da sie meistens gleichmäßig grau sind, bieten sie wenige Anhaltspunkte, um Klassifikationsentscheidungen zu treffen. Dadurch müssen viele Knoten des Entscheidungsbaums durchlaufen werden. Zum anderen machen sie große Teile des Bildes aus und haben damit einen großen Einfluß auf die Gesamtdauer.

Um schnell viele falsche Hypothesen zu verwerfen, erzeugen Viola und Jones [VJ04] zu ihrem Teilemodell einen kaskadierten Klassifikator, der wie bei Burge und Burger die Form eines zu einer Kette degenerierten Entscheidungsbaums hat. Die ersten Klassifikatoren in der Kette sind schnell, aber ungenau. Sie sind darauf spezialisiert, ohne großen Rechenaufwand viele falsche Treffer auszusortieren. Dabei müssen jedoch möglichst viele richtige Treffer zur nächsten Stufe weitergeleitet werden, da einmal zurückgewiesene Testbilder nicht weiter bearbeitet werden. Die hinteren Stufen müssen aufgrund der Vorauswahl der ersten Stufen nur noch eine geringe Anzahl an Bildern untersuchen. Sie enthalten daher aufwendigere und genauere Klassifikatoren. Die Klassifikatoren der Kaskade

werden jeweils durch Boosting erzeugt. Dabei werden solange neue Teile in den Klassifikator der jeweiligen Kaskade eingefügt, bis die gewünschten Raten für falsch Positive erreicht werden. Viola und Jones geben dazu an, daß Boosting eigentlich dazu dient, den Gesamtfehler der Erkennung zu minimieren, dabei aber keine Optimierung auf die Rate der falsch Positiven vornimmt. Die Autoren erzwingen dieses Verhalten, indem sie die Schwellen zur Erkennung absichtlich hoch einstellen. Und obwohl die Methode in der Praxis funktioniert, geben die Autoren an, daß noch unklar ist, wie sich dieser Eingriff auf die theoretische Leistungsfähigkeit des Boostings auswirkt.

Auch Ettelt [Ett06] versucht den hohen Aufwand zu verringern, den die Klassifikation aller Bildpositionen mit einem Entscheidungsbaum bedeutet. Er merkt dazu an, daß der Aufwand zunächst einmal von der Kameraauflösung abhängt, jedoch nicht unbedingt von dem Problem selbst. Er schlägt daher vor, als erstes die Auflösung auf das tatsächlich benötigte Maß zu reduzieren. Weiterhin stellt er fest, daß in den ersten Schichten des Baums noch keine hohe Genauigkeit nötig ist und dort ein grober Mustervergleich ausreicht. Eine höhere Genauigkeit sei erst in tieferen Stufen des Entscheidungsbaums nötig. Ettelt führt daher die Objekterkennung auf einer Bildpyramide durch (s. Abb. 3.24), was eine drastische Geschwindigkeitssteigerung bewirkt.

3.3 Fazit und Zielsetzung

Aktuelle Arbeiten behandeln die Modellierung von Objekten anhand von Objektteilen mit geometrischen Beziehungen untereinander [Sel01, WWP00, MLS06]. Die Gründe dafür sind sowohl technischer als auch ideeller Natur. Zum einen ergibt ein teilebasierter Ansatz eine gewisse Toleranz gegenüber Verdeckungen, Störungen und fehlenden Teile. Zum anderen ermöglicht eine lose Modellierung der Teilebeziehungen auch in gewissem Umfang die Behandlung von Objektdeformationen. Dabei sind immer wieder Einflüsse aus der Neurophysiologie sichtbar. Dies gilt am stärksten für die vielschichtige Teilehierarchie von Serre, Wolf und Poggio [SOP07, SWP05a, SWP05b, RP02], die fast ausschließlich am hierarchischen Aufbau des visuellen Kortex orientert ist und dabei gute Ergebnisse liefert. Ommer und Buhmann [OB06, OSB06] nennen zwei Hauptvorteile solch komplexer Teilehierarchien. Als erstes können allgemeingültige Teile in verschiedenen Objektzusammenhängen genutzt werden, was auch für komplexe Objektkategorien kompakte Modelle erlaubt. Als zweites schafft eine kompositionelle Hierarchie durch ihre Zwischenschichten die Verbindung von leicht berechenbaren, aber unspezifischen lokalen Merkmalen auf der einen Seite hin zu bedeutsamen, aber komplexen Objektkategorien auf der anderen Seite. Dies bedeutet einen Vorteil bei der Modellerzeugung, da zum einen Modelle für einfachere Teilkategorien erzeugt werden können, und die Modelle zum anderen gezielt auf die jeweilige Streuung innerhalb der Teilkategorien angepaßt werden können.

Derzeit behandeln jedoch die meisten Objekterkennungssysteme nur eine oder zwei Abstraktionsstufen [Sel01, CFH05, FPZ03]. Üblicherweise werden da-

bei operatorbasiert Teile ausgewählt und durch ein Geometriemodell zu einer einzigen Objektansicht [FPZ03] miteinander verbunden. Zum Aufbau einer Teilehierarchie liegen daher abgesehen von den Ergebnissen von Serre et al. [SOP07] und Ommer et al. [OB06] kaum Erkenntnisse vor. Immerhin haben Crandall et al. [CFH05] sowie Fergus et al. [FPZ06] den theoretischen Aufwand für die Arbeit mit verschieden stark vernetzten Teilemodellen geklärt und den praktischen Nutzen verglichen [CH06].

Große Unterschiede ergeben sich bei der Modellierung der Objektgeometrie. Während Serre et al. [SOP07] diese komplett vernachlässigen, trainieren Fergus et al. [FPZ03] relative Positionen und Skalierungen zwischen allen im Modell vorliegenden Teilepaaren. Da sich beide Modelle bezüglich der Anzahl an Abstraktionsebenen unterscheiden, ist ein direkter Vergleich schwierig. Aufgrund von neurowissenschaftlichen Erkenntnissen wäre in diesem Zusammenhang eine mit steigender Hierarchiestufe abnehmende räumliche Genauigkeit zu erwarten, die der entlang des ventralen Pfades beobachteten Vergrößerung der rezeptiven Felder der Nervenzellen entspricht. Bezüglich des Trainings ist festzustellen, daß die hohe Parameterzahl der Modelle die Suche nach einer optimalen Lösung erschwert. Es werden daher mehr oder weniger gierige heuristische Verfahren eingesetzt, beispielsweise Expectation Maximisation [FPZ06], Genetische Algorithmen [SK05] oder Boosting [VJ04]. Die Objekterkennung geschieht meistens über Mehrheitsentscheide, wobei gewisse theoretische Schwierigkeiten überbrückt werden müssen, oder über Verfahren aus dem Maschinenlernen.

Das Ziel der vorliegenden Arbeit ist nun, den hierarchischen Ansatz zur Objekterkennung genauer zu untersuchen und insbesondere folgende Fragestellungen zu untersuchen:

- Wodurch definieren sich Teile?
- Welche Faktoren sind für das Training ausschlaggebend?
- Wie werden mehrere Ansichten modelliert?

Der erste Punkt betrifft vor allem die Geometrie. Häufig werden Teile als Deskriptor modelliert, der keine feinere Zusammensetzung eines lokalen Bildbereichs mehr erkennen läßt. Dies ist eine akzeptable Methode, da für viele Deskriptoren gute Ergebnisse veröffentlicht wurden. Hier soll jedoch ein anderer Ansatz untersucht werden, um mehr über die Teilehierarchie zu erfahren. Daher wird ein Teil als Menge einfacherer Teile beschreiben. Die statistische Auswertung der geometrischen Abhängigkeiten zwischen den Teilen soll tiefere Erkenntnisse darüber liefern, was ein Teil ist und wie man es modellieren und trainieren kann. Dabei soll auch die Frage geklärt werden, welche geometrischen Toleranzen auf verschiedenen Abstraktionsebenen sinnvoll sind. Insbesondere betrifft dies Abwägungen zwischen der Geometriemodellierung oder einem Bag of Features.

Dies schließt auch das Training mit ein. Die genannten heuristischen Verfahren liefern zwar erwiesenermaßen gute Resultate, arbeiten allerdings mehr oder weniger als Black Box. Es wird daher nicht klar, aufgrund welcher Parameteroptimierungen das Ergebnis erzielt wurde und in welche Richtungen man die

Trainingsmethoden verbessern kann. In dieser Arbeit werden daher die Parameterräume innerhalb sinnvoller Bereiche erschöpfend untersucht, bzw. unterabgetastet, wenn dies rechentechnisch zweckmäßiger ist. Aus den Ergebnissen werden dann Regeln zur Parametrisierung des Modells abgeleitet, die nicht nur ein gutes Modell liefern, sondern dieses auch erklären.

Der letzte Punkt bezieht sich darauf, daß aktuelle Modelle meistens nur eine Ansicht optimieren. Es ist daher zu klären, wie eine Stichprobe zweckmäßig durch Ansichten abgedeckt werden kann und wie die einzelnen Ansichten modelliert werden.

Der in vorliegenden Arbeit beschriebene Ansatz basiert dabei unter anderem auf den folgenden Techniken:

- Erzeugung eines visuellen Alphabets allgemeingültiger Teile
- Einsatz der Merkmalsextraktion von Stommel und Kuhnert [SK05, SK06]
- Erweiterung der Merkmale um Punkte auf Skelettlinien
- Modellierung sternförmiger Teile-Beziehungen
- Mehrheitsentscheide zur Objekterkennung
- Modellierung mehrer Objektansichten durch einzelne Modelle

Der erste Punkt ist ein Merkmal biologischer Systeme und wird zudem in vielen aktuellen Ansätzen erfolgreich eingesetzt. Eine allgemeingültige Teilemodellierung ist die Grundlage für die Wiederverwendbarkeit von Teilen zur Beschreibung verschiedener Objekte und damit für ein kompaktes Modell. Darüberhinaus ist Allgemeingültigkeit die Grundlage zur Abstraktion von konkreten Bildern und damit auch zur Arbeit mit einem hierarchischem Modell.

Als einfachstes Teil auf der ersten Abstraktionsebene über dem Bild selbst dienen die von Stommel und Kuhnert [SK05, SK06] extrahierten Merkmalspunkte. Diese sind zwar einfacher aufgebaut als aktuelle Regionendetektoren [MLS05], lassen sich aber dichter zu komplexeren Teilen zusammensetzen. Dabei sind sie aufgrund ihrer niedrigen Dimensionalität leicht handhabbar und bieten eine Kombination verschiedener Merkmalsarten, was einer zu starken Anpassung an das verwendete Bildmaterial vorbeugt [LS03, BB97, BBM96]. Nicht zuletzt haben sie sich bereits schon zur Analyse realer Bilder bewährt [SK05, SK06], was bedeutet, daß die hier vorgestellten Ergebnisse nicht nur für die Cartoon-Datenbank gelten. Anstelle der reinen Intensitätsmerkmale werden hier allerdings für Flächen nur Punkte auf den Skelettlinien des Bildes ausgewählt, da sich diese prägnanter beschreiben lassen.

Teilebeziehungen werden aufgrund der geringen Komplexität nach dem Sternmodell modelliert, welches bereits sehr leistungsfähig ist [CH06]. Um dem von Fergus et al. [FPZ06] erwähnten Problem zu entgehen, daß ein fehlendes Referenzteil die Geometriemodellierung zunichte macht, wird für jede Teilekonstellation ein abstrakter Bezugspunkt eingeführt. Dieser ist für jedes erkannte Teil immer vorhanden und muß nicht detektiert werden. Aufgrund dieser Vorgehensweise erscheint eine Teileerkennung aufgrund von Mehrheitsentscheidungen

als vorteilhaft. Dabei akkumulieren erkannte Teile an der Position des jeweiligen Bezugspunktes Hinweise für das Vorhandensein des übergeordneten Teils. Hier gelten natürlich die von Grimson und Huttenlocher [GH88] gemachten Einschränkungen für die Hough-Transformation. Es existieren jedoch positive Beispiele [Sel01, MLS06], wo bei ausreichend vielen und komplexen Teilen eine zuverlässige Objekterkennung möglich ist. Mehrheitsentscheide erscheinen auch in Hinblick auf Bildszenen mit großen, nicht aufschlußreichen Flächen vorteilhafter als die sonst auch häufig eingesetzten Entscheidungsbäume, da diese hier recht lange brauchen.

Um die starken Variationen der Objekterscheinungen zu behandeln, die durch Verformungen des Objekts und perspektivische Transformationen entstehen, sollen mehrere prototypische Objektansichten identifiziert und separat modelliert werden. Die Prinzipien Aufzählen und Speichern werden durch ein weiteres visuelles Alphabet auf einer höheren Abstraktionsebene umsetzt. Der Ansatz ist teilweise motiviert durch die von Miyashita et al. [MSHM91, Miy93] und Nielsen et al. [NLR08] beschriebene neurophysiologische Mechanismen. Miyashita et al. zeigen, daß das Gedächtnis Zellen im inferioren Temporallappen des Affen beeinflußt. Die Autoren trainierten Zellen in diesem Teil des Gehirns auf willkürliche Muster und folgern, daß dies genutzt werden könne, um verschiedene Ansichten eines Objekts zu lernen. Nielsen et al. berichten zudem, daß Affen (offenbar im Gegensatz zu Menschen) rotierte Objekté offenbar durch jeweils verschiedene Mengen von visuellen Merkmalen für die jeweiligen Rotationswinkel repräsentieren.

In den folgenden Kapiteln wird das Modell und die Erkennungsmethode detailliert vorgestellt. Anschließend wird die Leistungsfähigkeit der entwickelten Methoden demonstriert, indem das Modell auf die Erkennung einer Cartoon-Figur trainiert wird und diese dann in ungelernten Bildern einer Cartoon-Datenbank gefunden wird. Anschließend wird das Verfahren mit einer weiteren Methode verglichen, die speziell auf die Cartoon-Vorlage zugeschnitten ist.

Kapitel 4

Modellierung von Objekten als Teile-Graph

Als Konsequenz zu den in der Literaturdurchsicht herausgestellten offenen Fragen und interessanten Detaillösungen wird ein hierarchisches Modell mit einer Baum- oder Waldstruktur vorgeschlagen.

Als erstes wird die Verbindungsstruktur des Modells definiert. Die Knoten des Modells entsprechen bestimmten Teilen eines Objekts. Verweise von einem Knoten auf andere legen fest, daß der Knoten die anderen als Unterstruktur enthält.

Nach der Definition der Verbindungsstruktur des Baums werden die geometrischen Abhängigkeiten zwischen den Knoten des Baums modelliert. Dazu werden für jeden Knoten die Relativpositionen der Kindknoten gespeichert. Mit zusätzlichen Geometrieparametern ähneln die inneren Knoten des Baums dem Constellation-Modell in der sternförmigen Variante. Blattknoten des Modells stellen lokale Merkmale dar.

Als letztes werden die Knoten mit Hilfe von Schlüsseln gekennzeichnet, was die Navigation in der Baumstruktur während der Objekterkennung ermöglicht.

Die Hauptmerkmale des Modells auf einen Blick sind:

- Mehrfachverwendung modellierter Muster
- Für jeden Knoten einstellbare Abhängigkeit von der Teilegeometrie
- Analyse des Objekterkennungsproblems auf verschiedenen Abstraktionsebenen
- Modellierbarkeit vieler Objektansichten gleichzeitig
- Unabhängigkeit von bestimmten Merkmalsdetektoren
- Anschaulichkeit, insbesondere Einsicht in die erlernten Zusammenhänge
- Robustheit gegenüber Verdeckungen, Störungen und geometrischen Verzerrungen

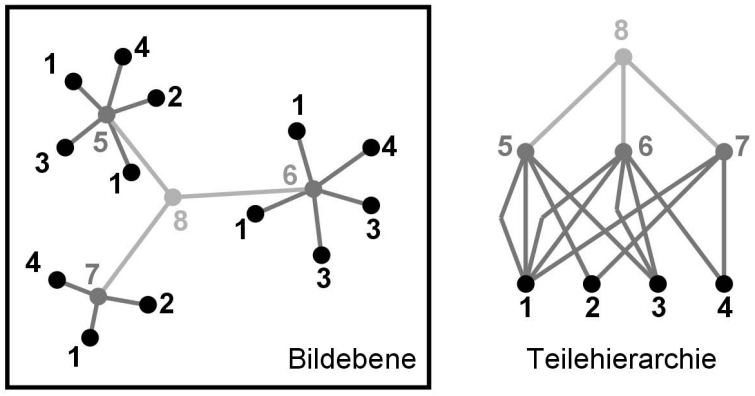


Abbildung 4.1: **Geometrische Teilebeziehungen innerhalb des hierarchischen Teilemodells.** Links: Auf der niedrigsten Ebene wird ein Objekt durch 13 lokale Merkmale in vier unterschiedlichen Ausprägungen beschrieben. Diese setzen sich zu drei abstrakteren Teilen zusammen, welche schließlich durch einen einzelnen Knoten repräsentiert werden. Rechts: Gleiche Ausprägungen von Knoten werden aus Platzgründen nur einfach gespeichert. Durch mehrfache Verweise mit entsprechenden geometrischen Attributen bleibt die Information jedoch erhalten.

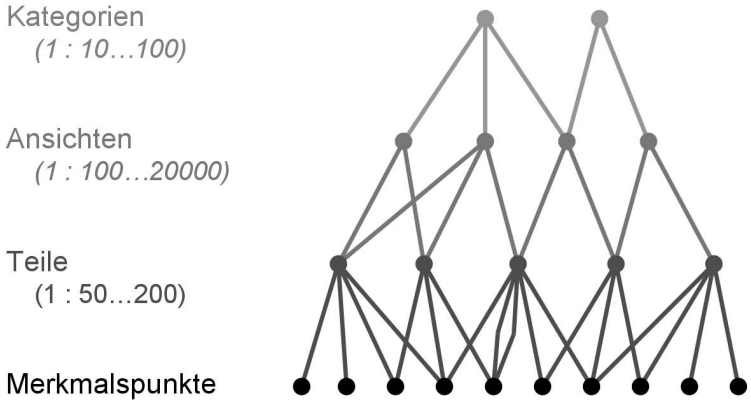


Abbildung 4.2: **Hierarchisches Teilemodell.** Jeder Abstraktionsebene kommt eine bestimmte Aufgabe zu. Die Zahlen in Klammern geben an, wieviele Knoten auf einer niedrigeren Ebene zu einem abstrakten Knoten zusammenlaufen.

- Gleichzeitige Detektion und Lokalisation

Das Modell wird genauer als Menge von Teilen

$$\{t_1, t_2, t_3, \dots\}$$

beschrieben, welche den Knoten der Baumstruktur entsprechen. Die Teileindices werden nicht nur eingesetzt, um verschiedene Teile zu unterscheiden, sondern vor allem, um Verweise zwischen Eltern- und Kindknoten anzugeben. Ein Knoten

$$t = \{m, U, L, \varsigma, \vartheta, \kappa\}$$

enthält dazu die Indices

$$U = \{u_1, u_2, u_3, \dots, u_m\}$$

der m Unterknoten auf einer niedrigeren Abstraktionsebene.

Die Wurzelknoten des Baums repräsentieren die zu erkennenden Objekte. Innere Knoten repräsentieren Teile von Objekten, Blattknoten repräsentieren nicht weiter teilbare lokale Bildmerkmale.

Ausdrücklich erlaubt sind die Fälle, daß mehrere Knoten den selben Knoten als Unterknoten haben, oder daß ein Knoten mehrfach als Unterknoten eines einzelnen anderen Knotens aufgeführt wird, im letzteren Fall zweckmäßigerweise in verschiedenen geometrischen Beziehungen. Solche Mehrfachverbindungen zwischen Knoten verschiedener Abstraktionsebenen erlauben die Verwendung einmal modellierter Muster in verschiedenen Kontexten. Dies ermöglicht das Erlernen eines visuellen Alphabets [Tan96], welches eine Abstraktion und Vereinfachung der Vielfalt auftretender lokaler Muster hin zu einer Menge universeller und aufgabenunspezifischer Teile darstellt. Diese Muster stellen eine Parallele zu den Eigenvektoren von Murase und Nayars [MN95] Eigenraum dar und eignen sich daher für eine vereinfachte Objektdarstellung. Dies entspricht auch der Idee einer spärlichen Objektkodierung im inferioren Temporallappen nach Rolls et al. [RTT97, FRAJ07].

Auf den Unterknoten wird eine sternförmige geometrische Abhängigkeit modelliert. Die Rolle des Referenzteils übernimmt dabei der Knoten t selbst. Die Positionen der Unterknoten hängen also alle von t ab, nicht jedoch voneinander. Die Positionen

$$L = \{x_1, y_1, x_2, y_2, \dots, x_m, y_m\}$$

der Unterknoten werden dabei relativ zur Position von t angegeben. Diese muß nicht selbst gespeichert werden, da sie sich während der Objekterkennung aus den Positionen der lokalen Merkmale auf der niedrigsten Ebene ergibt. Die Position eines Knotens relativ zu den detektierten Unterknoten ist frei wählbar und wird auf den Schwerpunkt der Unterknoten festgelegt. Die Objekterkennung verläuft dementsprechend in aufsteigender Richtung von der Merkmalsebene zu den abstrakteren Knoten. Detektion und Lokalisation sind abgeschlossen, wenn die Wurzelknoten des Modells erreicht sind. Im Gegensatz zum Constellation-Modell nach Fergus et al. [FPZ06] hängt die Geometriemodellierung hier nicht

von der Detektion des Referenzteiles ab. Die Positionen der Unterknoten können aufgrund der relativen Positionsmodellierung auch ohne eine bekannte Position für t verglichen werden. Der Fall, daß die Geometrie aufgrund eines nicht detektierten Referenzteils nicht modelliert werden kann, tritt daher nie auf.

Die Werte ς und ϑ sind Maße für die geometrische Genauigkeit und Vollständigkeit des Mustervergleichs während der Objekterkennung. Der Wert ς legt den Durchmesser von m Toleranzbereichen um die in L gespeicherten Positionen der Unterknoten fest. Unterknoten werden nur innerhalb dieser Bereiche detektiert. Die Schwelle ϑ legt fest, wieviele Unterknoten detektiert werden müssen, damit t erkannt wird. Mit diesen Parametern läßt sich der Geometrieeinfluß auf die Mustererkennung einstellen. Für kleine Durchmesser ς können Muster nur bei genauer Übereinstimmung der Geometrie erkannt werden. Für große Durchmesser relativ zur Streuung der Knotenpositionen L läßt der Geometrieeinfluß nach, d.h. ein Knoten nähert sich einem Bag of features an. Diese Modellierungsfreiheit erlaubt eine Abschätzung der Bedeutung der Geometrie für die Objekterkennung in der Trainingsphase.

Zuletzt enthält ein Knoten t noch einen Schlüssel κ. Für innere Knoten speichert dieser Schlüssel den Index eines Knotens, d.h. $t_i = \{\ldots, \kappa = i\}$. Für Blattknoten, d.h. die Merkmale der untersten Ebene, speichert der Schlüssel einen Deskriptor oder Attributvektor. Konzeptionell besteht kein Unterschied zwischen einem Index und einem Deskriptor: Da ein Deskriptor im Rechner immer diskret vorliegt, kann er in einen Index umgerechnet werden. Dazu werden alle auftretenden Ausprägungen abgezählt. Die Objekterkennung arbeitet genau nach diesem Prinzip. Bestimmte zusätzliche Maßnahmen dienen der Behandlung von Rauschen. Der Schlüssel κ ermöglicht daher die Identifikation eines Knotens und die Navigation zwischen verschiedenen Abstraktionsebenen. Grundsätzlich ist das Modell nicht an einen bestimmten Merkmalsdetektor und Deskriptor gebunden. Für hochdimensionale Deskriptoren sollte jedoch die Diskretisierung und Rauschbehandlung angepaßt werden.

Abbildung 4.1 gibt ein Beispiel für ein Modell mit drei Hierarchieebenen. Die Knoten 1–4 repräsentieren vier verschiedene Ausprägungen lokaler Bildmerkmale.

$$
\begin{aligned}
t_1 &= \{m = 0, U = \{\}, L = \{\}, \ldots, \kappa = 1\} \\
t_2 &= \{m = 0, U = \{\}, L = \{\}, \ldots, \kappa = 2\} \\
t_3 &= \{m = 0, U = \{\}, L = \{\}, \ldots, \kappa = 3\} \\
t_4 &= \{m = 0, U = \{\}, L = \{\}, \ldots, \kappa = 4\}
\end{aligned}
$$

Die Knoten 5–8 sind abstrakter und setzen sich aus mehreren Merkmalen zusammen. Wie an Knoten 5 und 6 gezeigt, dürfen mehrere Unterknoten dieselbe Merkmalsausprägung besitzen. In ihren geometrischen Beziehungen zu Knoten 5 und 6 unterscheiden sie sich jedoch.

$$
\begin{aligned}
t_5 &= \{m = 5, U = \{1, 4, 2, 1, 3\}, L = \{\ldots\}, \ldots, \kappa = 5\} \\
t_6 &= \{m = 5, U = \{1, 4, 3, 3, 1\}, L = \{\ldots\}, \ldots, \kappa = 6\} \\
t_7 &= \{m = 3, U = \{4, 2, 1\}, L = \{\ldots\}, \ldots, \kappa = 7\} \\
t_8 &= \{m = 3, U = \{5, 6, 7\}, L = \{\ldots\}, \ldots, \kappa = 8\}
\end{aligned}
$$

Das Beispiel ist etwas vereinfacht, da der Schlüssel κ in der Praxis noch einen Merkmalsdeskriptor und Verwaltungsinformationen enthalten kann.

Betrachtet man die räumliche Anordnung der Teile in Abbildung 4.1 links, dann stellt man fest, daß die Merkmalsvektoren an der Basis verschiedener Teilbäume des Modells räumlich nebeneinander liegen. Daraus ergibt sich, daß der Bildbereich, über den sich ein Teilbaum ausdehnt, umso größer ist, je höher der Wurzelknoten des Teilbaums im Gesamtmodell liegt. Dies hat Auswirkungen auf die Anzahl der Hierarchieebenen. Für den Fall, daß Teilbäume nicht überlappen dürfen, steigt die Ausdehnung exponentiell mit der Anzahl an Hierarchieebenen. Dies bedeutet, daß die Anzahl an Hierarchieebenen logarithmisch von der Bildgröße abhängt. Im vorliegenden Modell sind Überlappungen erlaubt und treten während des Trainings auch auf. Dadurch entfällt die Abhängigkeit zwischen Hierarchie und Bildgröße. Dies ist jedoch nur dadurch möglich, daß der fehlende räumliche Spielraum die Bedeutung Teilegeometrie im Vergleich zum eigentlichen Auftreten eines Teils in den Hintergrund drängt. Die Anzahl an Hierarchieebenen hängt auch von der Anzahl an Unterknoten ab. Für einen Binärbaum ergibt sich die größte Anzahl an Hierarchieebenen, Bäume mit mehr Unterknoten ergeben flachere Hierarchien. Da das vorgestellte Modell auf Mehrheitsentscheiden anhand des Schwellwerts ϑ basiert, ist eine größere Zahl von Unterknoten vorteilhaft für eine sichere Detektion. Diese Dinge lassen sich theoretisch schlecht ergründen. Experimentell hat sich jedoch die in Abbildung 4.2 dargestellten Modellstruktur als günstig erwiesen. Auf der untersten Ebene liegen die erwähnten lokalen Bildmerkmale. Die nächsthöhere Ebene repräsentiert das geforderte visuelle Alphabet in Form allgemeingültiger Teile. Aus diesen allgemeingültigen Teilen setzen sich Objektansichten zusammen. Diese entsprechen Verallgemeinerungen ähnlicher 2D-Ansichten von Objekten. Diese werden auf der höchsten Ebene durch Kategorieknoten zusammengefaßt. Modellierbar sind im übrigen auch Verbindungen von Knoten über mehrere Ebenen hinweg. Diese werden derzeit jedoch nicht trainiert, da die Bedeutung solcher Verbindungen noch unklar ist, obwohl sie im visuellen Kortex auch auftreten. Darüberhinaus ermöglicht das Modell auch Rückverbindungen von höheren auf niedrigere Ebenen. Dieser Fall wird in der vorliegenden Arbeit auch nicht behandelt, da es für nichtzyklische Modelle noch viel zu entdecken gibt.

Kapitel 5

Ablauf der Objekterkennung

Während der Objekterkennung werden für jeden Knoten des Modells Mehrheitsentscheide getroffen, die Ähnlichkeiten zur verallgemeinerten Hough-Transformation [Bal81], aber auch zu Radiale-Basisfunktionen oder Parzen-Fenstern aufweisen.

Zur Objekterkennung werden die Merkmale eines Bildes mit den Merkmalspunkten auf der niedrigsten Teileebene des Modells verglichen. Übereinstimmungen werden im Modell nach oben weitergereicht. Innere Knoten des Modells gelten als erkannt, wenn genügend Übereinstimmungen auf der tieferen Ebene vorliegen.

Um dies durchzuführen, wird zunächst eine Tabelle angelegt, die zu jedem Knoten die übergeordneten Knoten aufzählt. Die Tabelle ähnelt der r-Tabelle der erweiterten Hough-Transformation [Bal81] und vereinfacht die Navigation im Modell.

Als nächstes werden die Merkmale des Bildes mit den Blattknoten des Modells verglichen. Im Falle einer Übereinstimmung wird mit Hilfe der Tabelle ermittelt, in welchen direkt übergeordneten Teilen das Merkmal modelliert ist. Es werden die Hypothesen generiert, daß die übergeordneten Teile auftreten.

Die Hypothesen enthalten auch den Ort, an dem ein Teil auftreten soll. Wenn sich die Hypothesen an einer bestimmten Bildposition stark genug häufen, dann gilt ein Teil als erkannt. Dabei sind die Geometrieparameter zu beachten, die für jeden Knoten des Modells gespeichert sind.

Es ergibt sich der folgende Ablauf:

- Erzeugung einer Tabelle ähnlich der r-Tabelle der Hough-Transformation
- Merkmalsextraktion auf dem Testbild
- Erzeugung von Hypothesen bezüglich übereinstimmender Knoten
- Weiterleitung der Hypothesen auf abstraktere Modellebenen

- Terminierung an den Wurzelknoten des Modells

Der Ablauf wird nun detailliert dargestellt.

5.1 Informationsfluß im Modell

Um die Navigation zwischen verschiedenen Abstraktionsebenen des Modells zu vereinfachen, wird eine Tabelle

$$LUT : \kappa_i \rightarrow \{\text{alle } j | i \in U_j\}$$

angelegt, die zu jedem des Knotens t_i des Modells eine Liste von abstrakteren Knoten angibt, die t_i als Unterknoten haben Der Zugriff auf die Tabelle geschieht über den Schlüssel κ_i , der den Knoten t_i eindeutig identifiziert. Für innere Knoten des Modells wird κ einfach über einen fortlaufenden Knotenindex berechnet. Für Blattknoten, welche lokale Merkmale repräsentieren, bezeichnet κ die Ausprägung des Deskriptors. Dazu wird ein Deskriptor

$$\boldsymbol{D} = (d_1, d_2, d_3, \ldots, d_{n_D})$$

der Dimensionalität n_D mit den Quantisierungsstufen

$$\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_{n_D}$$

diskretisiert. Da die Anzahl der in einem Modell auftretenden diskreten Ausprägungen von Deskriptoren begrenzt ist, kann jeder Ausprägung ebenfalls eine eindeutige Knotenidentifizierung zugeordnet werden. In den Testbildern können dabei auch Merkmalsausprägungen auftreten, die keine Entsprechung im Modell haben, da sie in der Trainingsstichprobe nicht auftraten. Diese Fälle können allerdings erkannt werden. Die LUT liefert in diesem Fall eine leere Menge, d.h. diese Merkmale werden nicht weiter ausgewertet. Es ist außerdem zu beachten, daß einem Deskriptor aufgrund von Rauschen bei der Diskretisierung eine falsche Identifizierung zugeordnet werden kann. Daher werden für jeden Deskriptor auch die verrauschten Varianten

$$\tilde{\boldsymbol{D}} = (d_1 \pm \zeta_1, d_2 \pm \zeta_2, d_3 \pm \zeta_3, \ldots, d_{n_D} \pm \zeta_{n_D} 1) \tag{5.1}$$

in die LUT eingetragen. Dies entspricht allen Deskriptoren innerhalb eines durch die Quantisierungsstufen vorgegebenen Radius in der Manhattan-Norm. Bei hochdimensionalen Deskriptoren führt diese Vorgehensweise jedoch möglicherweise zu instabilen Ergebnissen [HAK00], obwohl die Manhattan-Norm sich in dieser Hinsicht schon gutmütiger verhält als der euklidische Abstand [AHK01]. Für hochdimensionale Deskriptoren bietet sich daher eine Diskretisierung mit Hilfe eines vorangehenden, an die Deskriptordimensionalität angepaßten Clusterungsschrittes an. Da in der vorliegenden Arbeit nur niederdimensionale Deskriptoren ($n_D \leq 3$) verwendet werden, ist dies hier noch nicht nötig.

Während der Objekterkennung wird die LUT genutzt, um zu einem detektierten Knoten t_i eine Liste von Hypothesen der Form

$$\boldsymbol{h} = (\kappa_j, x, y, c, \mathfrak{d}) \tag{5.2}$$

von Elternknoten t_j zu erzeugen. Dabei sind $x, y, c, \mathfrak{d}$ Eigenschaften, die sich aus t_j ermitteln lassen. Da diese Eigenschaften bereits vollständig während der Initialisierung der LUT bekannt sind, ergibt sich eine Geschwindigkeitssteigerung, wenn sie direkt dort eingetragen werden, statt während der Objekterkennung ermittelt zu werden. Die Tabelle erhält somit die Form

$$LUT : \kappa_i \rightarrow \{\text{alle } \boldsymbol{h} = (\kappa_j, x = x_c \in L_j, y = y_c \in L_j, c, \mathfrak{d}) | i = u_c \in U_j\}.$$

Hypothesen bekommen dadurch die Bedeutung eines Hinweises auf die Anwesenheit eines abstrakteren Knotens t_j aufgrund der Detektion des Knotens t_i. Mit Hilfe des Schlüssels κ_i des detektierten Knotens oder Merkmals t_i wird auf den Index j des Elternknotens t_j geschlossen. Dies geschieht durch das Auffinden des Index $u_c = i$ in der Liste von Unterknoten U_j des Knotens t_j. Der Wert c ist der Index des Unterknotens in der Unterknotenliste und ermöglicht die Bestimmung der relativen Position x_c, y_c zwischen t_i und t_j. Während der Objekterkennung basieren Hypothesen immer auf tatsächlich auftretenden Merkmalen. Die Position x, y, an der eine Hypothese ein mögliches Objekt anzeigt, ergibt sich daher aus der Position der Basismerkmale zusammen mit der relativen Lage aus dem Modell. Das heißt, daß in der LUT gespeicherte Hypothesen Relativpositionen angeben, die während der Objekterkennung erzeugten Hypothesen dagegen absolute Bildkoordinaten. Aus organisatorischen Gründen speichern Hypothesen zusätzlich die Tiefe $\mathfrak{d}$ des abstrakten Knotens t_j in der Hierarchie.

5.2 Nachweis von Modellknoten

Die Objekterkennung basiert auf der Auswertung von Hypothesen. Wenn die Auswertung einer Menge von Hypothesen bezüglich des Auftretens eines bestimmten Knotens zu einer positiven Entscheidung führt, gilt das Auftreten des Knotens als nachgewiesen. Die positive Entscheidung wird in Form einer *Instanz* eines Knotens gespeichert. Instanzen entsprechen bezüglich der gespeicherten Daten den Hypothesen. Der einzige Unterschied ist, daß die Positionen von Instanzen in absoluten anstelle von relativen Koordinaten angegeben wird.

Die Objekterkennung beginnt mit einer Merkmalsextraktion auf einem in Frage kommenden Testbild. Die so ermittelten lokalen Merkmale bestehen jeweils aus der Bildposition und einem Deskriptor. Merkmale, die Blattknoten des Modells entsprechen, werden als Instanzen dieser Knoten gespeichert. Die übrigen werden verworfen. Durch die Auswertung der im Modell vorgegebenen geometrischen Beziehungen werden nun zunehmend abstraktere Darstellungen erzeugt bis sich je nach Bildmaterial Instanzen von Kategorieknoten ergeben. Diese sind das Ergebnis der Objekterkennung. Im Gegensatz zu einigen in der Literaturdurchsicht vorgestellten Verfahren wird so eine gleichzeitige Detektion und Lokalisation erreicht.

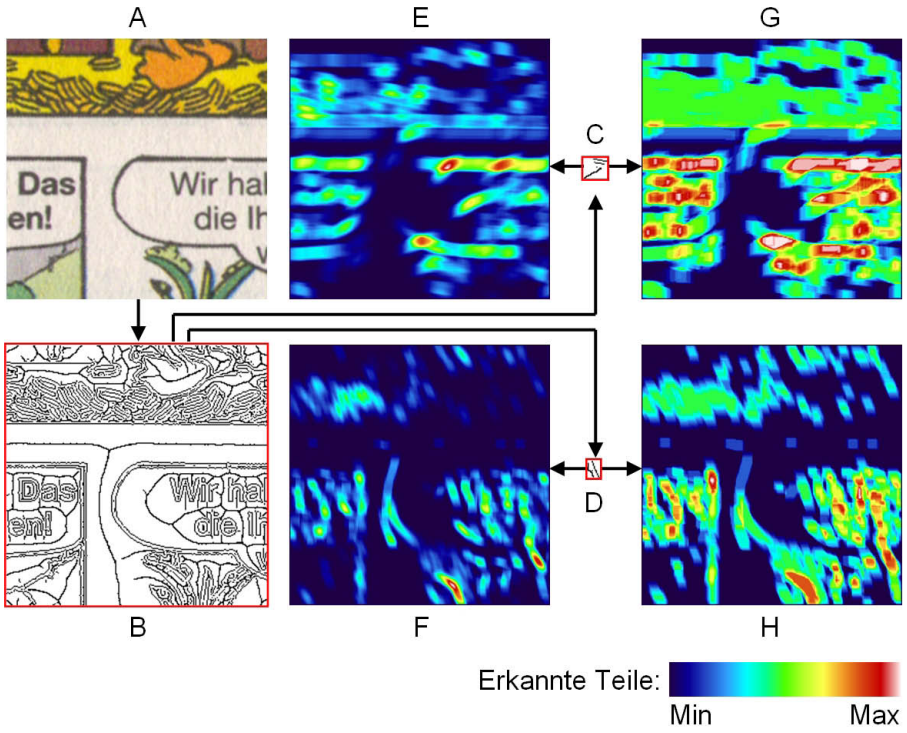


Abbildung 5.1: **Akkumulation von Hypothesen (E, F) verglichen mit der Akkumulation von Belegen (G, H).** Die Teile C und D werden im Merkmalsbild B gesucht. Die farbigen Plots E–H stellen jeweils die Anzahl der erkannten Unterteile über der Bildebene dar. Für die Akkumulation von Belegen ergeben sich mehr korrespondierende Bildbereiche. Da die ermittelten Korrespondenzen nicht von benachbarten Merkmalen beeinflußt werden, läßt sich auch der Schwellwert zur Erkennung zielgerichtet einstellen.

Instanzen eines abstrakteren Knoten t_j werden erzeugt, wenn genügend Belege für das Vorhandensein des Knotens angehäuft wurden. Ein Beleg entspricht einem detektierten Unterknoten. Wenn alle Belege für einen Knoten angesammelt wurden, wird durch den Vergleich mit dem jeweiligen Knotenschwellwert ϑ_j entschieden, ob t_j vorliegt oder nicht. Dieses Vorgehen läßt sich in Form von vier Schritten zusammenfassen:

- Herleitung von Hypothesen anhand von Instanzen
- Erbringung von Belegen für mögliche Instanzen
- Nachweis der Hypothesen durch Auszählen der Belege
- Erzeugung neuer Instanzen bei ausreichenden Belegen

Die Herleitung von Hypothesen geschieht durch die Auswertung der geometrischen Beziehungen zwischen Knoten und Unterknoten. Als Ausgangspunkt dient eine Instanz eines Knotens oder Merkmals t_i

$$(\kappa_i, x_i, y_i, c_i, \mathfrak{d}_i).$$

Ein Nachschlagen in der *LUT* liefert eine Liste aller Elternknoten von t_i einschließlich ihrer räumlichen Relativpositionen

$$\{(\kappa_{j_1}, x_{j_1}, y_{j_1}, c_{j_1}, \mathfrak{d}_{j_1}), (\kappa_{j_2}, x_{j_2}, y_{j_2}, c_{j_2}, \mathfrak{d}_{j_2}), (\kappa_{j_3}, x_{j_3}, y_{j_3}, c_{j_3}, \mathfrak{d}_{j_3}), \dots\}.$$

Indem der Schlüssel κ_i der Instanz durch den Schlüssel κ_j eines Elternknotens ersetzt wird, ergibt sich eine Hypothese für den Elternknoten. Die vermutete Position des Elternknotens ergibt sich durch Addition der Relativposition x_j, y_j auf die absolute Position der Instanz. Dies führt zu einer Liste von Hypothesen

$$\{(\kappa_{j_1}, x_i + x_{j_1}, y_i + y_{j_1}, c_{j_1}, \mathfrak{d}_{j_1}), (\kappa_{j_2}, x_i + x_{j_2}, y_i + y_{j_2}, c_{j_2}, \mathfrak{d}_{j_2}), \dots\}. \quad (5.3)$$

Die Hypothesen betreffen verschiedene abstrakten Knoten $t_{j_1}, t_{j_2}, t_{j_3}, \dots$ in möglicherweise unterschiedlichen Hierarchieebenen. Zur Auswertung von Hypothesen zu einem bestimmten Knoten liegen daher erst dann mit Sicherheit alle Daten vor, wenn alle niedrigeren Hierarchieebenen vollständig ausgewertet wurden.

Um einen Knoten $t_j = (m_j, \dots, \varsigma_j, \dots)$ nachzuweisen, wird nun auf alle Hypothesen

$$\mathfrak{H}_j = \{\boldsymbol{h} | \kappa = j\} \quad (5.4)$$

zugegriffen, die den Knoten vorhersagen. Die Vorhersage beruht letztendlich auf Instanzen der Unterknoten von t_j. Die Nummern dieser Unterknoten speichern Hypothesen in dem Wert c. Zu jedem Wert von $c = 1, 2, 3, \dots, m_j$ wird nun eine binäre Funktion

$$\mathfrak{b}_c(x, y) = \begin{cases} 1, & \text{falls } \exists \boldsymbol{h}_i = (..., x_i, y_i, c_i, ...) \in \mathfrak{H}_j | c_i = c \\ & \text{und } \|(x, y) - (x_i, y_i)\| \leq \varsigma_j / 2 \\ 0, & \text{sonst} \end{cases} \quad (5.5)$$

definiert, die für eine Bildkoordinate x, y den Wert Eins liefert, falls eine durch eine Instanz des Unterknotens $u_c \in U_j$ erzeugte Hypothese existiert, die im Umkreis von $\varsigma_j/2$ den Knoten t_j anzeigt. Wenn $\mathfrak{b}_c = 1$ ergibt, bedeutet dies, daß der Unterknoten t_{u_c} von t_j in der richtigen relativen Position auftritt. Die Funktion liefert in diesem Fall einen Beleg für eine Instanz von t_j. Gleichung 5.5 kann als radiale Basisfunktion mit einem rechteckigen Kern gedeutet werden. Da die Abstandsberechnung aus Geschwindigkeitsgründen nach der Manhattan-Norm geschieht, legt diese quadratische Bereiche in der Bildebene fest. Alternativ kann dies auch als Dilatation um die Position des gefundenen Knotens gedeutet werden.

Ob genügend Belege vorliegen, um Instanzen eines Knotens t_j zu erzeugen, wird mit Hilfe eines Akkumulators

$$\Xi_j(x, y) = \sum_c \mathfrak{b}_c(x, y) \tag{5.6}$$

entschieden, der für jede Bildposition angibt, wieviele Unterknoten dort gefunden wurden. Falls für eine Position x, y aufgrund des Kriteriums

$$\Xi_j(x, y) > \vartheta_j \tag{5.7}$$

eine ausreichende Anzahl von Belegen festgestellt wird, wird eine neue Instanz

$$(\kappa_j, x, y, \dots).$$

erzeugt. Die Hypothesen $\mathfrak{H}_j$ sind damit veraltet. Der Wert c der neuen Instanz muß dabei übrigens noch nicht festgelegt werden, da er bei der Auswertung der Geometrie aus der *LUT* ausgelesen und erst später beim Beleg neuer Hypothesen eine Rolle spielt.

Die Akkumulation von Unterteilen über der Bildebene entspricht weitgehend der erweiterten Hough-Transformation. Abgesehen davon, daß hier ein hierarchisches Teilemodell behandelt wird, grenzt jedoch vor allem die Unterscheidung zwischen Hypothesen und Belegen das beschriebene Verfahren von der Hough-Transformation ab. Da bei der Hough-Transformation Hypothesen akkumuliert werden, tritt schnell der Fall auf, daß mehr vermeintliche Teile gefunden wurden, als ein Objekt insgesamt besitzt [GH88]. Dabei treten insbesondere bei mehrdeutigen Bildstrukturen einzelne Teile mehrfach auf und werden dann mehrfach gezählt. Die Unterscheidung von Hypothesen und Belegen löst dieses Problem, da Teile eines Objekts immer nur einfach gezählt werden, auch wenn sie mehrfach auftreten. Da die Maximalzahl an akkumulierten Belegen durch die Anzahl der Teile begrenzt ist, kann auch die Schwelle ϑ einfacher eingestellt werden.

Abbildung 5.1 zeigt dies für die Detektion zweier Teile. Als Ausgangsbild dient hier das Hintergrundbild A. Die Merkmalsextraktion bringt Kanten- und Skelettmerkmale an den Bild B schwarz markierten Positionen hervor. Die gesuchten Teile sind in gleicher Weise in Bild C und Bild D dargestellt. Die Ergebnisse der Akkumulation von Hypothesen zeigen die Bilder E und F links von den gesuchten Teilen. Die Akkumulation von Belegen ist dagegen in den

Bildern E und F rechts dargestellt. Die Darstellungen sind jeweils auf das Maximum normiert. Für die Akkumulation von Hypothesen ergeben sich wenige, aber stark ausgeprägte Häufungen. Diese werden teilweise durch ähnliche Merkmale in der Nachbarschaft verstärkt, was bei der Festlegung der Schwellwerte ϑ berücksichtigt werden muß. Wird der Schwellwert jedoch höher eingestellt, werden mögliche korrespondierende Bildbereiche ignoriert, die keine verstärkende Umgebung besitzen. Im Vergleich dazu ergeben sich für Akkumulation von Belegen auch deutliche Häufungen an anderen passenden Bildpositionen. Die größten Häufungen sind dabei allerdings nicht so stark ausgeprägt wie im ersten Fall, was sich aufgrund der normierten Darstellung durch eine höhere mittlere Anzahl an gefundenen Unterteilen über das gesamte Bild zeigt. Dies ist jedoch kein Nachteil, da die Höhe der Maxima begrenzt ist und sich der Schwellwert gezielt darauf einstellen läßt.

Das vorgestellte Objekterkennungssystem bietet die Möglichkeit, die Akkumulation in Gleichung 5.7 durch Gewichte $w_1, w_2, w_3, \ldots, w_c$ zu erweitern. Die gewichtete Akkumulation

$$\Xi_j(x,y) = \sum_c w_c \mathfrak{b}_c(x,y)$$

erlaubt die Überprüfung der heuristischen Optimierungen von Selinger [Sel01] oder die Modellierung von Auftrittswahrscheinlichkeiten ähnlich Mikolajczyk et al. [MLS06]. Aufgrund der Ortstoleranz und der zweifachen Binarisierung in den Gleichungen 5.5 und 5.7 ist der pragmatische Ansatz von Selinger sicher erfolgversprechender. Rein formal würde sich auch eine Weiterentwicklung in Richtung Boosting [FS99] anbieten. Aufgrund des hohen Rechenaufwandes, der durch die zusätzlichen Multiplikationen erzeugt wird, werden Gewichtungen jedoch vorerst nicht weiter verfolgt.

5.3 Praktische Erwägungen

Die Erkennung der Knoten des Modells geschieht beginnend mit einfachen Bildmerkmalen nach Abstraktionsebenen geordnet. Dieses Vorgehen stellt sicher, daß beim Nachweis der Knoten immer alle Hypothesen vollständig vorliegen. Der Grund dafür ist, daß manuell Teilebäume erstellt werden können, bei denen die Unterknoten eines Knotens auf verschiedenen Hierarchieebenen liegen. Dies kann auftreten, wenn einzelne Unterknoten mehrere Elternknoten besitzen, von denen einige gleichzeitig Unterknoten des ersten Knotens sind. Aus diesem Grund kann es passieren, daß die LUT zu einem Schlüssel κ eine Liste von Hypothesen zu Knoten auf verschiedenen Abstraktionsebenen liefert. Abbildung 5.2 stellt diese Abhängigkeiten zwischen den Ebenen dar. Um zu gewährleisten, daß die Knoten in der richtigen Reihenfolge ausgewertet werden, enthalten die Hypothesen einen Wert $\mathfrak{d}$, der die Hierarchieebene des vermuteten Knotens angibt.

Innerhalb einer Ebene bestehen keine Abhängigkeiten zwischen den Knoten des Modells. Die Detektion von Knoten kann daher in beliebiger Reihenfolge

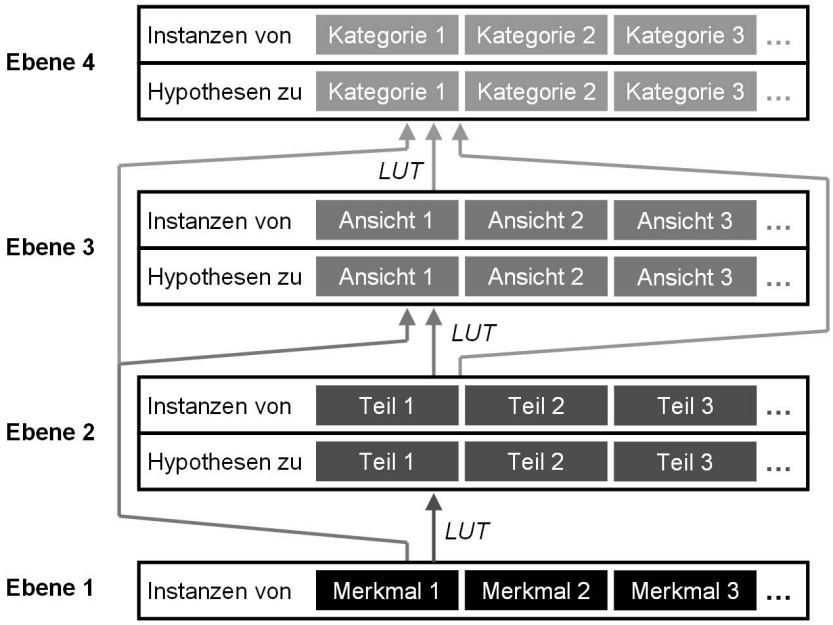


Abbildung 5.2: **Erkennung der Knotenhierarchie durch das Aufstellen von Hypothesen.** Instanzen von Modellknoten oder Merkmalen lassen Hypothesen bezüglich der Existenz ihrer Elternknoten zu. Diese liegen immer auf mindestens einer abstrakteren Ebene. Die Erzeugung und Auswertung von Hypothesen geschieht daher zweckmäßigerweise auf der niedrigsten Ebene und wird dann jeweils für die nächsthöhere Ebene fortgesetzt.

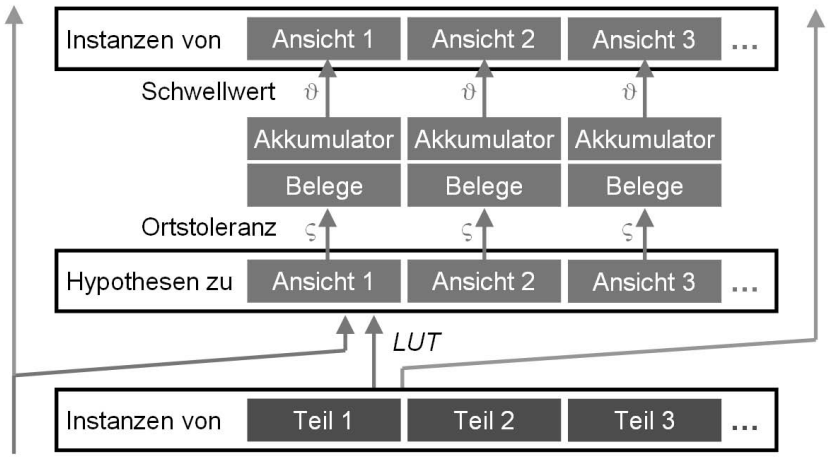


Abbildung 5.3: **Unabhängige Erkennung von Knoten einer Abstraktionsebene.** Innerhalb einer Ebene kann die Erkennung von Modellknoten in einer beliebigen Reihenfolge geschehen. Dazu werden separat für jeden Knoten, für den Hypothesen vorliegen, unter Berücksichtigung der Geometrie Belege gesammelt. Eine Schwellwertoperation entscheidet über das Anlegen neuer Instanzen des untersuchten Knotens.

geschehen. Wie Abbildung 5.3 zeigt, werden für jeden Knoten des Modells eigene Belegfunktionen und ein eigener Akkumulator erzeugt, die in der Software in Form von Matrizen implementiert sind. Abbildung 5.4 stellt den Inhalt der Akkumulatoren und das Ergebnis nach der Schwellwertbildung in Gleichung 5.7 graphisch dar.

Durch die getrennten Belegmatrizen und Akkumulatoren der Modellknoten ergeben sich zwei mögliche Strategien für die Auswertung der Hypothesen:

- Aufbewahrung aller Belege im Speicher des Rechners
- Zwischenspeicherung aller Hypothesen

Die erste Methode hat den Vorteil, daß Hypothesen sofort bei ihrer Erzeugung in die passenden Belegfunktionen eingetragen werden können. Dies geschieht durch das Setzen eines Bits an der durch die Hypothese vorgegebenen Position. Die Objekterkennung geschieht damit für alle Knoten parallel, was mögliche Optimierungen auf Hardwareseite zuläßt. Dabei ist jedoch zu bedenken, daß für jeden Knoten m Belegfunktionen mit den Ausmaßen eines Bildes gespeichert werden müssen. Diese sind als zweidimensionale Matrix über den Bildkoordinaten implementiert, wobei die Belegfunktionen für jede Koordinate ein Bit speichern und der Akkumulator 32 Bit. Aus Abbildung 4.2 ist ersichtlich, daß die Anzahl der Unterknoten m bis in die Größenordnung 10^4 reicht. Praxistaugliche Modelle ergeben daher schnell einen Speicherbedarf im Gigabytebereich.

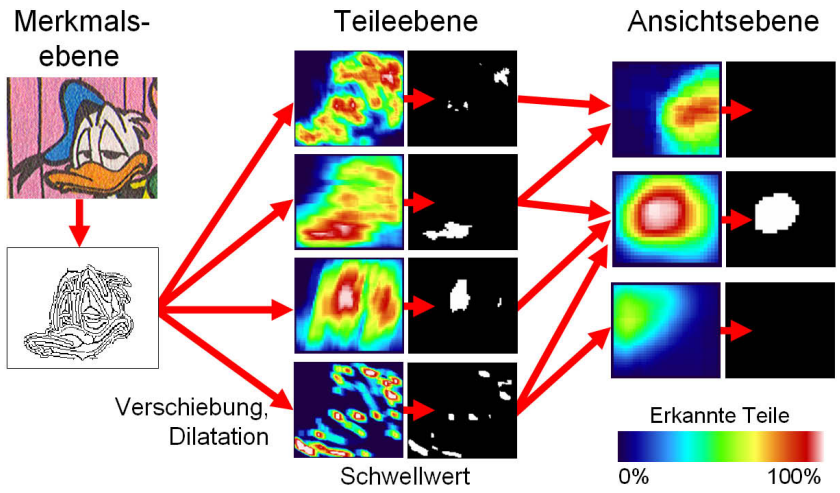


Abbildung 5.4: **Hierarchische Mehrheitsentscheidung während der Objekterkennung** Eine Merkmalsextraktion auf dem Testbild (links oben) liefert Instanzen von Merkmalsknoten des Modells. Das schwarz-weiße Bild links unten zeigt die Positionen solcher Merkmale in der Bildebene. Die Merkmalsknoten werden in Form von Belegen gemäß ihrer relativen Positionen als Unterknoten abstrakterer Knoten zu den Zentren der abstrakteren Knoten verschoben (siehe Gleichung 5.3). Eine gewisse Ortstoleranz entsteht durch Dilatation der sich ansammelnden Belege (Gleichung 5.5). Die farbigen Diagramme in der Mitte und rechts geben die Verteilung der Belege über dem Ort an. Jedes Diagramm steht hier für die räumliche Lage der Belege jeweils eines bestimmten Modellknotens. Durch das Auszählen (Gleichung 5.6) und Binarisieren (Gleichung 5.7) der Belege entstehen jeweils die schwarz-weißen Trefferbilder rechts neben den farbigen Diagrammen. Die weißen Stellen markieren die Positionen, an denen ein Knoten aufgrund einer ausreichen hohen Zahl von Belegen erkannt wurde. Durch eine Unterabtastung der Trefferbilder abhängig von der Ortstoleranz werden neue Instanzen der betreffenden Knoten erzeugt. Der Vorgang wird dann auf der nächsten Modellebene wiederholt. In dieser Abbildung ist die Ansichtsebene die höchste Ebene. Auf dieser Ebene zeigt nur das mittlere Trefferbild einen erkannten Knoten an. Die entsprechenden Punkte des Originalbildes werden daher der Sollklasse dieses Knotens zugeordnet.

Aus diesem Grund wird hier die zweite Strategie angewandt, bei der die Speicherung der Belegfunktionen dadurch umgangen wird, daß die Knoten des Modells sequentiell nachgewiesen werden. In dem Fall werden nur die Belegfunktionen für den aktuell auszuwertenden Knoten benötigt, was den Speicheraufwand um die Anzahl der Knoten im Modell reduziert. Dann müssen allerdings die Hypothesen nach der Erzeugung in Listen zwischengespeichert werden, da die entsprechenden Belegfunktionen erst später erzeugt werden. Wie sich experimentell zeigt, führt allerdings eine vollständige Auswertung aller Hypothesen schnell zu einem vergleichbaren Speicheraufwand. Shaked, Yaron und Kiryati [SYK94] zeigen jedoch, daß sich eine Hough-Tranformation auch mit einer kleinen Teilmenge aller Hypothesen durchführen läßt. In der vorliegenden Arbeit wird dieser Ansatz jedoch nicht intensiver untersucht. Stattdessen wird eine schnelle Entropiekodierung auf die Listen von Hypothesen angewandt, was das zeitaufwendige Auslagern von Speicherseiten vermeidet.

Als kritisch hat sich die Umdimensionierung der Listen herausgestellt. Da im Voraus nicht bekannt ist, wieviele Hypothesen das Nachschlagen von Instanzen 5.3 ergibt, müssen Hypothesen in dynamischen Liste gespeichert werden. Da Neudimensionierungen solcher Listen mit Speicheranforderungen an das Betriebssystem einhergehen, sollte bei zu kleinen Listen unbedingt auf eine exponentielle Vergrößerung geachtet werden. Es sollte auch beachtet werden, daß intensives Speichermanagement den zur Verfügung stehenden Adressraum unangenehm zerstückelt.

Im nächsten Schritt (Gleichung 5.4) müssen alle Hypothesen gefunden werden, die einen bestimmten Knoten vorhersagen. Dies kann über die in der Gleichung angegebene Bedingung $\kappa = j$ geschehen. Um die hierzu erforderliche Suche über alle Hypothesen zu vermeiden, wird jedem Knoten des Modells eine Liste zur Speicherung der Hypothesen zugeordnet. Beim Aufstellen von Hypothesen werden diese direkt in die passende Liste eingeordnet, sodaß der Suchschritt entfällt.

Die sequentielle Auswertung von Knoten gemäß den Abstraktionsebenen enthält einen weiteren verdeckten Suchschritt. Da Knoten, für die keine Hypothesen vorliegen, nicht bearbeitet werden müssen, geschieht die Auswahl zu untersuchender Modellknoten allein auf Basis der vorliegenden Hypothesen. Dabei wird der Knoten ausgewählt, für den die Hypothese die niedrigste Hierarchiestufe $\mathfrak{d}$ anzeigt. Um diese Suche zu vermeiden, wird eine weitere Liste eingeführt, die unter Angabe der Hierarchiestufe $\mathfrak{d}$ zu einer weiteren Liste aller Hypothesen auf dieser Stufe führt. Innerhalb dieser Liste sind die Hypothesen weiter nach einzelnen Knoten geordnet in Listen gespeichert. Der Einsatz von Listen verschiebt allerdings das Suchproblem nur in die Schlüsselverwaltung der Listen. Aus Effizienzgründen werden daher Listen eingesetzt, in denen Baumstrukturen verwendet werden, um eine schnelle Zuordnung von Schlüsseln und Listenelementen zu erreichen.

Die Berechnung der Belegfunktionen in Gleichung 5.5 geschieht durch das Markieren eines aufgetretenen Unterknotens mit einem gesetzten Bit an der entsprechenden Koordinate. Die Ortstoleranz ς wird durch einen lokalen Operator auf der Matrix hergestellt, indem alle Bits in der Umgebung ebenfalls gesetzt

werden. Aus Geschwindigkeitsgründen wird hier die in Form von ς vorliegende Skaleninformation zur Unterabtastung der Belegfunktion genutzt. Dies reduziert die Anzahl der erzeugten Instanzen erheblich, was zu einer geringeren Zahl von Hypothesen im Folgeschritt führt. Außerdem ermöglicht die Unterabtastung ein viel schnelleres Auszählen des Akkumulators in Gleichung 5.6. Auf die Objekterkennung selbst hat dieses Vorgehen dagegen keine Auswirkung, da mögliche doppelte Hypothesen innerhalb der Ortstoleranz in beiden Fällen nur einfach gezählt werden.

Kapitel 6

Training einzelner Abstraktionsebenen

Das Training beginnt mit der Auswahl geeigneter Merkmalspunkte. Diese werden im nächsten Schritt aufgrund einer statistischen Auswertung von gleichzeitig auftretenden Merkmalen zu Teilen zusammengesetzt. Anhand dieser Teile werden anschließend verschiedene Ansichten eines Objekts modelliert. Dabei ist das Ziel, die möglichen Erscheinungen eines Objekts abzudecken. Schließlich werden Kategorieknoten erzeugt, die verschiedene Ansichten zusammenfassen, um die Objekterkennung auf die Kriterien Sensitivität, Spezifität oder Korrektklassifikationsrate zu optimieren.

6.1 Extraktion lokaler Merkmale

Die Merkmalsextraktion geschieht durch Detektoren für Kanten, Ecken und Flächen. Das Ergebnis der Detektion ist jeweils eine Liste mit Bildkoordinaten, an denen diese Merkmale auftreten. Zu jeder dieser Koordinaten wird ein Deskriptor berechnet, der die lokale Bildumgebung passend zu dem gewählten Merkmal beschreibt. Für Kanten speichert der Deskriptor die Richtung des Gradienten. Für Ecken wird nur das Auftreten selbst, aber keine weitere Information gespeichert. Flächen werden durch Punkte auf Skelettlinien repräsentiert. Der entsprechende Deskriptor speichert die Orientierung der Skelettlinie, den Abstand zur nächsten Kante sowie die Helligkeit der Fläche an der Merkmalsposition. Die nächsten Abschnitte stellen die Verfahren zur Merkmalsextraktion genauer vor. Anschließend werden günstige Quantisierungsintervalle für die Werte der Deskriptoren bestimmt. Dabei ist abzuwägen, daß durch zu grobe Intervalle Information verlohren geht. Andererseits erhöhen zu feine Intervalle den Speicheraufwand, da für jede Merkmalsausprägung Hypothesen zu abstrakteren Teilen aufgestellt werden.

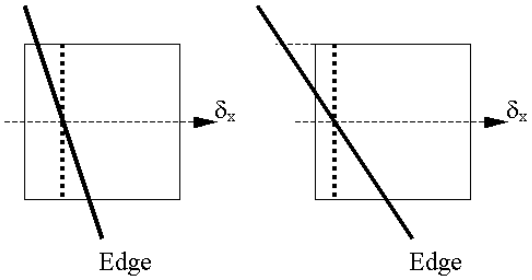


Abbildung 6.1: **Lage einer Kante auf einem quadratischen Bildpunkt [SK04].** Falls eine vertikale (horizontale) Kante die seitlichen (oberen und unteren) Ränder eines Pixels überquert, treten bei der Interpolation mittels einer Parabel Fehler auf, da Licht von beiden Seiten der Kante auf ein dem Zentrumspunkt benachbartes Sensorelement fällt. Das Symbol δ_x entstammt der angegebenen Quelle und stellt die Kantenposition in horizontaler Richtung dar.

6.1.1 Kantenmerkmale

Generell werden alle Bilder vor der Merkmalsextraktion mit einem 5×5-Binomialfilter geglättet, um die Unebenheiten der Bildwerte ι aufgrund des groben Halbtonverfahrens beim Druck der Cartoon-Vorlagen (s. Abb. 6.3 oder auch Anhang B.1) auszugleichen.

Die Kantendetektion geschieht anschließend über die Schätzung des Gradienten

$$\nabla = (\partial\iota/\partial x, \partial\iota/\partial y)$$

durch die Faltung des Bildes mit der Filtermatrix [−0,5, −1, 0, 1, 0,5] (bzw. ihrer Transponierten) einmal in vertikaler und einmal in horizontaler Richtung. Dies entspricht einer Faltung mit der Maske [−1, −1, 1, 1] und einer zusätzlichen Verschiebung um einen halben Bildpunkt. Für je drei horizontal oder vertikal aufeinanderfolgende Bildpunkte mit den Werten ι_1, ι_2 und ι_3 wird über die Bedingung

$$\iota_1 < \iota_2 \text{ und } \iota_2 \geq \iota_3 \tag{6.1}$$

entschieden, ob der mittlere Punkt auf einer Kante liegt. Ist dies der Fall, wird die Bildfunktion lokal durch eine Parabel an ι_1 bis ι_3 angenähert. Die Position der Kante ist durch das Maximum der Parabel gegeben und läßt sich geschlossen als

$$x^* \text{ bzw. } y^* = (\iota_3 - \iota_1)/(2(2\iota_2 - \iota_1 - \iota_3))$$

in horizontaler bzw. vertikaler Richtung berechnen. Unter der Annahme idealer, quadratischer Pixel treten dabei Fehler auf, wenn eine Kante nicht exakt vertikal oder horizontal ausgerichtet ist und dabei über zwei benachbarte Punkte quer zur Hauptrichtung der Kante verläuft. Abbildung 6.1 verdeutlicht dies. Im linken Bild verläuft die fett dargestellte Kante durch den als Quadrat dargestellten Zentrumspunkt, nicht aber durch den linken Nachbarpunkt. Die Fläche links

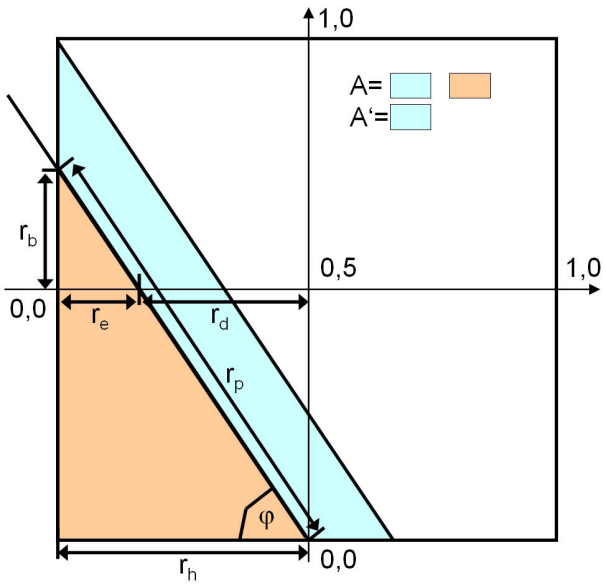


Abbildung 6.2: **Pixelmodell zur Bestimmung der tatsächlichen Lage einer Kante.** Das Quadrat stellt einen Bildpunkt dar, wobei die Mitte des Quadrats die Koordinate $(0{,}5, 0{,}5)$ hat. Die schräge Linie links stellt die gesuchte richtige Kante dar. Diese läßt sich durch die Subpixelverschiebung r_d und den Kantenwinkel φ angeben.

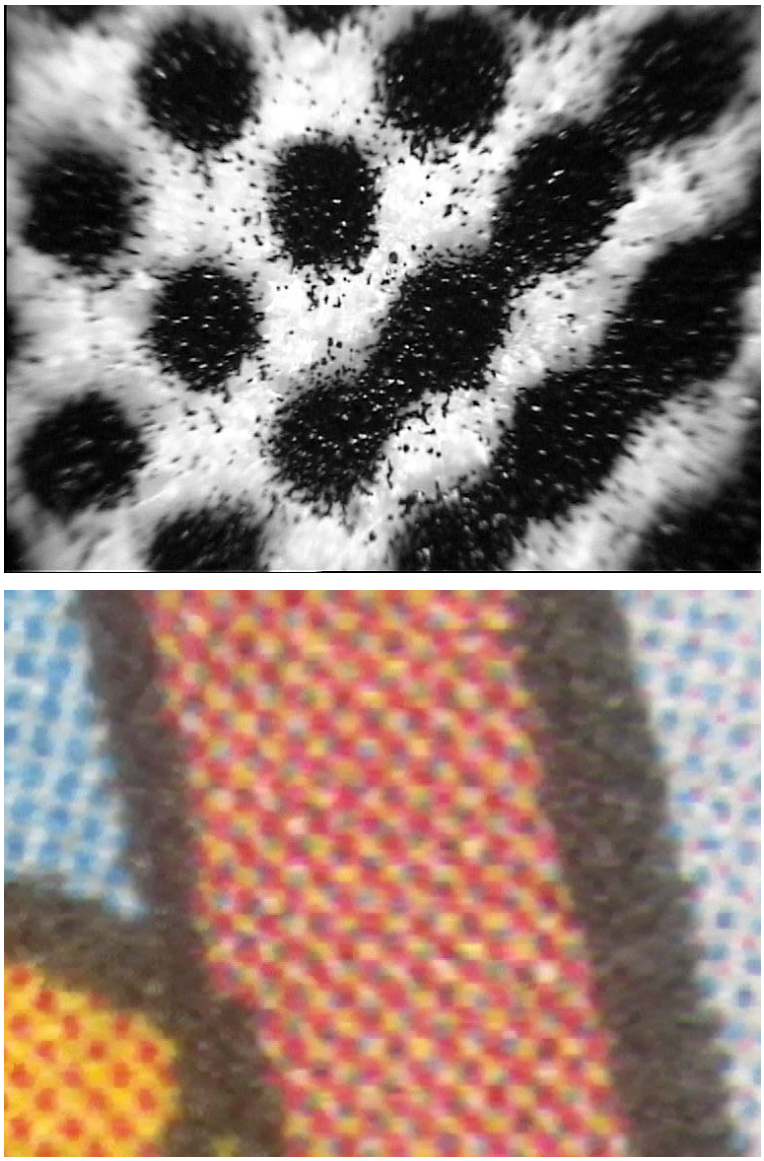


Abbildung 6.3: **Vergrößerungen von Halbtondrucken.** Oben: Mikroskopaufnahme eines rechteckigen Ausschnitts mit einer Höhe von 0,8 mm auf dem Papier. Der Ausdruck geschah auf einen Schwarz-Weiß-Laserdrucker (HP Laser Jet 1300). Deutlich ist zu erkennen, daß im Druck runde Punkte angestrebt werden, obwohl andere Formen möglich sind. Die kleinsten schwarzen Punkte sind einzelne Tonerpartikel, die fast immer etwas streuen. Unten: Scan eines Comics [Dis] mit 2400 dpi. Die Höhe des Bildausschnitts beträgt etwa 3 mm auf dem Papier.

der Kante innerhalb des Zentrumspunkts ist für die schräge Kante und eine exakt vertikale Kante, hier gepunktet eingezeichnet, gleich. Für die Helligkeiten ι_1, ι_2 und ι_3 ergeben sich für die schräge Kante die gleichen Werte wie für die exakt vertikale Kante. Im rechten Fall ergibt sich jedoch eine Abweichung, da für die schräge Kante Licht von ihrer rechten Seite auf den linken Sensor fällt, für eine vertikale Kante jedoch nicht. Da dies bei der Approximation durch die Parabel nicht berücksichtigt ist, wird die Kante zu weit links detektiert. Die Detektion entspricht daher der exakt vertikalen Kante, sodaß die tatsächliche Position der schrägen Kante rechts der fett gezeichneten liegt, und zwar an einer Position, welche die gleiche Fläche auf der linken Seite für eine schräge und eine vertikale Kante ergibt.

Mit dem in Abbildung 6.2 gezeigten Pixelmodell läßt sich dies jedoch korrigieren. Für die Fläche A des Pixels links der gesuchten Kante ergibt sich ein kritischer Wert A', für den die Kante genau durch eine Ecke des Pixels verläuft. Dieser läßt sich mit der Kantenrichtung

$$\varphi = \arctan \frac{\partial \iota / \partial y}{\partial \iota / \partial x}$$

gemäß

$$A' = 0{,}5 / \tan \varphi$$

berechnen. Aus der Gleichung

$$A = \frac{1}{2} {r_p}^2 |\sin \varphi| |\cos \varphi|$$

ergibt sich die Länge r_p der gesuchten Kante innerhalb des Pixels als

$$r_p = \sqrt{\frac{2A}{|\sin \varphi| |\cos \varphi|}}.$$

Für die Höhe r_b des Schnittpunkts der Kante mit der linken Pixelseite über der Pixelmitte ergibt sich dann die Gleichung

$$r_b = r_p \sin \varphi - 0{,}5,$$

mit der sich der horizontale Abstand r_e zwischen der Kante und dem linken Pixelrand auf der mittleren Pixelhöhe als

$$r_e = r_b / |\tan \varphi|$$

berechnen läßt. Aus diesem läßt sich wiederum die tatsächliche Subpixelverschiebung

$$r_d = 0{,}5 - r_e$$

berechnen und es ergibt sich

$$r_d = 0{,}5 - \left(\sqrt{\frac{2A}{|\sin \varphi| |\cos \varphi|}} |\sin \varphi| - 0{,}5 \right) \frac{1}{|\tan \varphi|}.$$

Mit den so berechneten Subpixelverschiebungen lassen sich die über die Kriterien 6.1 ermittelten ganzzahligen Koordinaten korrigieren. Die Berechnung für Schnitte mit der rechten Pixelseite sowie die Behandlung horizontaler Kanten geschieht auf die gleiche Weise.

Dieser Operator hat sich bereits bei der Segmentierung besonders kleinteiliger Objekte [SK04] in realen Kamerabildern bewährt. Der Deskriptor von Kantenpunkten speichert nur die Kantenrichtung. Der Gradientenbetrag eignet sich weniger, da er aufgrund der starken Helligkeitsschwankungen an Kanten unsicher ist, wie bei früheren Arbeiten zur Stereoanalyse festgestellt wurde [KLSK07]. Aus diesem Grund wird der Gradientenbetrag wohl auch nicht im SIFT-Operator [Low99] verwendet. Stattdessen basiert der SIFT-Operator ebenfalls vollständig auf Gradientenrichtungen.

Bei der Anwendung des Operators auf die Cartoon-Datenbank beschränkt allerdings die Bildqualität die theoretisch hohe Präzision des Verfahrens. Zum einen wurden die Bildwerte durch den 5×5-Binomialfilter bereits stark manipuliert. Zum anderen ist das Pixelmodell schlecht an die Aufgabe angepaßt, sowohl was das Einscannen angeht als auch den Druck der Vorlagen. Die optischen Eigenschaften des verwendeten Scanners sind weitgehend unbekannt und werden daher zusammen mit der Erscheinung der Objekte modelliert. Was den Druck der Vorlagen angeht, setzen Verlage und Druckerhersteller bei Grauwertausdrucke überwiegend auf runde Punktformen. Ein Beispiel zeigt 6.3. Bei Farbausdrucken werden mehrere runde, verschiedenfarbige Punkte in einem quadratischen Feld untergebracht, um eine bestimmte Farbe aus wenigen Pigmentarten zu mischen. Selbst wenn man also einen Sensor mit quadratischen Elementen hätte, der die gedruckten Punkte perfekt aufnehmen könnte, wären die Punkte trotzdem rund oder ungleichmäßig über die Fläche verteilt. Die Genauigkeit des quadratischen Pixelmodells wird daher nicht erreicht.

Dies stellt allerdings keinen Nachteil dar, denn das Verfahren ist genauer als die Bildvorlage. Es gehen also keine Informationen verloren. Zum anderen wird die zur Objekterkennung nötige Ortsgenauigkeit beim Erlernen der Teile auf das Bildmaterial und die zu lösende Aufgabe eingestellt. Darüberhinaus ist ein an Cartoon-Vorlagen angepaßtes Modell nicht erstrebenswert, da der Merkmalsdetektor allgemeingültig sein soll. Eine starke Spezialisierung würde andere Anwendungsbereiche ausschließen.

6.1.2 Eckenmerkmale

Ecken werden als Punkte auf Kanten betrachtet, an denen die Krümmung ein lokales Maximum hat. Dazu wird für jeden Kantenpunkt (x, y) aus der vorangegangenen Kantenextraktion die lokale Krümmung bestimmt. Diese wird als Differenz der Gradientenrichtung an zwei Punkten im Abstand von ± 1 Pixel senkrecht zum Gradienten um (x, y) berechnet. Das lokale Maximum wird über den Betrag aller in einem 5×5-Fenster um (x, y) berechneten Krümmungen bestimmt.

Der Eckenoperator ist ein Nebenprodukt der Kantenerkennung. Aufgrund der Kantenverdünnung in Gleichung 6.1 ist die Lokalisation etwas besser als

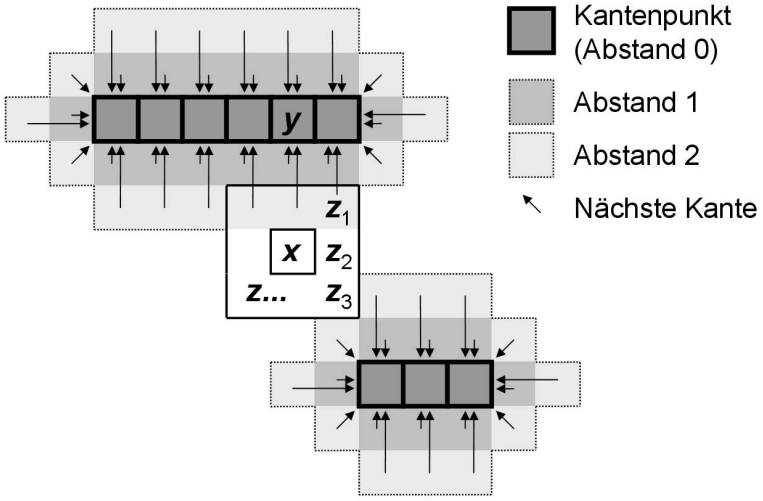


Abbildung 6.4: **Abstandstransformation.** Zunächst wird eine Kantenextraktion durchgeführt, die in diesem Beispiel zwei Konturen liefert (dunkelgrau markiert). Die Abstandstransformation wird jeweils in einem 1 Pixel breiten Rand um bereits bearbeitete Bereiche (hellgrau markiert) durchgeführt. Die Abstandstransformation ordnet jedem Punkt der Umgebung einer Kontur den Kantenpunkt mit dem geringsten euklidischen Abstand zu (Pfeilmarkierungen). Dazu wird zu jedem Punkt x auf dem Rand eine Liste von nahen Kantenpunkten ermittelt, indem bereits bestehende Zuordnungen von Nachbarpunkten z untersucht werden. Der Kantenpunkt aus der Liste mit dem geringsten Abstand zu x wird dann x zugeordnet. Der Abstand wird für jeden Punkt gespeichert.

beim Moravec-Operator, der die Eckenposition über die Größe des Filterfensters verschmiert.

Als Deskriptor von Eckenmerkmalen wird eine Konstante gespeichert, d.h. das bloße Auftreten eines Eckpunkts. Da Ecken sehr viel seltener sind als Kanten, sind sie auch für die Bildung von Teilen beim Training der nächsten Hierarchiestufe eher von untergeordneter Bedeutung. Da Ecken potentiell informationshaltiger sind als Kanten, könnte der Nachteil der geringeren Häufigkeit durch eine entsprechende stärkere Gewichtung ausgeglichen werden. Die zuverlässige Erkennung von Ecken ist jedoch auch mit aufwendigeren Verfahren [SB95, För86, DH97, FT98, LCH98, Ebn98] schwierig, sodaß von einer starken Spezialisierung auf das Thema zugunsten des hierarchischen Teile-Ansatzes abgesehen wird.

6.1.3 Flächenmerkmale

Die Berechnung von Flächenmerkmalen beruht auf einer Skelettierung der Eingabebilder. Das Ergebnis der Skelettierung sind die Skelettlinien, welche die Mit-

te von Flächenumrandungen repräsentieren. Da Punkte auf Skelettlinien etwa so häufig vorkommen wie Kantenpunkte, harmonisieren beide Beschreibungen gut miteinander, sodaß auf zusätzliche Normierungen verzichtet werden kann.

Eine gängige algorithmische Deutung der *Mitte von Flächenumrandungen* besteht darin, zunächst ein Binärbild zu erzeugen, in dem alle Kanten den Wert Null und alle Flächen den Wert Eins haben. Durch wiederholte Erosion des Binärbildes werden die Flächen dann bis auf einen Durchmesser von einem Pixel zu verdünnen. Um den hohen Zeitaufwand einer wiederholten Filterung des kompletten Bildes über mehrere Erosionsschritte zu vermeiden, wird hier allerdings auf eine alternative Deutung zurückgegriffen.

Abstandstransformation

Ein genaueres Kriterium für einen Punkt auf einer Skelettlinie ist, daß sich um den Punkt ein Kreis ziehen läßt, der vollständig innerhalb der Fläche liegt und den Rand der Fläche an mindestens zwei Punkten berührt, die den gleichen Abstand vom Mittelpunkt haben. Um dieses Kriterium umzusetzen, müssen die Abstände aller Punkte von der nächsten Kante berechnet werden, was unter dem Begriff *Abstandstransformation* zusammengefaßt wird.

Eine einfache, aber sehr präzise Methode der Berechnung besteht darin, alle Punkte des Bildes zu durchlaufen, dabei für jeden Punkt den Abstand zu allen Kantenpunkten zu bestimmen und jeweils das Minimum zu speichern. Von der Rechenzeit liegt diese Methode je nach Implementierung, Bildgröße und Rechner immer noch in der Größenordnung mehrerer 10 Minuten. Daher wird hier ein Verfahren eingesetzt, bei dem für jede im vorigen Schritt extrahierte Kontur eine Umgebung bestimmt wird, die ausgehend von den bloßen Kanten iterativ vergrößert wird. Die Abstandstransformation wird jeweils in den neu hinzugekommenen Randbereichen der Kantenumgebung durchgeführt. Dies hat den Vorteil, daß die Minimierung des Abstands zu einer Kante lokal erfolgen kann, da immer eine Zuordnung zu einem Kantenpunkt besteht. Der Aufwand wird dadurch erheblich reduziert.

In jedem Schritt wird eine Menge

$$E_r = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n_E}\}$$

von n_E Randpunkten behandelt, die an eine bereits behandelte Kantenumgebung mit einem maximalen Kantenabstand von $r - \partial r$ angrenzt. Dabei bedeutet *angrenzen*, daß alle Randpunkte mindestens einen Nachbarpunkt besitzen, für den der Abstand von der nächsten Kante bereits berechnet wurde, oder selbst auf so einem Punkt liegen. Die Randpunkte $E_{r=0}$ können ohne Sonderfall mit den Kantenpunkten selbst initialisiert werden. Der Abstand eines Punktes von der nächsten Kante wird in einer Abstandsmatrix $I_{\mathfrak{D}}$ gespeichert. Für Kantenpunkte $\boldsymbol{y}$ wird die Abstandsmatrix gemäß

$$I_{\mathfrak{D}}(\boldsymbol{y}) = 0$$

vorbelegt. Die übrigen Punkte werden durch ein spezielles Symbol als 'undefiniert' markiert. Dazu bietet sich beispielsweise die Bildgröße n_x erhöht um 1

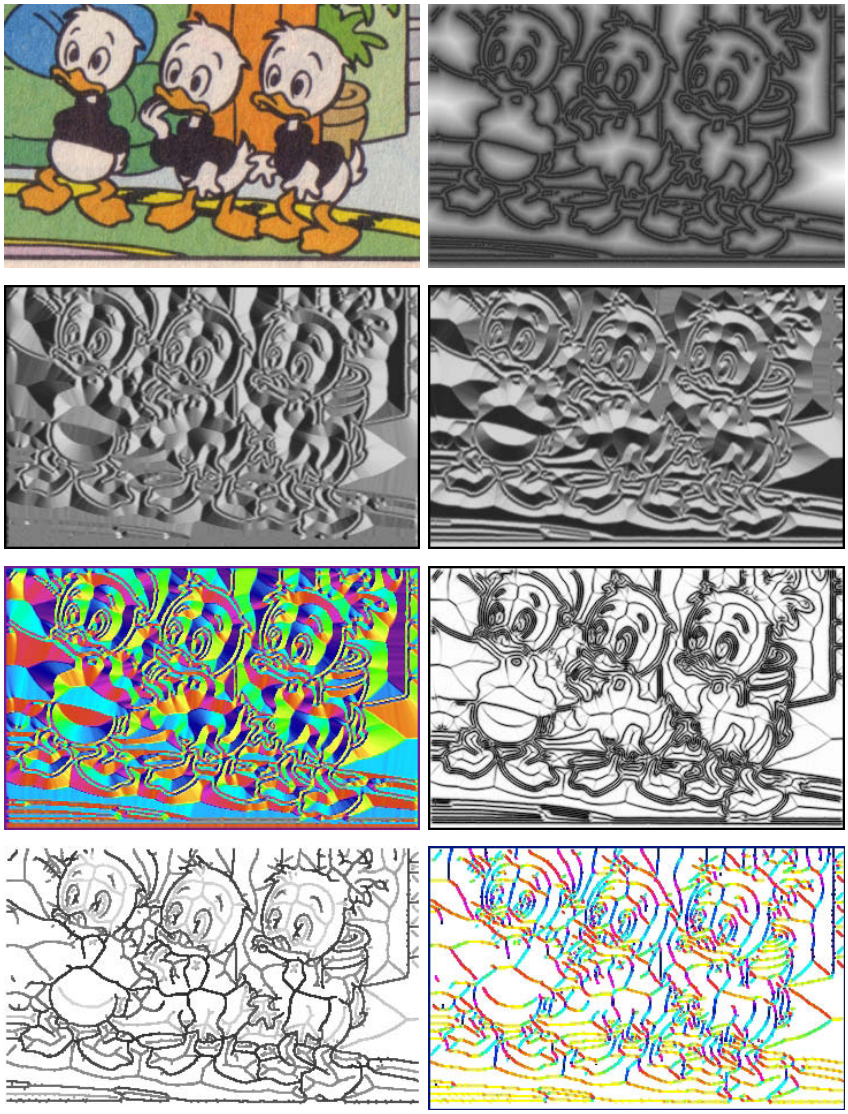


Abbildung 6.5: **Abstandstransformation und Skelettierung.** Der Reihe nach: 1. Originalbild [Dis]. 2. Abstandstransformation: Je heller, desto größer der Abstand von einer Kante. 3. Horizontale Änderung des Abstands. 4. Vertikale Änderung des Abstands. 5. Richtung der Abstandsänderung als Farbe kodiert. 6. Betrag der Abstandsänderung. Je heller, desto größer die Änderung. Der Betrag ist minimal an Kanten und Skelettlinien. 7. Intensität an Skelettpunkten: Je heller die Skelettlinie, desto heller sind die Punkte im Originalbild. Die Skelettlinien sind dicker gezeichnet, um ein besseres Druckbild zu erzeugen. 8. Richtung von Skelettlinien farblich kodiert.

an, da keine Abstände auftreten können, die größer als die Anzahl der Punkte des Bildes sind. Die Koordinaten der den Bildpunkten zugeordneten nächsten Kantenpunkte werden in den Matrizen I_h und I_v für die horizontale bzw. vertikale Komponente gespeichert. Diese werden für Kantenpunkte $\boldsymbol{y}$ ebenfalls mit den entsprechenden Koordinaten initialisiert, sodaß

$$(I_h(\boldsymbol{y}), I_v(\boldsymbol{y})) = \boldsymbol{y}.$$

Nicht initialisierte Werte spielen keine Rolle, da sie nie ausgelesen werden.

Abbildung 6.4 zeigt nun beispielhaft den Ablauf der Abstandstransformation für die Iteration $r = 3$. Zunächst wird ein Randpunkt $\boldsymbol{x} \in E_r$ ausgewählt. Dieser besitzt acht Nachbarpunkte $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \ldots, \boldsymbol{z}_8$. Für mindestens einen Nachbarn wurde die Abstandstransformation bereits durchgeführt. Für jeden Nachbarn $\boldsymbol{z}$ wird nun der Abstand zur nächsten Kante

$$r_z = I_{\mathfrak{D}}(\boldsymbol{z})$$

bestimmt. Falls r_z definiert ist, d.h. dem Punkt $\boldsymbol{z}$ bereits ein Kantenpunkt $\boldsymbol{y}$ zugeordnet wurde, wird geprüft, ob dieser Kantenpunkt auch für den Randpunkt $\boldsymbol{x}$ der nächste Nachbar ist. Dazu wird die Koordinate des $\boldsymbol{z}$ zugeordneten Kantenpunktes

$$\boldsymbol{y} = (I_h(\boldsymbol{z}), I_v(\boldsymbol{z}))$$

ausgelesen und der euklidische Abstand

$$r_y = \|\boldsymbol{x} - \boldsymbol{y}\| \tag{6.2}$$

zwischen dem Randpunkt $\boldsymbol{x}$ und dem Kantenpunkt $\boldsymbol{y}$ des Nachbarn berechnet. Falls $\boldsymbol{x}$ noch kein Kantenpunkt zugeordnet wurde oder ein bereits zugeordneter Kantenpunkt einen größeren Abstand hat, d.h.

$$I_{\mathfrak{D}}(\boldsymbol{x}) = \text{undef. oder } I_{\mathfrak{D}}(\boldsymbol{x}) > r_y,$$

wird $\boldsymbol{y}$ als nächster Kantenpunkt übernommen. Dazu wird für $\boldsymbol{x}$ der Kantenabstand aktualisiert, sodaß

$$I_{\mathfrak{D}}(\boldsymbol{x}) = r_y,$$

und die Kantenkoordinate gespeichert, wodurch

$$(I_h(\boldsymbol{x}), I_v(\boldsymbol{x})) = \boldsymbol{y}.$$

Das Durchlaufen aller Nachbarn eines Randpunktes $\boldsymbol{x}$ liefert schließlich den minimalen Kantenabstand $r_x = I_{\mathfrak{D}}(\boldsymbol{x})$.

Als nächstes werden die Randpunkte $E_{r+\partial r}$ für die nächste Iteration $r + \partial r$ bestimmt. Die Liste der Randpunkte ist zunächst leer, d.h.

$$E_{r+\partial r} = \{\}.$$

Dabei kann aufgrund der Auswahl von Randpunkten in einem quadratischen Raster über die direkte Nachbarschaft zu bereits behandelten Rändern mit einem geringeren Kantenabstand der Widerspruch auftreten, daß der minimale ermittelte Kantenabstand über den Rand hinausragt, d.h.

$$r_x > r. \tag{6.3}$$

In dem Fall wird das Rechenergebnis verworfen und $\boldsymbol{x}$ für die nächste Iteration gespeichert, d.h.

$$I_{\mathfrak{D}}(\boldsymbol{x}) = \text{undef.} \tag{6.4}$$

und

$$E_{r+\partial r} = E_{r+\partial r} + \{\boldsymbol{x}\}. \tag{6.5}$$

Andernfalls können alle Punkte $\boldsymbol{z}$ in der 8-Nachbarschaft zum neuen Rand hinzugefügt werden, sofern für sie noch kein Abstand berechnet wurde:

$$E_{r+\partial r} = +\{\text{alle } \boldsymbol{z} | \|\boldsymbol{z} - \boldsymbol{x}\| \leq \sqrt(2), \boldsymbol{x} \in E_r, I_{\mathfrak{D}}(\boldsymbol{z}) = \text{undef.}\} \tag{6.6}$$

Wenn alle Randpunkte $\boldsymbol{x} \in E_r$ behandelt wurden, wird die nächste Iteration für alle Randpunkte in $E_{r+\partial r}$ gestartet. Der Algorithmus terminiert, wenn alle Bildpunkte behandelt wurden, d.h. keine Randpunkte mehr vorliegen

$$n_E = 0 \text{ bzw. } E_{r+\partial r} = \{\}.$$

Die auf den ersten Blick umständlich wirkenden Gleichungen 6.3 bis 6.5 hängen eng mit der Wahl der Randbreite ∂r zusammen. Sie sind für die numerische Stabilität des Verfahrens von entscheidender Bedeutung, da sie die verschiedenen Abstandsnormen aus Gleichung 6.2 und Gleichung 6.6 in Einklang bringen. Der Manhattan-Abstand aus Gleichung 6.6 ergibt sich dabei aus der 8-Nachbarschaft in einem quadratischen Punktraster. Die Gründe dafür sind ausschließlich technischer Natur. Der euklidische Abstand in Gleichung 6.2 und die Einhaltung der Randbreite ∂r in Gleichung 6.3 stellt dagegen sicher, daß die Kantenumgebungen stets in alle Richtungen gleich wachsen. Dies ist wichtig, da die Umgebungen bei der Manhattan-Norm beispielsweise in diagonaler Richtung um den Faktor $\sqrt{2}$ schneller wachsen als in senkrechter oder horizontaler Richtung. Wenn dann eine diagonal gewachsene und eine vertikal oder horizontal gewachsene Umgebung aufeinander stoßen, liegen die Berührungspunkte nicht in der Mitte zwischen den Kanten, sondern im Verhältnis $\sqrt{2} : 1$. Dies führt zu falschen Ergebnissen bei der anschließenden Skelettierung. Die Randbreite sollte daher auf einen Wert

$$\partial r < \sqrt{2}$$

eingestellt werden, damit sich ein gleichmäßiges Wachstum ergibt. In dieser Arbeit wird $\partial r = 0{,}7$ gewählt, da zu kleine Werte keinen numerischen Vorteil bieten, aber die Berechnung aufgrund der Rückschritte in den Gleichungen 6.4 und 6.5 verlangsamen. Abbildung 6.5 rechts oben zeigt eine fertig berechnete Abstandsmatrix.

Skelettierung

Mit Hilfe der Abstandsmatrix $I_{\mathfrak{D}}$ lassen sich nun die Skelettlinien finden. Nach der oben genannten Deutung von Skelettpunkten als Kreismittelpunkte müssen nun Punkte gefunden werden, die zu mindestens zwei Kantenpunkten den gleichen Abstand besitzen. Da jedem Punkt jedoch nur ein Kantenpunkt zugeordnet wurde, müssen dazu die Kantenpunkte zweier benachbarter Punkte untersucht werden, was zu einer kleinen Ungenauigkeit führt. Die Abstände zu den zwei Kantenpunkten muß nun für Skelettpunkte gleich sein. Unter Berücksichtigung der geringeren Genauigkeit, die hier nur im Pixelbereich liegt, ist diese Bedingung jedoch auch erfüllt, wenn die beiden zugeordneten Kantenpunkte selbst benachbart sind und damit keine separaten Berührungspunkte für den Kreis darstellen. Darüberhinaus ergeben sich aus der geringen Genauigkeit auch geringe Abweichungen im Kantenabstand.

Eine Lösung dieses Problems besteht darin, nach Punkten zu suchen, die lokale Maxima bezüglich des Kantenabstands zweier entgegengesetzt benachbarter Punkte sind. In diesem Fall sind die Nachbarpunkte immer unterschiedlichen Kanten zugeordnet. Um die numerische Stabilität zu erhöhen, müssen noch zwei weitere Bedingungen erfüllt sein. Zum einen sollte sich der Kantenabstand auf dem Skelettpunkt nicht ändern, d.h. die Ableitung des Abstands sollte ausreichend klein sein. Zum anderen sollte der Kantenabstand einen gewissen Mindestwert haben, da sich nahe an Kanten keine stabilen Ergebnisse erzielen lassen. Für einen Skelettpunkt $\boldsymbol{x} = (x, y)$ ergeben sich daher die Bedingungen

$$I_{\mathfrak{D}}(x-1, y-1) < I_{\mathfrak{D}}(\boldsymbol{x}) > I_{\mathfrak{D}}(x+1, y+1) \tag{6.7}$$

$$\vee \quad I_{\mathfrak{D}}(x-1, y) < I_{\mathfrak{D}}(\boldsymbol{x}) > I_{\mathfrak{D}}(x+1, y) \tag{6.8}$$

$$\vee \quad I_{\mathfrak{D}}(x, y-1) < I_{\mathfrak{D}}(\boldsymbol{x}) > I_{\mathfrak{D}}(x, y+1) \tag{6.9}$$

$$\vee \quad I_{\mathfrak{D}}(x+1, y-1) < I_{\mathfrak{D}}(\boldsymbol{x}) > I_{\mathfrak{D}}(x-1, y+1) \tag{6.10}$$

(lokales Maximum des Kantenabstandes) und

$$\|\nabla I_{\mathfrak{D}}(\boldsymbol{x})\| < \vartheta_{\nabla} \tag{6.11}$$

$$\vee \quad I_{\mathfrak{D}}(\boldsymbol{x}) > \vartheta_r \tag{6.12}$$

(kleine Ableitung und hoher Kantenabstand). Die Schwellen ϑ_1 und ϑ_2 wurden experimentell an das verwendete Bildmaterial angepaßt. Anschließend wird noch ein Nachverarbeitungsschritt durchgeführt, der hin und wieder auftretende isolierte Punkte entfernt, die keiner Skelettlinie angehören. Abbildung 6.5 zeigt ein Beispiel für eine Skelettierung. Die Bedeutung der Bedingungen 6.7 bis 6.10 ist aus dem Bild rechts oben ersichtlich. Die Bilder in der zweiten und dritten Reihe stellen den Gradienten $\nabla I_{\mathfrak{D}}$ der Abstandsmatrix dar. Die Bedeutung von Bedingung 6.11 und 6.12 demonstriert insbesondere das rechte Bild in der dritten Reihe: Der Betrag des Gradienten ist minimal für Kanten und für Skelettlinien. Um Kanten von der Erkennung als Skelettlinien zu schützen, wurde Bedingung 6.12 eingeführt.

Zu jedem detektierten Skelettpunkt wird ein dreidimensionaler Deskriptor erzeugt. Dieser speichert die Intensität des entsprechenden Pixels im Original-

Messung	Anzahl Merkmale	Mittlere Richtung	Standardabweichung
1	286	278°	5,436
2	430	270°	5,508
3	51	141°	12,456
4	236	269°	3,96
5	250	90°	3,564
6	177	292°	3,96
7	178	323°	6,768
8	89	220°	5,94

Tabelle 6.1: Rauschen der Gradientenrichtung für verschiedene gerade Kanten

bild, den Kantenabstand $I_{\mathfrak{D}}(\boldsymbol{x}$ als Maß für die Flächengröße und die Ausrichtung der Skelettlinie als Maß für die Ausrichtung der Fläche. Um die Richtung der Skelettlinie zu bestimmen, wird ein 5×5 Pixel großes Fenster um jeden Skelettpunkt gelegt und der Winkel der Hauptträgheitsachse aller Skelettpunkte in diesem Fenster berechnet. Die letzte Reihe aus Abbildung 6.5 zeigt die Intensität und Richtung der Deskriptoren. Die Berechnung der Skelettierung einschließlich der Deskriptoren dauert für das angegebene Beispiel etwa 1s auf einem AMD Athlon XP mit 2GHz.

6.1.4 Quantisierung der Deskriptoren

Bei der Bestimmung der Anzahl an Quantisierungsstufen $\zeta_1, \zeta_2, \ldots$ für die Diskretisierung der Deskriptoren spielen drei Dinge eine Rolle:

- Das Rauschen des Bildes und Rechenungenauigkeiten bei der Merkmalsextraktion
- Die zeichnerische Freiheit bei karikaturenhaften Abbildung realer Vorlagen
- Die Beeinflussung nachfolgender Verarbeitungsschritte bezüglich Trennschärfe und Verallgemeinerbarkeit

Zunächst werden daher die Eigenschaften der Stichprobe untersucht.

Eigenschaften des Bildmaterials

Das Merkmalsrauschen aus dem ersten Punkt ist einfach zu bestimmen. Zur Vermessung des Rauschens der Gradientenrichtung bei Kantenmerkmalen werden kleine Bildausschnitte ausgewählt, die gerade Kanten enthalten. Auf diesen Ausschnitten wird eine Merkmalsextraktion durchgeführt. Von den resultierenden Merkmalen werden dann manuell die ausgewählt, die auf den geraden Kanten liegen. Für jede gerade Kante wird dann das Histogramm über die Gradientenrichtung berechnet. Abbildung 6.6 zeigt ein Beispiel. Für eine Meßreihe mit

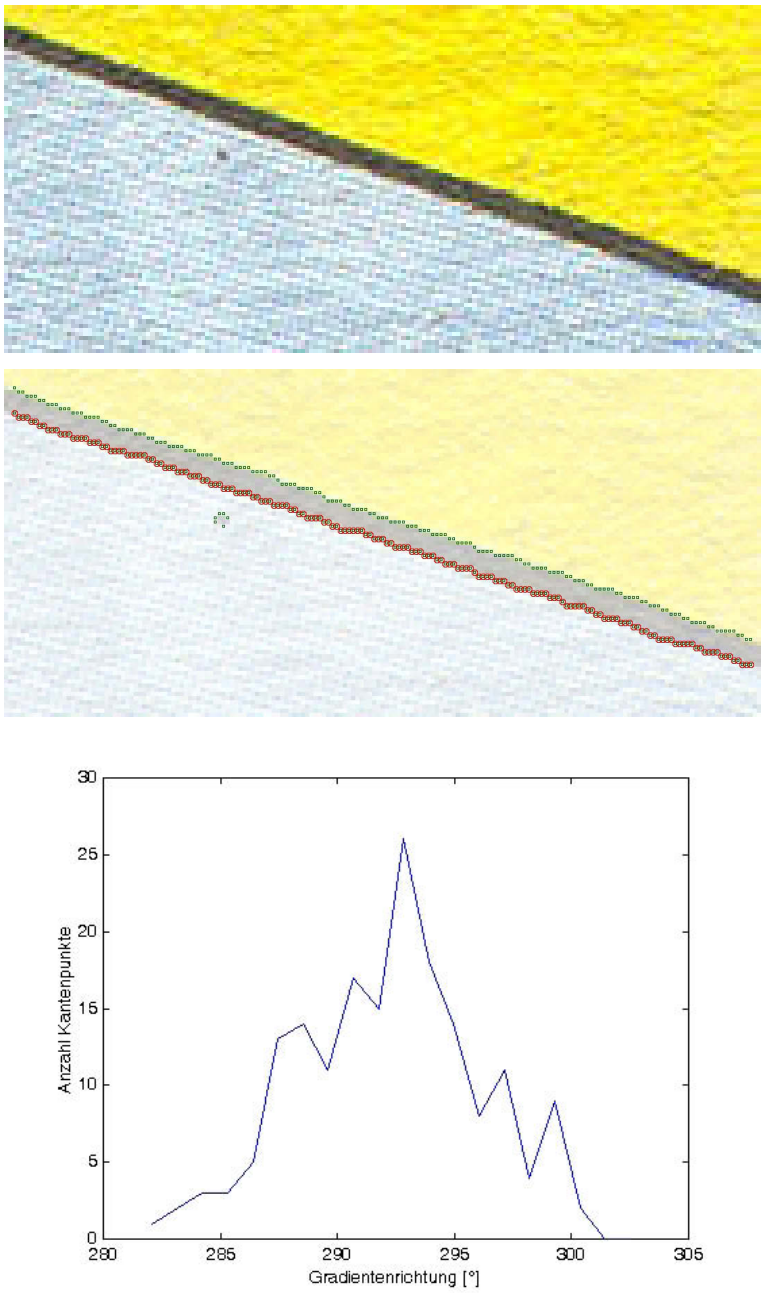


Abbildung 6.6: **Rauschen von Kantenmerkmalen.** Oben: Originalbild, 181×83 Pixel. Mitte: Extrahierte Kantenpunkte sind grün dargestellt. Die zur Vermessung benutzten Punkte sind rot umrandet. Unten: Histogramm über die Gradientenrichtungen der markierten Punkte. Die mittlere Richtung beträgt 292 Grad bei einer Standardabweichung von 3,96.

Messung	Anzahl Merkmale	Mittlere Richtung	Standardabweichung
9.	414	186°	12,456
10.	359	204°	11,664
11.	185	176°	11,484
12.	143	178°	12,24
13.	175	325°	5,976
14.	122	316°	10,044

Tabelle 6.2: Rauschen der Orientierung von Skelettlinien

Messung	Anzahl Merkmale	Mittlere Intensität	Standardabweichung
15.	414	0,87	0,0162
16.	359	0,87	0,0165
17.	185	0,82	0,0178
18.	143	0,57	0,0365
19.	175	0,79	0,0174
20.	122	0,48	0,0208

Tabelle 6.3: **Rauschen der Intensität von Pixeln auf Skelettlinien.** Die Intensitätswerte sind auf das Intervall von 0 bis 1 normiert.

acht Ausschnitten aus verschiedenen Scans der Cartoon-Datenbank ergeben sich die in Tabelle 6.1 dargestellten Ergebnisse. Es zeigt sich, daß die Standardabweichung um den Wert 5,9 schwankt. Dabei scheint das Bild in Messung 3 ein Ausreißer zu sein. Für die willkürliche, aber vernünftige Forderung, daß etwa 95 Prozent aller Merkmale richtig erkannt werden sollen, müssen Gradientenrichtungen aus dem Wertebereich von $\pm 2 \cdot$ Standardabweichung auf die drei benachbarten Quantisierungsintervalle abgebildet werden, die der Klassifikator zur Erkennung eines Merkmals überprüft. Für die mittlere Standardabweichung von 5,9 ergibt sich daraus eine Quantisierungsstufe der Größe $4 \cdot 5{,}9/3 \approx 7{,}9$. Dies sind etwa 45 Quantisierungsstufen bezogen auf die vollen 360 Grad. Für die schlechteste Messung ergeben sich immer noch 22 Stufen.

Die Quantisierungsstufen für die Orientierung von Skelettlinien und die Intensität von Bildpunkten auf Skelettlinien werden genauso ermittelt. Die Ergebnisse zeigen die Tabellen 6.2 und 6.3. Für das schlechteste Ergebnis bei der Orientierung von Skelettlinien ergeben sich 22 Stufen, für das schlechteste Ergebnis der Intensität 36 Stufen.

Die Pixelintensität schwankt nicht nur über Punkte innerhalb homogener Bereiche. Wie die in Tabelle B.2 (Anhang) dargestellten Messungen zeigen, schwanken die Mittelwerte homogener Bereiche für gleiche Farben zusätzlich noch über verschiedene eingescannte Comic-Seiten. Die Standardabweichungen der vermessenen Schwankungen liegen für die wichtigsten Farben Weiß, Orange

Messung	Anzahl Merkmale	Mittlere Richtung	Standardabweichung
21.	504	264°	14,652
22.	295	170°	9,396
23.	207	280°	4,86
24.	208	108°	7,38
25.	169	138°	10,296
26.	242	256°	7,596

Tabelle 6.4: Streuung der Gradientenrichtung aufgrund von zeichnerischer Freiheit.

und Schwarz bei 0,036 bzw. 0,034 und 0,030 (bezogen auf den Wertebereich 0 bis 1). Daraus ergeben sich nur noch 21 bis 25 Quantisierungsstufen.

Wie Abbildung 6.7 zeigt, sehen die Ergebnisse für den Kantenabstand der Skelettpunkte anders aus. Da die Abstandstransformation jeden Bildpunkt einem eindeutigen Kantenpunkt zuordnet, werden Skelettpunkte, die definitionsgemäß den gleichen Abstand zu mehreren Kantenpunkten besitzen, einem Kantenpunkt willkürlich zugeordnet. Aufgrund des am Pixelrasters orientierten Algorithmus geschieht dies nur mit Pixelgenauigkeit. Das Histogramm über den Kantenabstand für eine Skelettlinie innerhalb eines Rechtecks ist daher bimodal und zeigt für beide Seiten des Rechtecks eine Häufung. Die Quantisierung kann daher auch höchstens mit Pixelgenauigkeit geschehen.

Neben den Ungenauigkeiten, die aus der Bildreproduktion und -verarbeitung resultieren, haben auch die zeichnerischen Stilmittel einen Einfluß auf die Quantisierung von Merkmalen. Die meisten Besonderheiten der Cartoon-Stichprobe lassen sich schwer quantisieren und können nur als Teil der Erscheinung hingenommen werden. Speziell bei der Bestimmung von Kantenorientierungen fällt jedoch auf, daß in den verwendeten Comics aus dramaturgischen Gründen Kanten oft krumm gezeichnet werden, die eigentlich gerade sein müßten (siehe Abbildung 6.8). Diese zeichnerischen Variationen lassen sich ebenfalls mit der oben beschriebenen Methode quantisieren. Die Ergebnisse zeigt Tabelle 6.4. Für Messung 21, die auf dem in Abbildung 6.8 gezeigten Ausschnitt beruht und ein extremes Beispiel ist, ergeben sich 18 Quantisierungsstufen. Für das zweithöchste Ergebnis (Messung 25) ergeben sich 26 Stufen. Dies liegt am unteren Ende der Genauigkeit für unverzerrt gezeichnete Geraden und stellt damit einen geringeren Einfluß dar, als erwartet.

Zusammengefaßt ergeben die beschriebenen Messungen, daß das vorliegende Bildmaterial die Unterscheidung von

- mindestens 18 Kantenrichtungen,
- mindestens 21 Intensitätswerte von Skelettpunkten,
- mindestens 22 Orientierungen von Skelettlinien

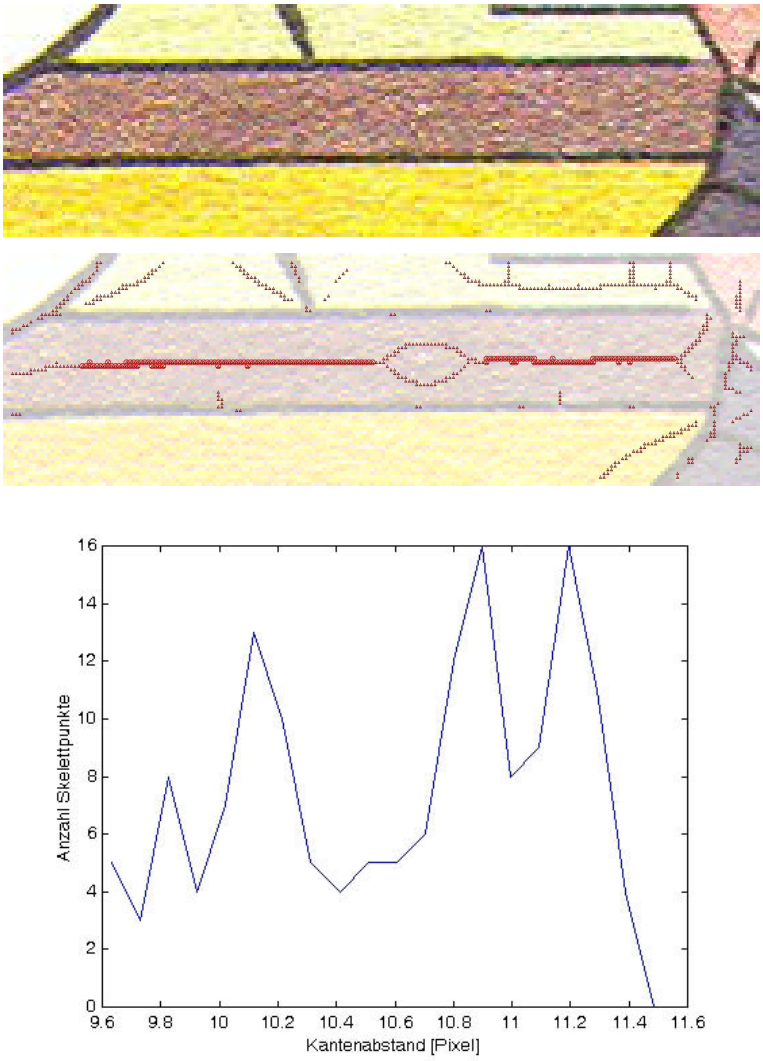


Abbildung 6.7: **Rauschen des Kantenabstands von Skelettpunkten.** Oben: Originalbild, 206×63 Pixel. Mitte: Die extrahierten Skelettpunkte sind dünn rot markiert. Die dicke rote Markierung zeigt die zur Messung benutzten Merkmale in der Mitte des braunen Rechtecks. Der Kantenabstand ist theoretisch konstant über die Breite des Rechtecks. Unten: Histogramm über den Kantenabstand. Es zeigen sich zwei Maxima im Abstand von genau einem Pixel. Diese kommen durch Rechenungenauigkeiten zustande, wenn ein Skelettpunkt entweder der Ober- oder der Unterkante des Rechtecks zugeordnet werden muß. Die Histogramme variieren etwas, woraus sich hier weitere, nicht systematisch auftretende hohe Maxima ergeben.

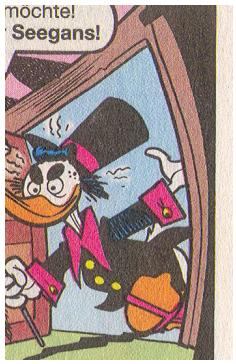


Abbildung 6.8: **Übertriebene Perspektive als zeichnerisches Stilmittel.** Der Türrahmen ist perspektivisch stark gekrümmt. Diese Krümmung wirkt als zusätzliches Rauschen auf der Gradientenrichtung.

- und pixelgenauen Kantenabständen

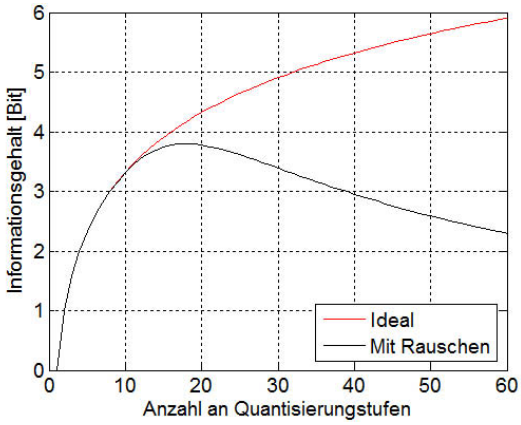
erlaubt.

Optimierung der Quantisierungsfeinheit

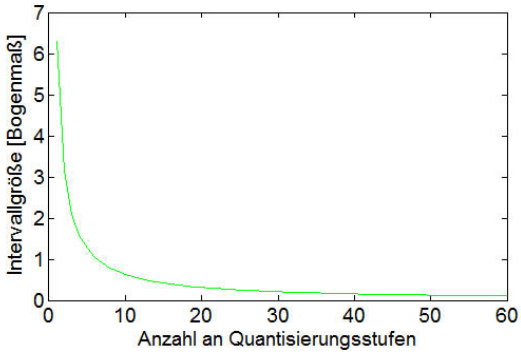
Nachdem geklärt wurde, wieviel Information in dem Datenmaterial vorliegt, stellt sich nun die Frage, wie diese am besten ausgewertet wird. Für die Wahl der Quantisierungsstufen ζ ist dabei nicht nur der Informationsgehalt der extrahierten Merkmale von Bedeutung, sondern auch die Toleranz des Verfahrens gegenüber Rauschen und der Speicheraufwand des Modells.

Auf der einen Seite sind feine Quantisierungsstufen günstig für die Objekterkennung, da die im Bildmaterial vorliegende Information vollständig in die Berechnungen eingeht. Andererseits steigt bei feinen Quantisierungstufen die Gefahr, daß einzelne Merkmale nicht erkannt werden, da die Meßwerte aufgrund von Rauschen in die falsche Quantisierungsstufe eingeordnet werden. Die höhere Anzahl verschiedener Merkmalsausprägungen bei einer feinen Quantisierung der Meßwerte bedeutet zudem einen höheren Speicherbedarf für die Tabelle LUT, in der jede auftretende Merkmalsausprägung einer Menge von abstrakteren Knoten des Modells zugeordnet wird (vgl. Abschnitt 5.1). Eine übermäßig feine Quantisierung erhöht außerdem den Aufwand zur Speicherung der Hypothesen, da viele redundante Fallunterscheidungen vorgenommen werden. Für eine grobe Quantisierung ist eine bessere Toleranz gegenüber Rauschen und ein geringerer Speicheraufwand zu erwarten. Andererseits wird die im Bildmaterial vorhandene Information schlecht genutzt. Um herauszufinden, bei welcher Anzahl an Quantisierungsstufen sich die optimale Informationsausnutzung ergibt, wird nun die Erkennung eines Objekts an einem einzelnen Merkmal simuliert.

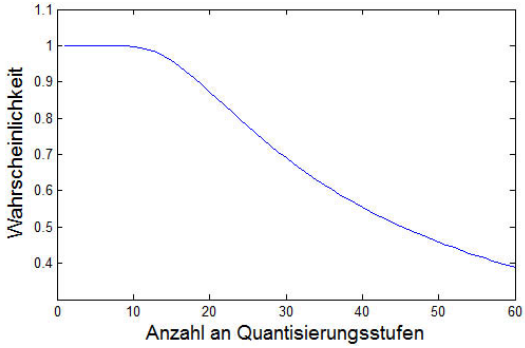
Bei der Unterscheidung von ζ Quantisierungsstufen hat eine Messung idealerweise einen Informationsgehalt von $\log_2 \zeta$ Bits (s. Abb. 6.9a). Dieser wird



a) Informationsgehalt einer Messung

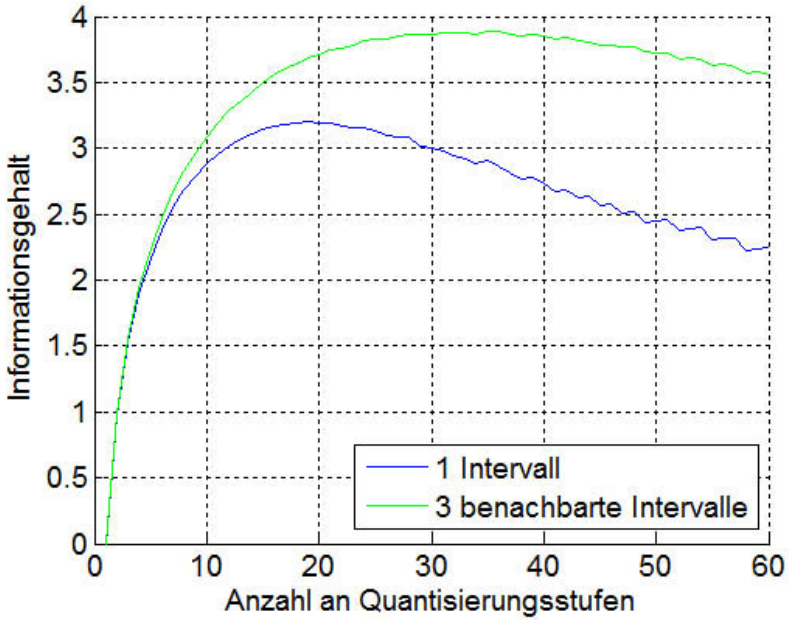


b) Intervallgröße abhängig von der Quantisierungsfeinheit

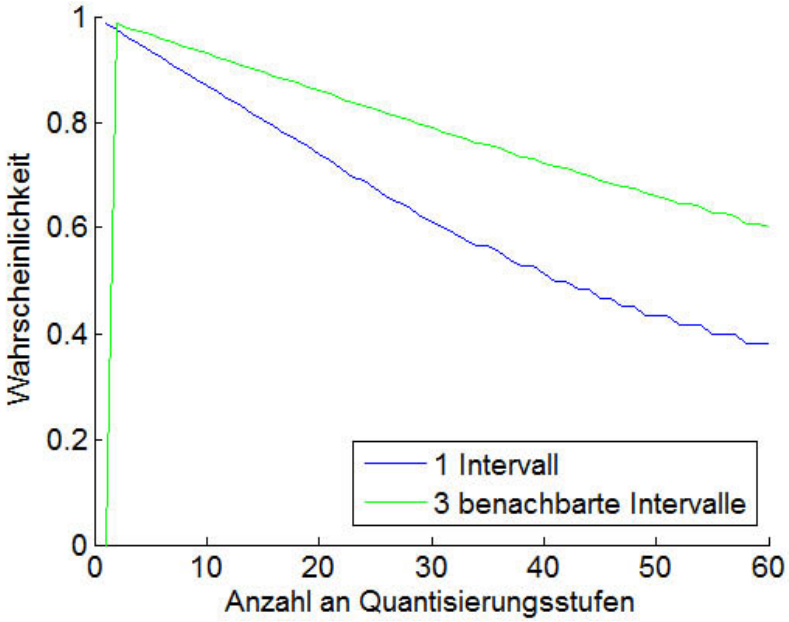


c) Wahrscheinlichkeit des Auftretens eines Meßwerts im richtigen Intervall

Abbildung 6.9: **Simulation der Informationsausnutzung durch verschiedene Quantisierungen eines Merkmals.** a) Theoretischer (rote Linie) und praktisch erzielbarer (schwarze Linie) Informationsgehalt einer Messung. Der praktisch erzielbare Informationsgehalt wird dadurch begrenzt, daß aufgrund von Rauschen bei feinen Quantisierungsstufen die Wahrscheinlichkeit sinkt (c), daß ein Meßwert in das zu erwartende Werteintervall fällt (b).



a) Informationsgehalt einer Messung



b) Wahrscheinlichkeit des Auftretens eines Meßwerts im modellierten Intervall

Abbildung 6.10: **Simulation der Informationsausnutzung bei Betrachtung eines einzelnen bzw. je 3 benachbarter Werteintervalle.** a) Erreichter Informationsgehalt für verschiedene Quantisierungsstufen. b) Wahrscheinlichkeit der Erfassung des Meßwerts der korrekten Klasse.

jedoch durch Rauschen verringert. Basierend auf den in Tabelle 6.1 dargestellten Meßergebnissen zur Gradientenrichtung wird ein normalverteiltes Rauschen mit einer Standardabweichung von $\sigma = 5{,}9$ angenommen. Durch die Diskretisierung der Gradientenrichtung ergeben sich ζ Winkelintervalle, in die der Meßwert fallen kann. Abbildung 6.9b zeigt die Intervallgröße $2\pi/\zeta$ über der Quantisierungsfeinheit. Es wird vereinfachend angenommen, daß die Intervalle so liegen, daß der Mittelwert der Gaußverteilung auf ein Intervall zentriert ist, welches den richtigen Meßwert repräsentiert. Die Wahrscheinlichkeit p_{richtig}, daß dieses Intervall tatsächlich den tatsächlichen Wert enthält, ergibt sich durch die Integration der Normalverteilung innerhalb des Intervalls, d.h.

$$p_{\text{richtig}} = \int_{-\zeta/2}^{\zeta/2} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{d}{\sigma}\right)^2} \mathrm{d}d$$

(s. Abb. 6.9c). Der maximal erreichbare Informationsgehalt ergibt sich aus dem Produkt des idealen Informationsgehalts und der Wahrscheinlichkeit, mit einem Quantisierungsintervall die tatsächliche Merkmalsausprägung zu messen. Wie Abbildung 6.9a zeigt, entspricht das Ergebnis für große Quantisierungsstufen weitgehend dem idealen Informationsgehalt. Werden die Quantisierungsstufen allerdings verkleinert, überwiegt ab einem bestimmten Punkt die steigende Fehlerrate. Es ergibt sich ein Maximum bei $\zeta = 18$ Quantisierungsstufen mit einem Informationsgehalt von 3,79 Bits, welches 0,38 Bits unter dem Ideal liegt.

Für die Objekterkennung sind also sowohl zu feine als auch zu grobe Quantisierungsstufen nachteilig. Wie sehen nun die Ergebnisse aus, wenn wie in Gleichung 5.1 vorgeschlagen auch die Intervalle neben dem korrekten Intervall berücksichtigt werden?

Um diese Frage zu beantworten, wird eine genauere Simulation durchgeführt. Dabei soll zudem berücksichtigt werden, daß verschiedene Werteintervalle unterschiedliche Klassen von Objekten anzeigen können. Es wird wieder davon ausgegangen, daß ein kontinuierlicher Merkmalswert durch normalverteiltes Rauschen gestört ist. Die Wahrscheinlichkeit p_{fd}, durch das Quantisierungsintervall d einen bestimmten Merkmalswert f zu erfassen, wird durch die Integration der Gaußfunktion über das Intervall berechnet, d.h.

$$p_{fd} = p_f \int_{-\zeta/2}^{\zeta/2} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{d-f}{\sigma}\right)^2} \mathrm{d}d.$$

Es wird außerdem angenommen, daß alle Merkmalswerte mit der gleichen Wahrscheinlichkeit p_f auftreten. Die Wahrscheinlichkeit p_{richtig}, daß ein Meßwert im Quantisierungsintervall d korrekt ist, d.h. der richtigen Objektklasse angehört, ergibt sich aus der Integration von p_{fd} über alle Merkmalswerte f innerhalb des Intervalls. Abbildung 6.10b (blaue Kurve) zeigt den Verlauf von p_{richtig} über der Quantisierungsfeinheit. Die Multiplikation der berechneten Wahrscheinlichkeit mit dem Informationsgehalt ergibt die in Abbildung 6.10a durch die blaue Kurve dargestellte Funktion. Der maximale Informationsgehalt ergibt sich für 19 Quantisierungsstufen mit 3,20 Bits. Der abweichende Informationsgehalt im

Vergleich zur vorigen Simulation resultiert aus der anderen Interpretation des korrekten Intervalls. Der Unterschied liegt darin, daß in der ersten Simulation von einem einzelnen korrekten Merkmalswert ausgegangen wird, auf den das Intervall zentriert ist. In der zweiten Simulation wird dagegen eine Klassifikationsentscheidung aufgrund eines ganzen Intervalls getroffen, sodaß alle Werte innerhalb des Intervalls als korrekt betrachtet werden.

Um bei der Klassifikation auch die benachbarten Intervalle zu berücksichtigen, wird p_{fd} nun über die Merkmalswerte f innerhalb von drei benachbarten Intervallen integriert. Bei der Berücksichtigung mehrerer Intervalle steigt die Wahrscheinlichkeit, durch eine Messung auch gestörte Merkmalswerte zu erfassen (s. Abb. 6.10b, grüne Kurve). Dies ermöglicht eine feinere Quantisierung der Merkmale. Wie Abbildung 6.10a zeigt, ergibt sich so ein maximaler Informationsgehalt von 3,89 Bits bei 35 Quantisierungsstufen.

Als nächstes wird untersucht, wie weit die Simulationsergebnisse für echtes Bildmaterial zutreffen. Aufgrund der zusätzlichen Verfahrensparameter und der in der Simulation nicht berücksichtigten Eigenschaften des Bildmaterials können hier Abweichungen auftreten.

Auswirkung der Quantisierung auf die Objekterkennung

Im folgenden wird untersucht, wie sich die Merkmalsquantisierung auf die Erkennung eines kleinen und einfachen Objekts bei Störeinflüssen auswirkt. Die Schwelle ϑ und die Ortstoleranz ς treten als freie Parameter auf und werden zur Quantisierungsfeinheit in Beziehung gesetzt. Neben dem Rauschen spielt auch die Variation der Stichprobe eine Rolle. Daher werden anschließend Experimente mit verschiedenen Stichprobenelementen durchgeführt. Als letztes wird geprüft, ob sich die Orientierung von Skelettlinien grundsätzlich anders verhält als die Kantenorientierung.

Um den Einfluß der Merkmalsquantisierung bei der Erkennung gestörter Bilder zu untersuchen, wird ein Stichprobenbild skaliert und mit Rauschen versehen. Das Bild in der Originalgröße von 181×158 Pixel wird mit 7 Skalierungsfaktoren zwischen 0,4 und 1,6 vervielfältigt. Auf jeder Größenstufe werden 5 Bilder erzeugt, auf die normalverteiltes Rauschen mit Standardabweichungen von 0,0 bis 0,2 addiert wird. Die Standardabweichung bezieht sich auf den normierten Dynamikumfang der Bildintensitäten von Null bis Eins. Einige dieser Testbilder sind in Abbildung 6.11 oben dargestellt. Die etwa 60 Kantenpunkte der linken Augenbraue des Originalbildes werden als Modell gespeichert. Eine graphische Darstellung zeigt Abbildung 6.11 in der Mitte. Das Modell enthält außer den Kantenpunkten nur einen Teileknoten, der die Ortstoleranz ς und die Schwelle ϑ enthält.

Die Parameter ς, ϑ und die Anzahl der Quantisierungsstufen werden systematisch variiert. Für die Ortstoleranz wird der Wertebereich von 3–15 Pixel untersucht. Da die Augenbraue eine Größe von etwa 17×19 Pixel hat, gehen bei einer Ortstoleranz von 15 Pixel kaum noch Geometrieinformationen in die Objekterkennung ein. Das Modell arbeitet dann praktisch als Bag-of-features. Für die Schwelle wird der Wertebereich von 20% bis 90% überprüft. Die Anzahl der

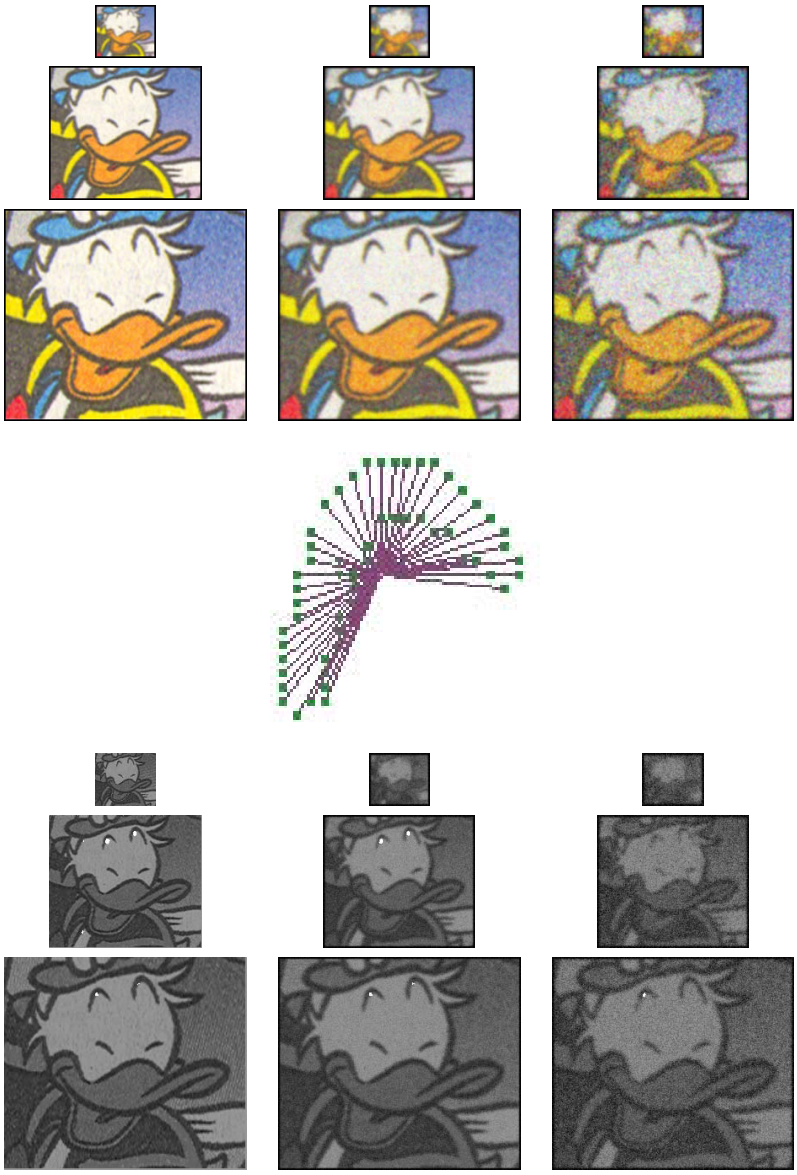


Abbildung 6.11: **Erkennung einer Augenbraue bei Skalierung und Rauschen.** Oben: Skalierung des Originalbildes mit Faktoren zwischen 0,4 und 1,6 und Addition von normalverteiltem Rauschen mit Standardabweichungungen zwischen 0,0 und 0,2 (bezüglich des auf $[0 \ldots 1]$ normierten Wertebereichs). Mitte: Modell zur Erkennung der linken Augenbraue des Stichprobenbildes. Ein Teileknoten faßt etwa 60 Kantenmerkmale zusammen. Die Merkmale entstammen dem gezeigten Stichprobenbild (unskaliert, ohne Rauschen). Unten: Trefferpositionen für die Parametrisierung $\vartheta = 0{,}5$ und $\varsigma = 5$ des Teileknotens (weiße Markierung). Es wurden 35 Kantenorientierungen unterschieden.

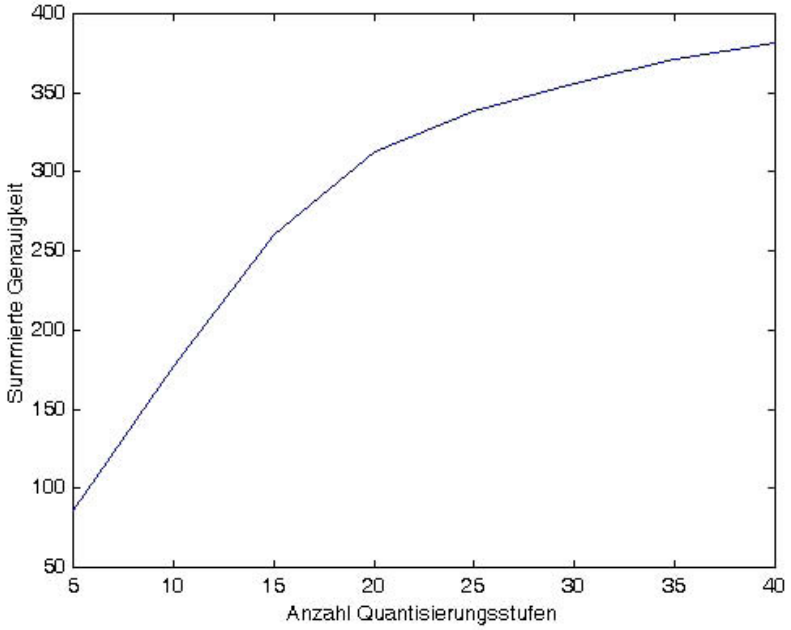


Abbildung 6.12: **Genauigkeit der Augenbrauenerkennung summiert über viele Messungen.** Der Plot zeigt für jede Quantisierungsstufe die Summe der Genauigkeiten für $\varsigma = 3 \ldots 15$ und $\vartheta = 20\% \ldots 90\%$.

Quantisierungsstufen wird zwischen 5 und 40 variiert, was innerhalb der in den vorigen Versuchen ermittelten Genauigkeit von höchstens 18–45 Stufen liegt. Die ausgewählten Wertebereiche decken damit den plausiblen Parameterraum vollständig ab.

Um richtige Treffer zu erkennen wird manuell ein rechteckiger Bereich von 17×19 Pixeln um die linken Augenbrauen in den Testbildern festgelegt. Treffer innerhalb des Bereichs zählen als richtig, andere als falsch. Auf dieser Basis wird für jede Parametereinstellung die Genauigkeit der Erkennung als Quotient aus richtigen Treffern und falschen Treffern berechnet.

Um einen Eindruck von den Ergebnissen der Objekterkennung zu vermitteln, sind in Abbildung 6.11 unten einige Trefferbilder dargestellt für $\varsigma = 5$ und 35 Quantisierungsstufen, eine der besten Einstellungen bei einer Schwelle von $\vartheta = 50\%$. Die erkannten Positionen sind weiß markiert. Abgesehen davon, daß oft auch die rechte Augenbraue erkannt wurde, ergeben sich keine falschen Treffer. In 23 der 35 Testbildern wird die Augenbraue erkannt. Bei weniger Quantisierungsstufen werden teilweise mehr Augenbrauen erkannt, allerdings steigt auch die Zahl falscher Treffer.

Die vollständigen Ergebnisse sind im Anhang in Tabelle B.3 angegeben. Um eine kompakte und von den Knotenparametern ς und ϑ unabhängige Darstellung zu erreichen, wird der dreidimensionale Parameterraum auf die Achse für die Quantisierungsfeinheit projeziert. Dazu werden für jede Anzahl an Quantisierungsstufen die Genauigkeiten aller getesteten Schwellen und Ortstoleranzen aufsummiert. Das Ergebnis ist in Abbildung 6.12 dargestellt. Es zeigt sich, daß die Genauigkeit der Objekterkennung unabhängig von den gewählten übrigen Parametern mit einer feineren Quantisierung steigt. Ab etwa 20 Quantisierungsstufen treten nur noch geringe Verbesserungen ein. Die Ergebnisse stimmen damit grob mit dem in Abbildung 6.10a dargestellten Simulationsergebnis überein, wobei das Maximum allerdings bei 40 statt bei 35 Quantisierungsstufen liegt. Da keine feineren Merkmalsunterteilungen untersucht wurden, ist unklar, ob sich dort wie von der Simulation vorhergesagt schlechtere Ergebnisse zeigen.

In den Daten zeigt sich zudem noch eine starke Abhängigkeit von den in dieser Darstellung vernachlässigten Knotenparametern. Die Abbildung 6.13 zeigt daher die Genauigkeit abhängig von der Quantisierungsfeinheit und der Ortstoleranz für drei ausgewählte Schwellen. Für die niedrige Schwelle von 50 Prozent ergibt sich ein Maximum bei einer geringen Ortstoleranz $\varsigma = 5$ Pixel und einer feinen Quantisierung von 40 Stufen. Mit steigender Ortstoleranz und gröberer Quantisierung nimmt die Genauigkeit ab, wobei die Ortstoleranz der stärkere Einfluß ist. Das ist auch kein Effekt der Skalierung, da die dargestellten Höhenlinien fast ausschließlich die horizontalen Begrenzungen des Diagramms schneiden und dieses im übrigen den plausiblen Parameterbereich vollständig zeigt. Bei höheren Schwellwerten, d.h. höheren Anforderungen an die Vollständigkeit der erkannten Kantenpunkte, wandert das Maximum in Richtung höherer Ortstoleranzen. Gleichzeitig sinkt die Anzahl der notwendigen Quantisierungsstufen, die Genauigkeit fällt erst unter 10–15 Quantisierungsstufen stark ab. Der Grund dafür ist wahrscheinlich, daß bei einer niedrigen Schwelle nur wenige Merkmale erkannt werden müssen, um das gesamte Objekt

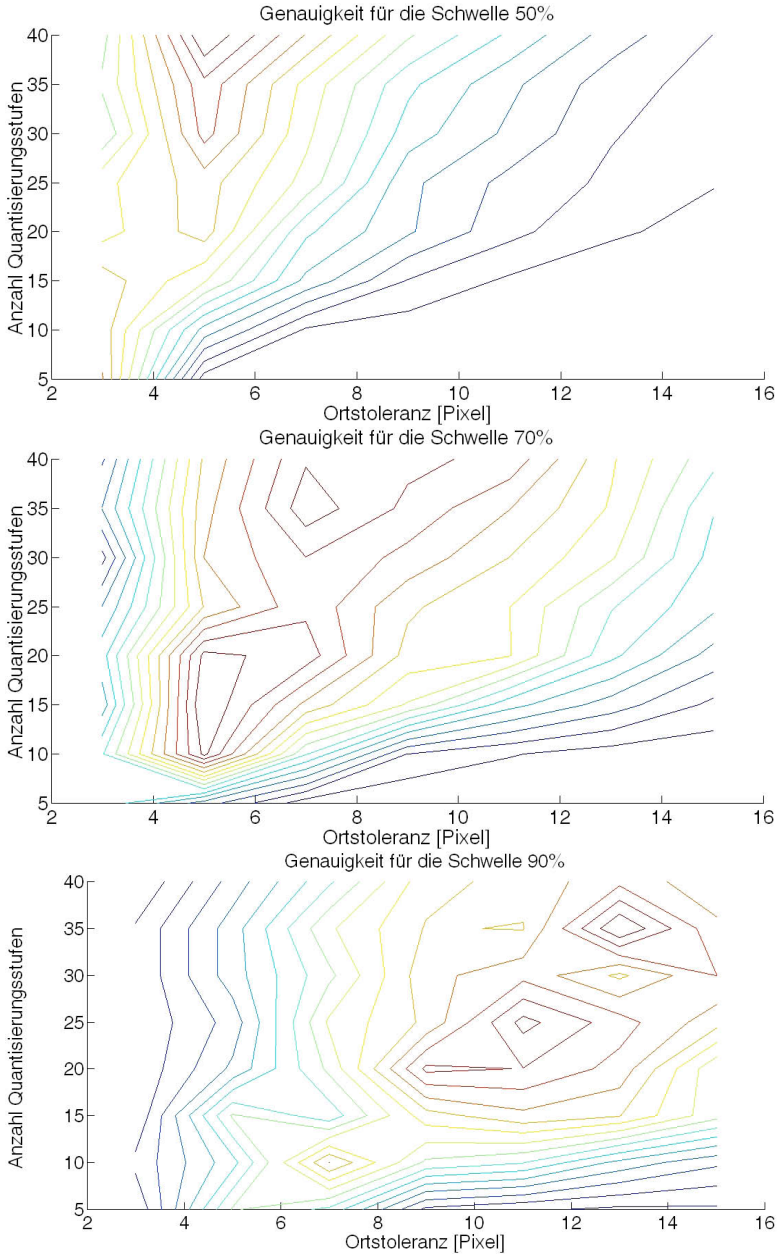


Abbildung 6.13: **Genauigkeit der Augenbrauenerkennung abhängig von der Ortstoleranz und der Schwelle.** Die Diagramme zeigen die Genauigkeit der Objekterkennung in Form von Höhenlinien abhängig von der Ortstoleranz. Die Schwelle zur Erkennung liegt im ersten Diagramm bei 50%, im zweiten bei 70% und im dritten bei 90%.

zu erkennen. Wenn die Erkennung dennoch zuverlässig sein soll, müssen diese wenigen Merkmale genau übereinstimmen. Dies wird zum einen durch eine gute geometrische Übereinstimmung in Form einer niedrigen Ortstoleranz und zum anderen über eine feine Unterscheidung verschiedener Merkmalsausprägungen erreicht. Da die Diagramme ein kontinuierlich wanderndes langgestrecktes Maximum zeigen, liegen die insgesamt besten Genauigkeiten offenbar auf einem ebenen Unterraum der drei überprüften Parameter. Die insgesamt besten Ergebnisse zeigen sich für $\varsigma = 5$ Pixel, $\vartheta = 40\%$ und 25 Quantisierungsstufen, sowie für $\varsigma = 10$ Pixel, $\vartheta = 50\%$ und 40 Quantisierungsstufen. Die starke Abhängigkeit von dem Geometrieparameter ς spricht gegen ein Bag-of-Features-Modell auf Merkmalsebene. Die Merkmalsausprägung ist allerdings ebenfalls bedeutsam.

Im Vergleich mit der Simulation zeigen sich insgesamt starke Abweichungen. Zwar zeigt insbesondere die Abbildung 6.13 unten auch einen Abfall der Genauigkeit bei feineren Quantisierungsstufen wie von der Simulation vorhergesagt, die Position des Maximum hängt jedoch stark von den Knotenparametern ab. Eine Verschiebung in Richtung gröberer Quantisierungsstufen ist aufgrund des höheren Rauschanteils möglich, der aus der Stichprobe stammt. Bei Verschiebungen in Richtung feinerer Quantisierungen können die Geometrieparameter eine Rolle spielen. Wenn in der Stichprobe benachbarte Merkmale ähnliche Ausprägungen besitzen, können nicht erkannte Merkmale möglicherweise durch Nachbarmerkmale kompensiert werden und so die Wahrscheinlichkeit der korrekten Klassifikation erhöhen. Allgemein bedeutet dies, daß sich durch die Messung mehrerer zu einem Modell zusammengefaßter Merkmale theoretisch ein höherer Informationsgehalt ergibt. Dieser kann wiederum durch Abhängigkeiten zwischen den gemessenen Merkmalen verringert werden. Qualitätiv sprechen die in Abbildung 6.13 gezeigten Ergebnissen mit dieser Interpretation überein, da sie für feine Merkmalsquantisierungen die höchsten Genauigkeiten liefern, wenn die Schwelle niedrig oder die Ortstoleranz hoch ist.

Neben der pixelbezogenen Genauigkeit stellt auch die Anzahl der erkannten Objekte ein Kriterium dar. Dazu wird für jede Quantisierung die beste Parametereinstellung bezüglich Schwellwert und Ortstoleranz ermittelt. Für diese Parametrisierung wird die Anzahl der erkannten Augenbrauen ermittelt. Die Ergebnisse zeigt Abbildung 6.14. Offenbar spielt die Feinheit der Quantisierung keine Rolle was die Anzahl der Augenbrauen angeht. Bezüglich der Genauigkeit zeigt sich ein Maximum bei 25 Quantisierungsstufen.

Da die Objekterkennung nicht nur auf eine einzige Abbildung eines Donald-Kopfes trainiert werden soll, sondern auf viele verschiedene Objektansichten, wird nun untersucht, wie sich verschiedene Merkmalsquantisierungen auf die Objekterkennung in unterschiedlichen Bildern auswirkt. Als zu erkennendes Objekt dient wieder eine Augenbraue (siehe Abbildung 6.15 oben). Das Modell wird genau wie in dem (wenige Absätze) zuvor beschriebenen Versuch erzeugt. Als Testbilder dienen 24 Abbildungen von Donald-Köpfen, in denen eine Augenbraue ähnlich der des Referenzbildes vorhanden ist. Abbildung 6.15 zeigt einen Teil der Testbilder. Die Bereiche für eine korrekte Erkennung werden wieder manuell markiert. Wie im vorhergehenden Versuch wird für jede Parametrisierung aus dem Bereich $\varsigma = 3 \ldots 15$ Pixel, $\vartheta = 20\% \ldots 90\%$ sowie Quantisierungen zwi-

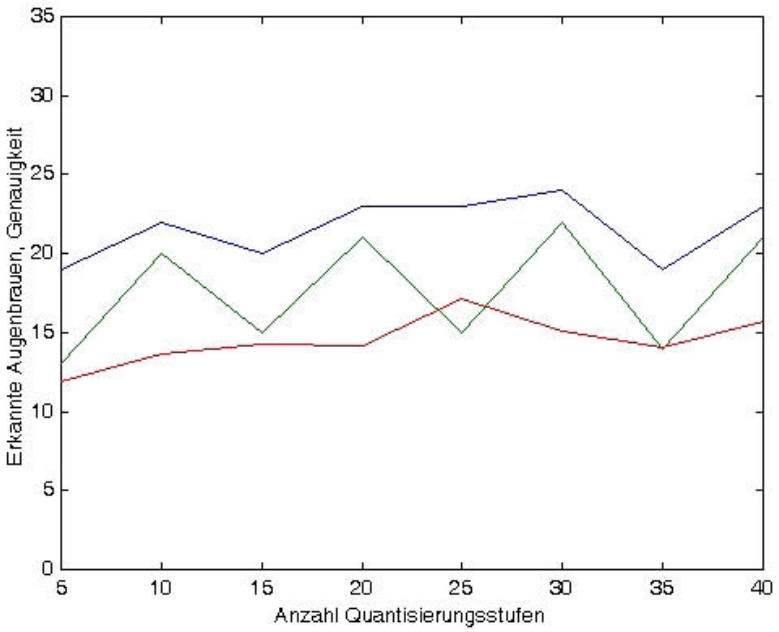


Abbildung 6.14: **Anzahl der erkannten Augenbrauenerkennung abhängig von der Anzahl an Quantisierungsstufen.** Für jede Quantisierungsfeinheit wird die beste Parametrisierung bezüglich Ortstoleranz und Schwelle ausgewählt. Die rote Linie gibt die Genauigkeit der Erkennung der besten Parametrisierungen an. Die blaue Linie zeigt die Anzahl der in den 35 Testbildern erkannten linken Augenbrauen an. Die grüne Linie zeigt Treffer auf der rechten Augenbraue an.

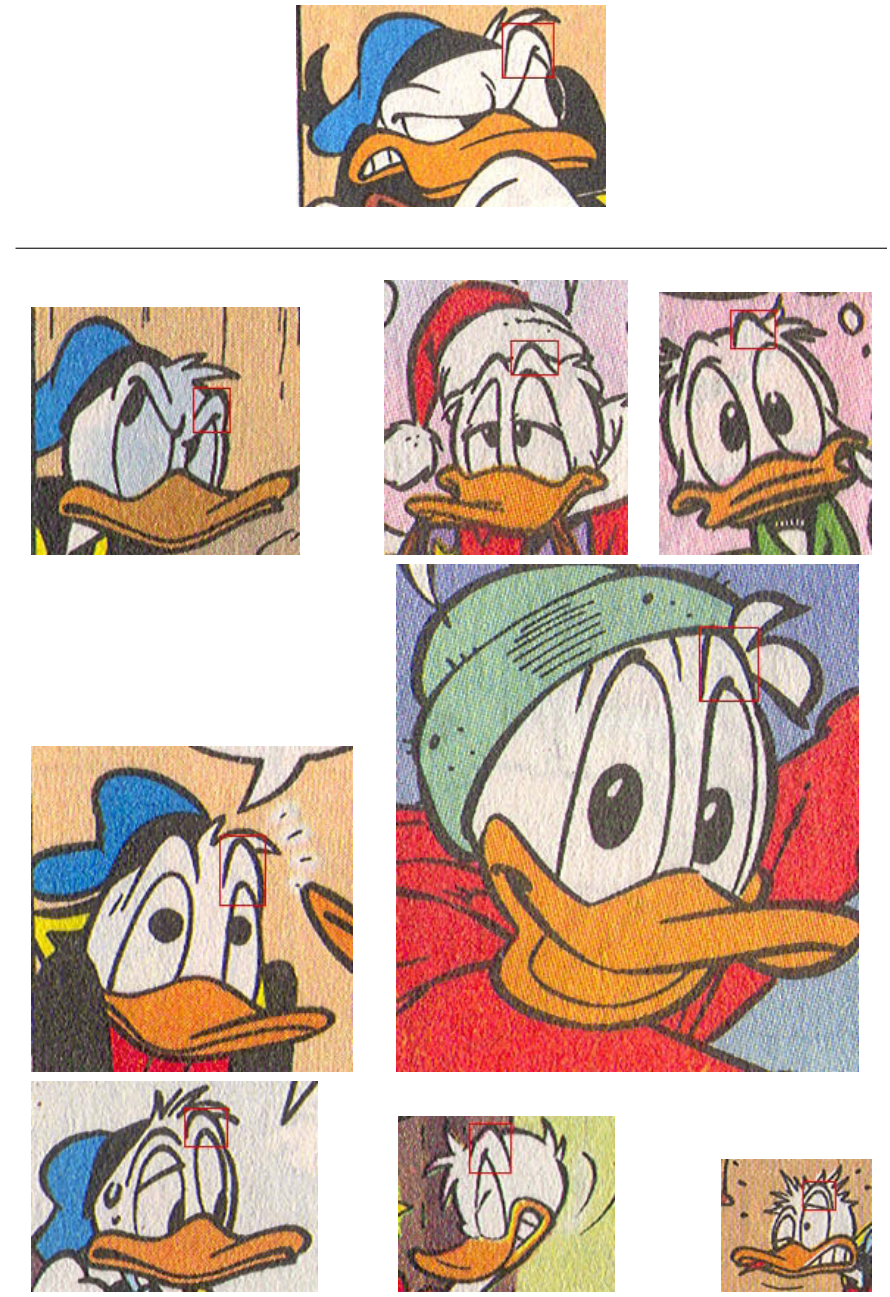


Abbildung 6.15: **Erkennung einer Augenbraue in verschiedenen Donald-Bildern.** Das einzelne Bild oben ist das Referenzbild. Zu den Kantenmerkmalen der markierten Augenbraue wird ein Modell erzeugt. Das Modell wird nun in 24 weiteren Donald-Bildern gesucht. Hier sind acht dieser Bilder dargestellt. Der Bereich für richtige Treffer wurde jeweils manuell markiert.

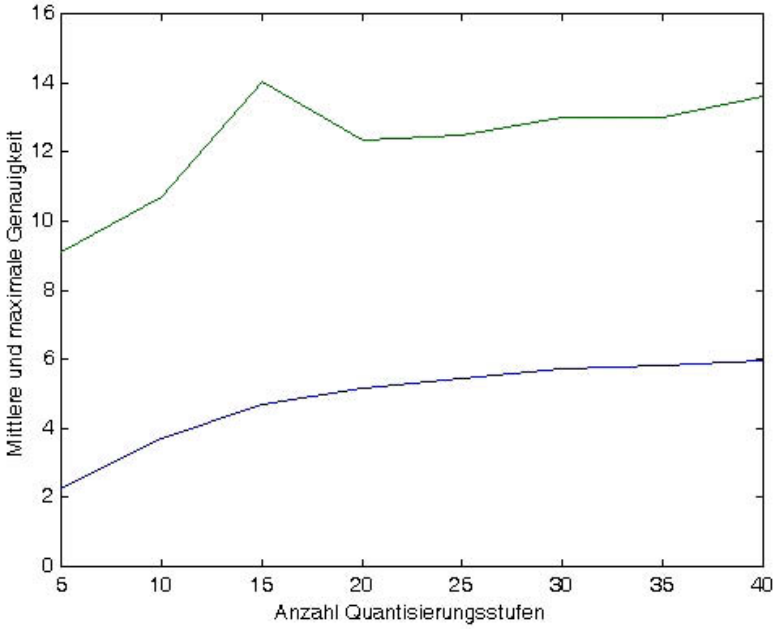


Abbildung 6.16: **Genauigkeit der Erkennung von 24 Augenbrauen.** Die obere grüne Linie gibt die maximale Genauigkeit an, die für eine bestimmte Anzahl an Quantisierungsstufen über alle getesteten Parameter ς und ϑ erzielt wurde. Die untere blaue Linie gibt dagegen die mittlere Genauigkeit über alle Parameter ς und ϑ an.

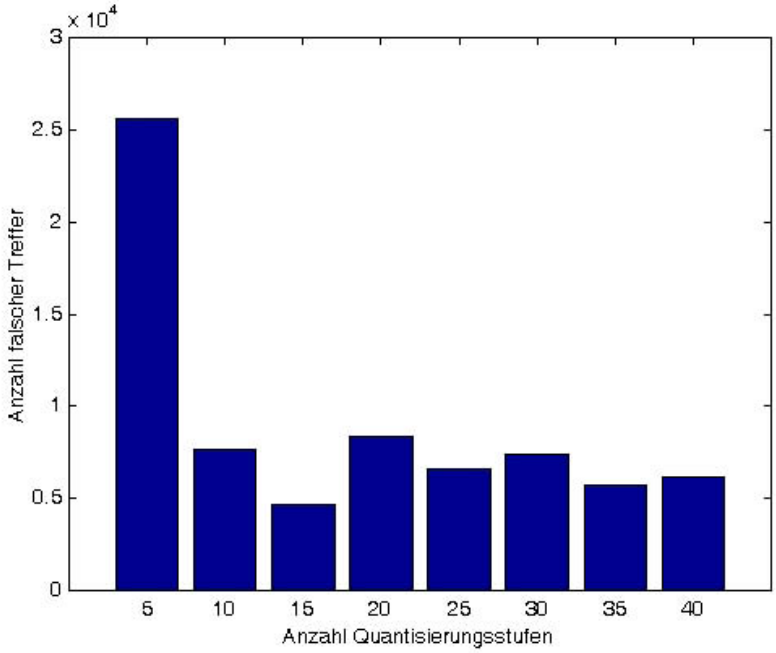


Abbildung 6.17: **Mittlere Anzahl falscher Treffer über die 5 besten Parametrisierungen für jede Anzahl an Quantisierungsstufen.** Für jede Quantisierungsfeinheit wurden 5 Parametrisierungen bezüglich des Schwellwerts und der Ortstoleranz gesucht, bei denen alle 24 Augenbrauen erkannt wurden, und die jeweils die wenigsten falschen Treffer lieferten. Das Diagramm zeigt nun die mittlere Zahl falscher Treffer jeweils über die 5 besten Parametrisierungen für jede Quantisierungsfeinheit.

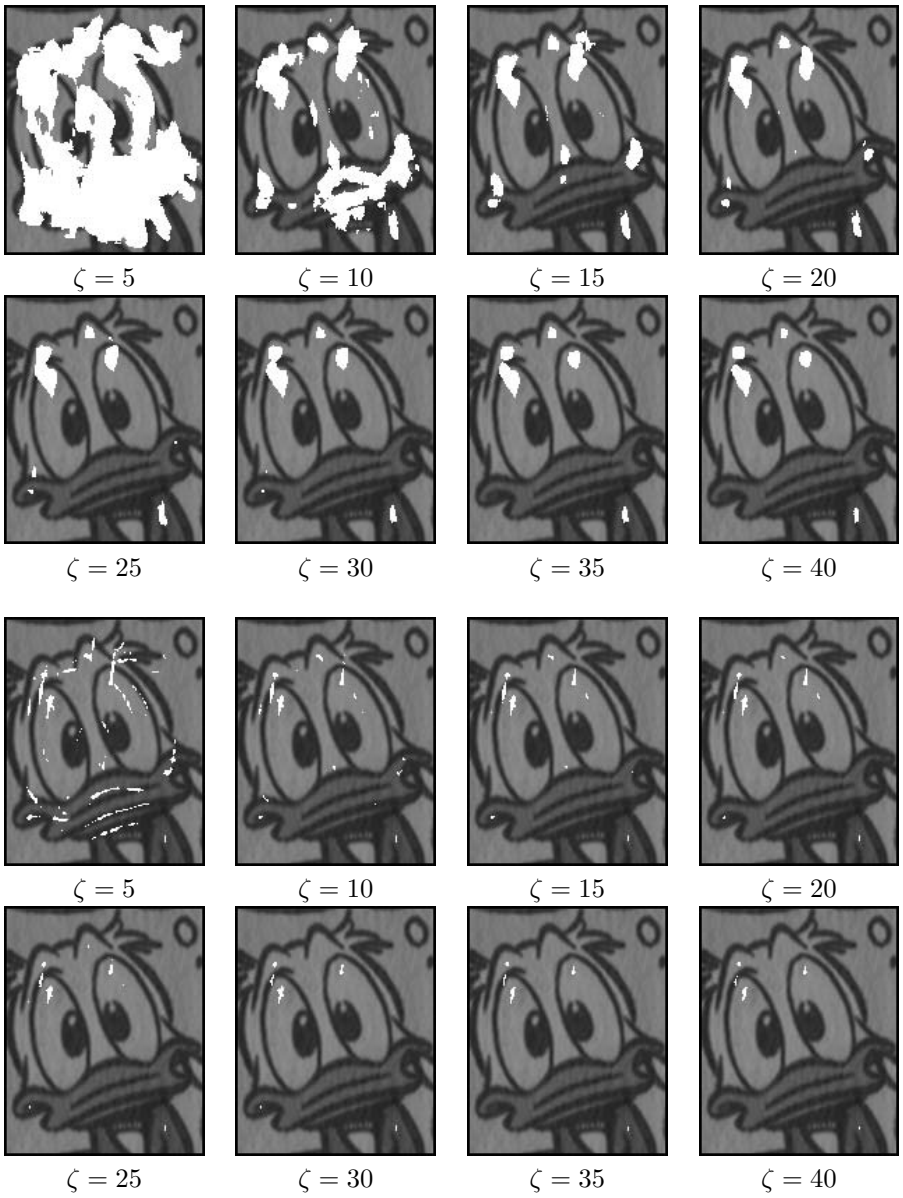


Abbildung 6.18: **Trefferbilder der Augenbrauenerkennung für verschiedene Quantisierungen.** Die Variable ζ gibt jeweils die Anzahl der Quantisierungstufen an. Für die oberen Bilder wurde die Ortstoleranz auf 9 Pixel und die Schwelle auf 60% eingestellt. Für die unteren Bilder wurde die Ortstoleranz auf 3 Pixel und die Schwelle auf 30% eingestellt. Die Anzahl der Quantisierungstufen ist unter jedem Bild angegeben. Ab einer gewissen Mindestzahl an Quantisierungstufen stellen sich gleichbleibende Ergebnisse ein.

schen 5 und 40 Stufen die Genauigkeit berechnet. Die Ergebnisse sind in Tabelle B.4 des Anhangs angegeben.

Um die Ergebnisse bezüglich der Quantisierungstufen zusammenzufassen, wird für jede Quantisierungsfeinheit einmal die mittlere und einmal die maximale Genauigkeit bezüglich des Parameterbereichs der Ortstoleranz ς und der Schwelle ϑ bestimmt. Diese sind in Abbildung 6.16 dargestellt. Wie im vorigen Versuch (vgl. Abb. 6.12) steigt die Genauigkeit im Durchschnitt mit einer feineren Quantisierung der Kantenmerkmale, wobei die Steigung bei einer feineren Quantisierung ebenfalls nachläßt. Bei 20 Quantisierungsstufen sind schon 80 Prozent des Maximalwerts bei 40 Stufen erreicht, bei 25 Quantisierungsstufen bereits 90 Prozent.

Der große Unterschied zwischen der mittleren und der maximalen Genauigkeit deuten allerdings an, daß die Ortstoleranz und der Schwellwert wieder einen großen Einfluß auf das Ergebnis haben. Wie die Abbildung zeigt, ergibt sich bereits für 15 Quantisierungsstufen die maximale Genauigkeit über alle Modell- und Verfahrensparametrisierungen. Dabei spielt sicher auch eine Rolle, daß die Testbilder nun eine größere Variation zeigen. Diese spiegelt sich offenbar auch in den Merkmalsausprägungen wieder.

In der Praxis ist bei der Bewertung jedoch nicht unbedingt von Bedeutung, wieviele Pixel auf dem Objekt korrekt erkannt wurden, sondern vielmehr eine einfache Entscheidung, ob ein bestimmtes Objekt vorliegt oder nicht. Um die Versuche noch realistischer zu gestalten wird daher für jede Parametrisierung die Anzahl der erkannten Augenbrauen ermittelt. Die Ergebnisse sind in Tabelle B.5 des Anhangs dargestellt. Bei den falschen Treffern zählt jedoch weiterhin jedes Pixel. Die entsprechenden Ergebnisse zeigt Tabelle B.6.

Zusammengefaßt läßt sich sagen, daß sich für den größten Teil des getesteten dreidimensionalen Parameterraums alle 24 gesuchten Augenbrauen erkennen lassen. Nur für sehr restriktive Einstellungen der Ortstoleranz und des Schwellwerts ergeben sich weniger Treffer. Die Zahl der falschen Treffer steigt dagegen sowohl mit einer steigenden Glättung als auch mit einem sinkenden Schwellwert sehr schnell an. Die Anzahl der Quantisierungstufen scheint über den gesamten Parameterraum gesehen einen deutlich geringeren Einfluß zu haben. Aufgrund der hohen Anzahl falscher Treffer bei den meisten Parametereinstellungen ist nur ein kleiner Teil der möglichen Parametrisierungen interessant. Um den Einfluß der Quantisierungen bei diesen interessanten Einstellungen zu erfassen, werden nur für jede Quantisierungsfeinheit jeweils die besten 5 Ergebnisse ausgewählt. Die Gütebedingung zur Auswahl der besten Parametrisierungen ist, daß bei möglichst wenigen falschen Treffern alle 24 Augenbrauen erkannt werden sollen. Das alternative Gütekriterium aus den vorangegangenen Versuchen ist hier weniger geeignet, da es falsche Treffer sehr stark bewertet. Dadurch ergibt sich die höchste Genauigkeit oft für Parametrisierungen, bei denen fast keine Treffer erzielt werden. Der Mangel an positiven Ergebnissen ist jedoch für die Klärung des Einflusses der Quantisierung auf die Erkennung nachteilig. Möglicherweise ergeben sich noch bessere Konfigurationen bei einem niedrigeren Recall, d.h. weniger als 24 erkannten Testbildern. Die Anzahl der erkannten Testbilder fällt

jedoch über dem Parameterraum schnell ab, was eine einheitliche Festlegung einer kleineren Bilderanzahl erschwert.

Für jede Quantisierungsfeinheit wird nun die Zahl der falschen Treffer über die jeweils besten 5 Parametrisierungen gemittelt. Das Ergebnis zeigt Abbildung 6.17. Wie zu sehen ist, liefert eine grobe Quantisierung mit fünf Stufen die schlechtesten Ergebnisse. Für feinere Quantisierungen ergeben sich deutlich bessere Ergebnisse, die allerdings auf einem konstanten Niveau verharren. Zusammengefaßt bedeutet dies, daß für relevante Einstellungen der Schwelle und der Ortstoleranz eine Mindestanzahl von 10 Quantisierungsstufen erforderlich ist. Feinere Quantisierungen führen jedoch weder zu mehr richtigen Treffern, da hier aufgrund des Entwurfs des Experiments das Maximum erreicht war, noch zu weniger falschen Treffern. Möglicherweise drängt auch die pragmatische Forderung nach einem hohen Recall die Objekterkennung in Richtung einer groben Quantisierung. Übergroße Quantisierungsintervalle können hier die Chance erhöhen, ein Objekt zu detektieren.

Einen weiteren Einblick in die Bedeutung der Quantisierung liefert die Abbildung 6.18. Diese stellt die Trefferbilder über der Anzahl an Quantisierungsstufen für zwei ausgewählte Einstellungen bezüglich Ortstoleranz und Schwellwert dar. Für feinere Quantisierungen als 10 Stufen ergeben sich nur noch minimale Änderungen.

Nachdem die Frage nach der Quantisierung von Kantenmerkmalen geklärt wurde, stellt sich die Frage, ob für die anderen Merkmale die gleichen Gesetzmäßigkeiten gelten. Möglicherweise verhalten sich Skelettmerkmale grundsätzlich anders als Kanten. Um diese Frage zu beantworten, werden die in den Tabellen B.4, B.5 und B.6 dargestellten Versuchsergebnisse zur Erkennung von Augenbrauen anhand von Kantenmerkmalen für die Nutzung von Skelettmerkmalen neu berechnet. Als Referenzobjekt dient jetzt der Schnabel eines Donald-Bildes, da Schnäbel relativ flächig sind und daher viele Skelettmerkmale enthalten. Die Skelettmerkmale für den Schnabel des in Abbildung 6.19 oben gezeigten Referenzbildes werden wie im vorigen Versuch zu einem Modell zusammengefaßt, indem ein übergeordneter Teileknoten erstellt wird. Dieser wird in einer Stichprobe von 19 Donald-Bildern mit ähnlich großen und ähnlich ausgerichteten Schnäbeln gesucht. Die richtige Position wird wieder jeweils manuell markiert. Beispiele aus der Stichprobe zeigt Abbildung 6.19 unten. Der Parameterraum umfaßt in diesem Versuch wieder die Schwelle und die Ortstoleranz des Teileknotens. Die Quantisierung wird jetzt allerdings bezüglich der Orientierung der Skelettmerkmale untersucht. Aufgrund der Erfahrungen aus dem vorigen Versuch wird ein geringerer Wertebereich von 4 bis 24 Quantisierungsstufen untersucht. Die Anzahl der erkannten Schnäbel pro Parametrisierung ist in Tabelle B.7 angegeben. Die Zahl der falschen Treffer zeigt Tabelle B.8.

Zur Auswertung der Ergebnisse werden für jede Quantisierungsfeinheit wieder die besten fünf Ergebnisse aus allen Parametrisierungen bezüglich der Schwelle und der Ortstoleranz bestimmt. Als beste Ergebnisse sind wie im vorhergehenden Versuch die Parametrisierungen zu verstehen, bei denen alle 19 Schnäbel mit einer minimalen Anzahl falscher Treffer erkannt werden. Die mittlere Zahl falscher Treffer über die jeweils besten fünf Parametrisierungen

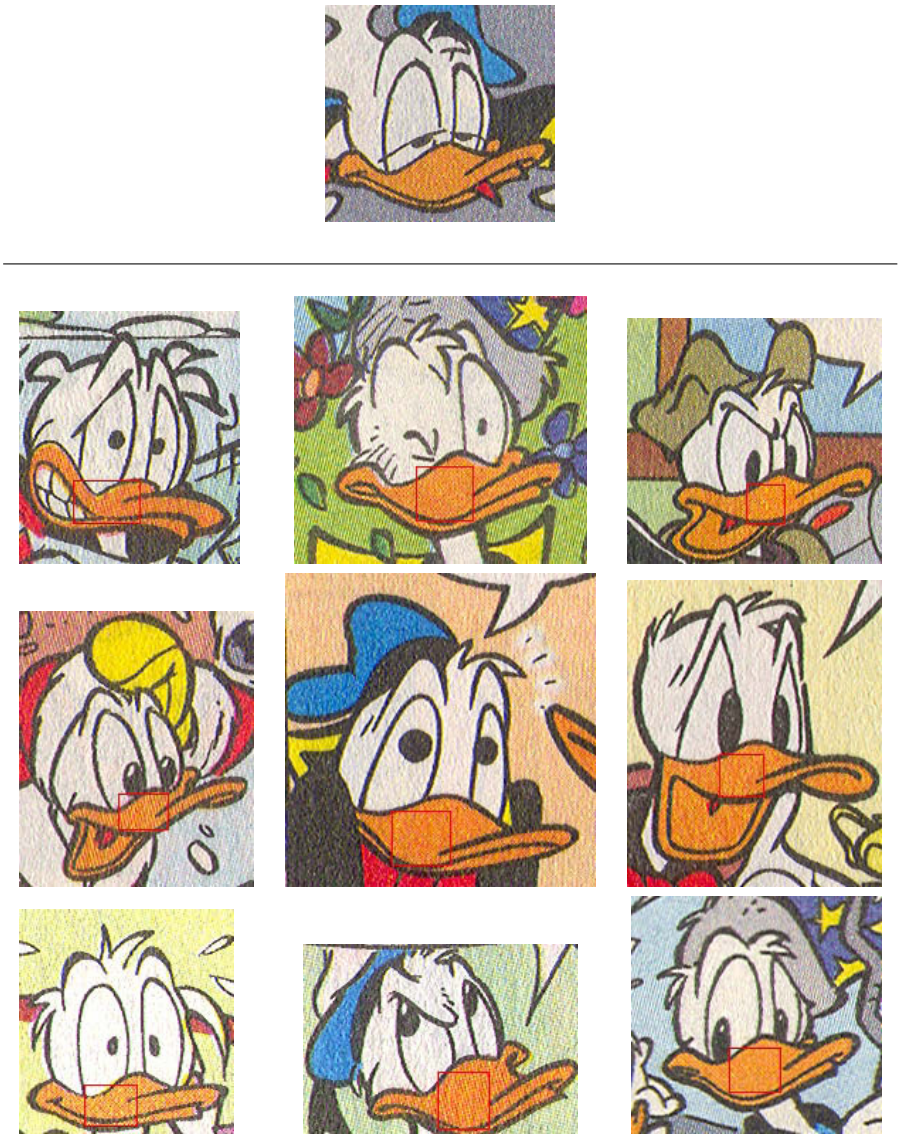


Abbildung 6.19: **Erkennung des Schnabels in verschiedenen Donald-Bildern.** Das einzelne Bild oben ist das Referenzbild. Aus den Skelettmerkmalen des Schnabels wird ein Modell erzeugt, das in 19 Testbildern gesucht wird. Neun dieser Testbilder sind hier dargestellt. Die Testbilder wurden so ausgewählt, daß die Figur wie im Referenzbild nach rechts schaut und die Größe ähnlich ist. Der Bereich für richtige Treffer wurde wieder manuell markiert.

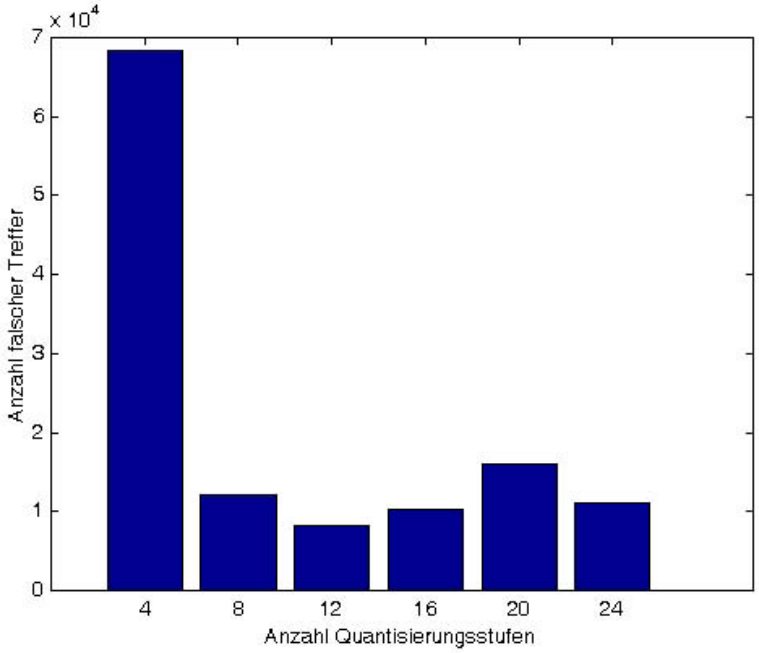


Abbildung 6.20: **Mittlere Anzahl falscher Treffer über die 5 besten Parametrisierungen für jede Anzahl an Quantisierungsstufen.** Für jede Quantisierungsfeinheit wurden 5 Parametrisierungen bezüglich des Schwellwerts und der Ortstoleranz gesucht, bei denen alle 19 Schnäbel erkannt wurden, und die jeweils die wenigsten falschen Treffer lieferten. Das Diagramm zeigt nun die mittlere Zahl falscher Treffer jeweils über die 5 besten Parametrisierungen für jede Quantisierungsfeinheit.

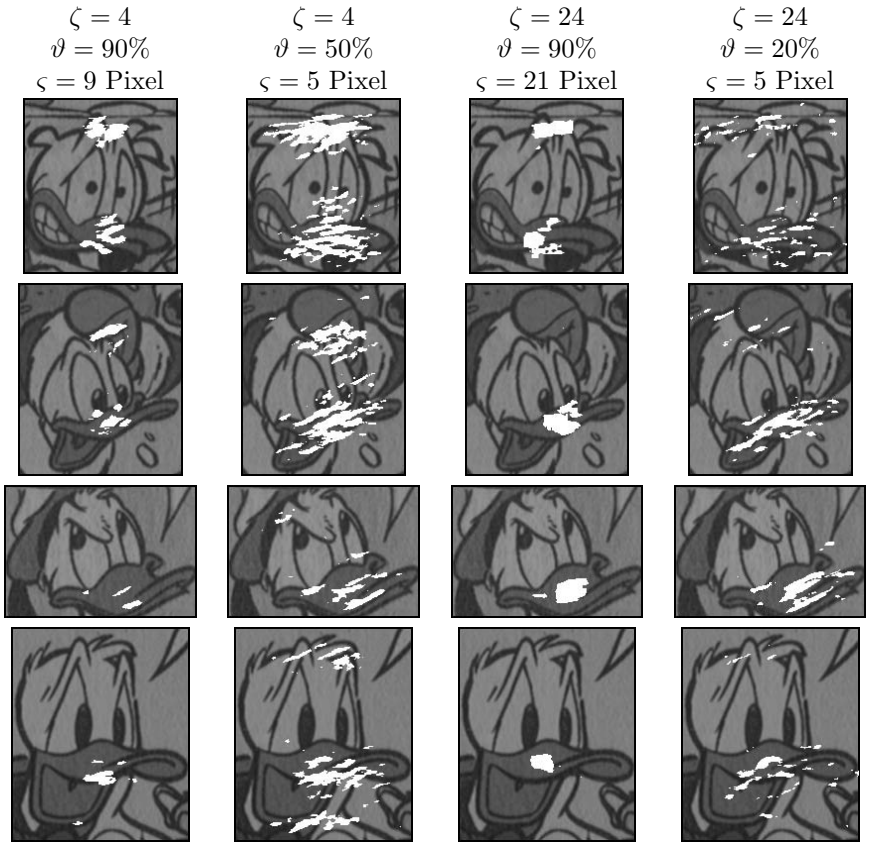


Abbildung 6.21: **Trefferbilder der Schnabelerkennung für verschiedene Quantisierungen.** Die Trefferbilder in der ersten und zweiten Spalte entstanden durch die Quantisierung mit $\zeta = 4$ Stufen. Für die dritte und vierte Spalte wurden 24 Quantisierungsstufen unterschieden. Die über den Spalten angegebenen Parametrisierungen bezüglich des Schwellwerts und der Ortstoleranz gehören zu den erwähnten 5 besten Ergebnissen für die betreffende Anzahl an Quantisierungsstufen. Ein Vergleich zwischen den Spalten links und rechts zeigt leicht zielgenauere Ergebnisse für die feinere Quantisierung. Im direkten Vergleich scheint jedoch die höhere Schwelle der Spalten 1 und 3 einen größeren Einfluß zu haben, was sich an den kompakten Treffern auf den Schnäbeln zeigt.

zeigt Abbildung 6.20. Offenbar genügen 8 Quantisierungsstufen für eine optimale Erkennung. Die Ähnlichkeit zu den Ergebnissen für die Kantenrichtung (Abbildung 6.17) ist klar erkennbar. Skelettmerkmale verhalten sich bezüglich der Quantisierung offenbar genauso wie Kantenmerkmale.

Die Abbildung 6.21 zeigt schließlich noch einige Trefferbilder für vier der besten ermittelten Parametrisierungen. Bei der Erzeugung der Trefferbilder der linken beiden Spalten wurden 4 Quantisierungsstufen für die Orientierung von Skelettlinien unterschieden, bei den rechten Spalten dagegen 24 Stufen. Obwohl die Trefferbilder rechts präziser erscheinen, ist der Einfluß der Quantisierung gering im Vergleich zwischen den unterschiedlichen Parametrisierungen bezüglich Ortstoleranz und Schwellwert. Diese Beobachtung deckt sich ebenfalls mit den Ergebnissen aus dem vorigen Versuch. Bei der Bewertung der Trefferbilder ist im übrigen zu beachten, daß die Parametrisierungen hinsichtlich eines guten Vergleichs verschiedener Quantisierungsfeinheiten ausgewählt wurden. Für andere Parametrisierungen ergeben sich nochmals kompaktere und ungestörtere Trefferbilder.

6.1.5 Zwischenfazit

Auf der Teileebene des hierarchischen Objektmodells geschieht das Training durch eine Merkmalsextraktion. Die Knoten des Modells werden durch drei verschiedene Merkmalsdetektoren erzeugt. Der erste Detektor erkennt Kantenpunkte mit Subpixelgenauigkeit. Die durch diesen Detektor erzeugten Knoten repräsentieren lokale Kantenstücke einer bestimmten Orientierung. Aufbauend auf den Ergebnissen des ersten Detektors ermittelt der zweite Detektor Eckenmerkmale. Modellknoten speichern jedoch außer dem Auftreten der Ecken keine weiteren Informationen. Der dritte Detektor erkennt Flächen und beschreibt diese durch Punkte auf Skelettlinien. Der vorgestellte Detektor basiert auf einem Verfahren zur Abstandstransformation, das hinsichtlich Verarbeitungsgeschwindigkeit und numerischer Stabilität optimiert ist. Die Ortsgenauigkeit liegt hier jedoch nur bei einem Pixel. Die zugehörigen Modellknoten repräsentieren Kreisflächen eines bestimmten Durchmessers und einer bestimmten Helligkeit. Darüberhinaus geben sie die Ausrichtung der umgebenden Fläche an, da diese in der Regel selbst kein Kreis ist.

Da die Merkmalsausprägungen während der Objekterkennung grob diskretisiert werden, muß zur Festlegung passender Quantisierungsstufen das Rauschen auf den Merkmalen für die Cartoon-Stichprobe bestimmt werden. Die Messungen ergeben, daß sich jeweils etwa 20 Kantenrichtungen, Intensitätswerte von Skelettpunkten und Orientierungen von Skelettpunkten, sowie pixelgenaue Kantenabstände unterscheiden lassen. Es zeigt sich jedoch, daß bei der Erkennung unterschiedlicher Bilder mit dem gleichen Modell zusätzliche Merkmalsschwankungen auftreten, die aus Variationen des modellierten Objekts über verschiedene Bilder hinweg resultieren. Für Kantenmerkmale wurde ermittelt, daß die Unterscheidung von mehr als 10 Orientierungen keine Vorteile mehr bringt. Dies gilt genauso für Flächenmerkmale, wo die Unterscheidung von 8 Orientierungen bereits optimale Ergebnisse liefert. Die Ergebnisse zeigen jedoch auch, daß die

Feinheit der Quantisierung weniger relevant ist als die Einstellung der Ortstoleranz und der Schwelle in den Modellknoten.

Die hohe Abhängigkeit von der Ortstoleranz spricht eindeutig gegen ein Bag-of-features-Modell auf der Merkmalsebene. Vielmehr scheint der überwiegende Anteil der zur Objekterkennung benötigten Information aus der Geometrie zu stammen. Aufgrund der schlechten Ergebnisse für extrem grobe Quantisierungen ist jedoch auch ein allein auf geometrischen Informationen basierendes Modell nachteilig. Die Abwägung zwischen diesen beiden Extremen wird beim Training auf Teileebene genauer untersucht.

6.2 Identifikation von kritischen Variablen für die Modellierung von Teilen

Die zweite Hierarchieebene faßt Gruppen von Merkmalen zu Teilen zusammen. Die Frage ist nun, nach welchen Kriterien diese Zusammenstellung erfolgen soll, damit möglichst gute Teile erzeugt werden. Wie der folgende Abschnitt beschreibt, verhindern lokale Maxima in der Klassifikationsgüte, das Modell durch die schrittweise Hinzunahme von Merkmalen zu optimieren. Es wird daher nach anderen Variablen gesucht, die für die Teile-Modellierung relevant sind. Zunächst wird ein Zusammenhang zwischen Ortstoleranz und Schwellwert festgestellt. Es stellt sich dann heraus, daß für die Objekterkennung die Modellierung der Geometrie von großer Bedeutung ist und hier offenbar die räumliche Nähe von Merkmalen eine besondere Rolle spielt. Um diese auswerten zu können, wird ein Clusterungsverfahren eingeführt, das räumlich benachbarte Merkmale zu Teilekandidaten zusammenfaßt. Diese werden provisorisch zu Ansichtsmodellen kombiniert, um den Parameterraum zu sondieren. Auf diese Weise werden Abhängigkeiten zwischen der Teilegröße, der Größe der Stichprobenelemente, der Ortstoleranz und der Repräsentativität der trainierten Elemente für die Stichprobe ermittelt. Die neuen Erkenntnisse regen die Erzeugung eines visuellen Alphabets von Teilen an.

6.2.1 Gierige Optimierung von Teilen

Ein einfacher Ansatz ist das inkrementelle Zusammenstellen von Merkmalen. Die Grundidee ist, zunächst alle extrahierten Merkmale bezüglich ihrer Qualität zur Objekterkennung zu bewerten. Das beste Merkmal wird ausgewählt. Als nächstes werden alle Kombinationen des ausgewählten Merkmals mit einem anderen Merkmal überprüft. Falls sich eine Zweierkombination ergibt, die besser ist als das erste Merkmal alleine, wird überprüft, ob sich die Qualität durch Hinzunahme eines dritten Merkmals weiter verbessert. Dieser Vorgang wird so lange wiederholt, bis das Modell nicht weiter verbessert werden kann. Da das Verfahren in jedem Schritt die optimale Lösung anstrebt, wird diese Vorgehensweise *gierige* Optimierung genannt. Gierige Optimierungen eignen sich bekanntermaßen nicht, wenn der Lösungsraum lokale Nebenmaxima aufweist.

Abbildung 6.22: **Testobjekt zur Modelloptimierung.** Für jede Bildkoordinate können Merkmale erzeugt werden. An Kanten dient die Gradientenrichtung als Merkmal, auf Flächen dagegen die Rot-, Grün- und Blau-Werte.

Da Rückschritte zu schlechteren Modellen nicht vorgesehen sind, können lokale Maxima nicht wieder verlassen werden.

Fergus et al. [FPZ07] geben an, daß die Modellerzeugung für die Objekterkennung solche lokale Maxima enthält. Aus diesem Grund untersucht ihr Expectation-Maximization-Algorithmus nicht nur die beste Modellkonfiguration, sondern alle. Stommel und Kuhnert [SK06] halten in ihrem durch Genetische Algorithmen inspirierten Trainingsverfahren ebenfalls mehrere suboptimale Modellkonfigurationen vor, um einer frühen Konvergenz zu suboptimalen Lösungen zu entgehen.

Die Abbildungen 6.23 und 6.24 geben einen Eindruck von dem Problem. Dargestellt ist die schrittweise Erzeugung eines Modells für das in Abbildung 6.22 gezeigte gelbe Auto. Das linke Bild zeigt die Positionen der ausgewählten Merkmale in Form grüner Punkte an. Verschiedene Kombinationen von Merkmalen werden anhand eines Fitness-Kriteriums [SK05] bewertet, das auf dem Test des Modell an einer Datenbank mit Autos und Hintergrundbildern beruht. Die maximal erreichbare Fitness liegt bei dem Wert 1 für eine optimale Erkennungsleistung. Die linken Bilder zeigen neben den ausgewählten Merkmalen für jede Bildposition die Fitness an, die sich ergibt, wenn an der betreffenden Stelle ein zusätzliches Merkmal ausgewählt wird. Je heller der Punkt dargestellt ist, desto höher ist die Fitness. Die Diagramme auf der rechten Seite geben die Fitness über dem Ort noch einmal als 3D-Diagramm an. Die Auswahl startet mit einem Merkmal. Schrittweise wird immer ein weiteres Merkmal für die Position hinzugenommen, an der die Fitness maximal ist. Falls mehrere gleich gute Positionen existieren, wird eine dieser Positionen zufällig ausgewählt. Die Bilder zeigen, daß die Fitness über dem Raum möglicher Modellkonfigurationen nicht glatt verläuft, sondern viele Ausreißer zeigt. Eine deutliche Präferenz für eine bestimmte Merkmalsposition zeigt sich erst bei der Auswahl des fünften Merkmals. Nach acht Schritten wird ein Plateau erreicht und die Optimierung endet mit einer Fitness von 0,63 für das Modell. Wie Stommel und Kuhnert [SK05]

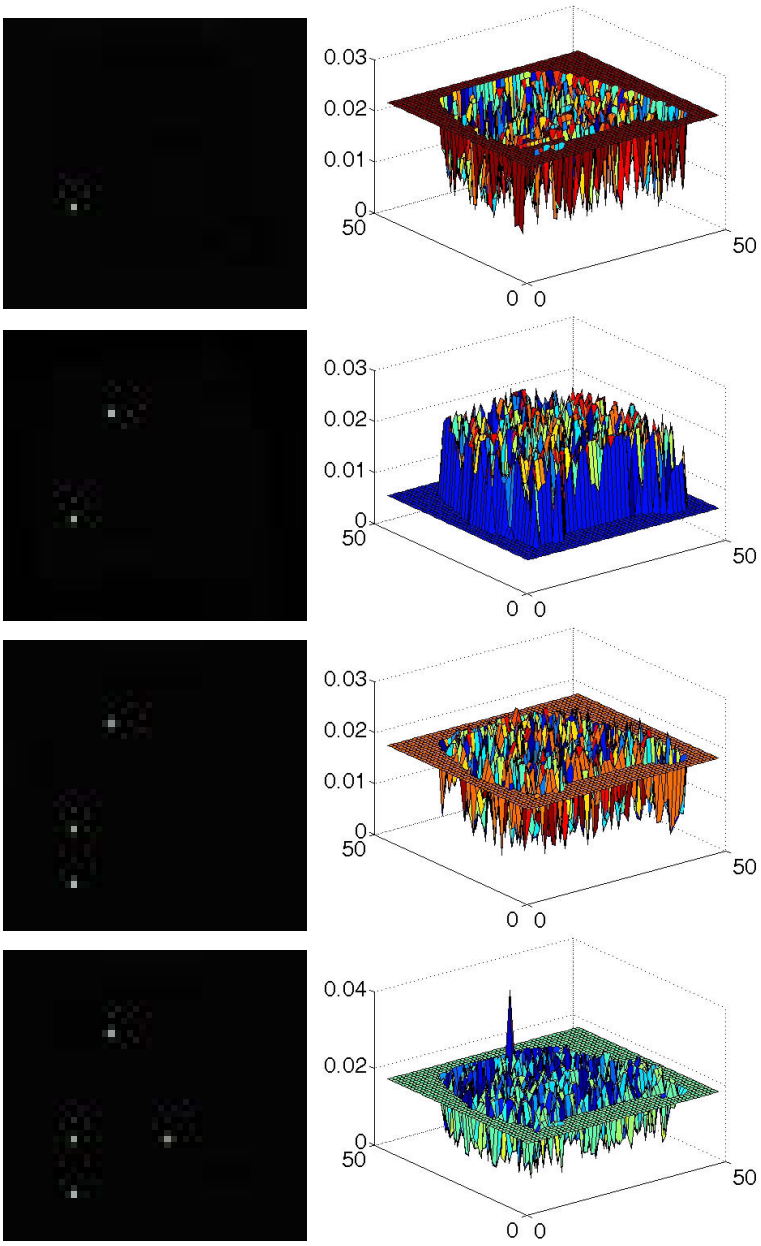


Abbildung 6.23: **Gierige Optimierung einer Merkmalszusammenstellung, Schritte 1–4.** Bei der Auswahl der ersten Merkmale ist kein deutliches Maximum zu erkennen. Die Fitness des Modells ist sehr niedrig.

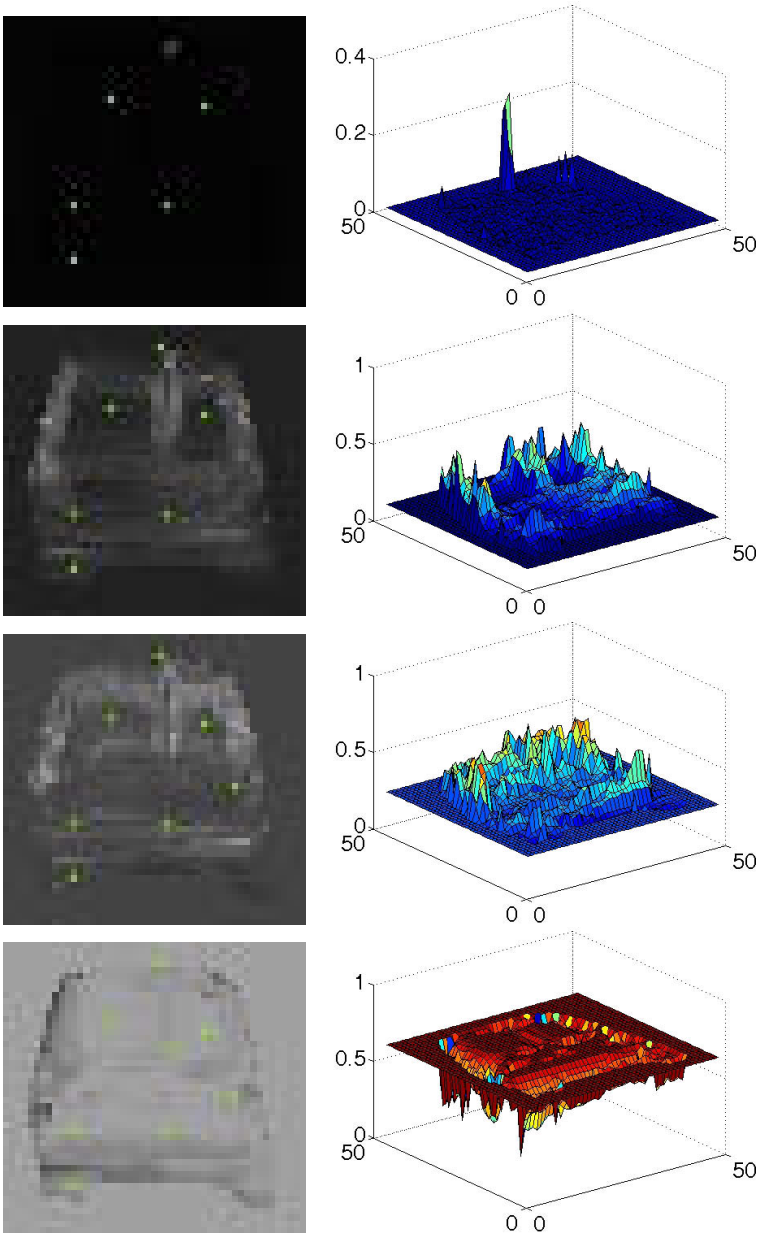


Abbildung 6.24: **Gierige Optimierung einer Merkmalszusammenstellung, Schritte 5–8.** Ab einer Größe von 6 Merkmalen erreicht das Modell eine nennenswerte Fitness. Der Verlauf der Fitness über den Bildkoordinaten zeigt allerdings mehrere Maxima. Diese sind in der Regel suboptimal und beenden die Optimierung früh.

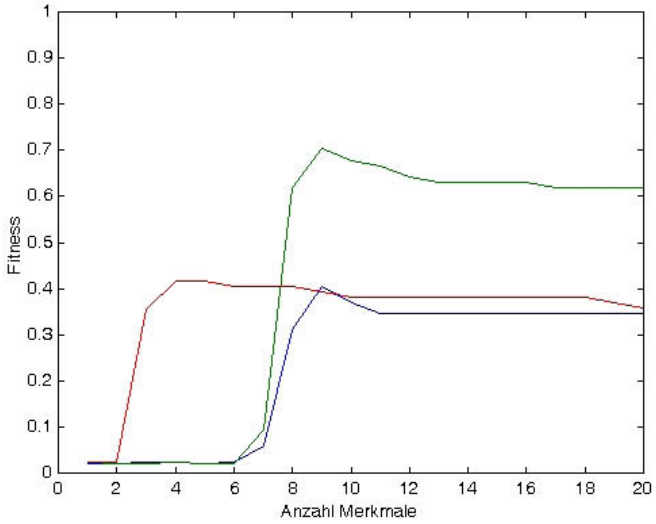


Abbildung 6.25: **Verlauf der Fitness bei gieriger Optimierung.** Der Trainingsverlauf wird für die Erzeugung von drei Modellen gezeigt. Nach jeweils wenigen Optimierungsschritten erreicht die Fitness ein suboptimales Maximum und fällt wieder ab, wenn weitere Merkmale hinzugenommen werden. Eine Optimierung mit einem genetischen Algorithmus erzeugt dagegen Modelle mit einer Fitness von über 0,9.

zeigen, kann jedoch eine Fitness von über 0,9 erreicht werden. Die Optimierung ist also in einem suboptimalen Maximum zu Halt gekommen. Abbildung 6.25 zeigt den Verlauf der Fitness für drei weitere Optimierungen. Die Methode eignet sich anscheinend nicht für die Erzeugung guter Modelle, da sich eine Menge von Merkmalen offenbar anders verhält als Teile der Menge. Dies erinnert an die Versuche von Albright und Gross [AG90], komplexe Stimuli durch die Kombination von Fourierdeskriptoren zu erzeugen, um Zellreaktionen im inferioren Temporallappen hervorzurufen. Hier führte auch erst der entgegengesetzte Ansatz von Tanaka [Tan96] zum Erfolg, bei dem komplexe Stimuli, für die eine Zellreaktion gefunden wurde, schrittweise vereinfacht wurden.

Da die inkrementelle Modellerzeugung offenbar schlecht an das Problem angepaßt ist, bietet sich der bereits erfolgreich eingesetzte genetische Algorithmus oder ein anderes weniger strikt optimierendes Verfahren an. Obwohl diese Optimierungsverfahren zu guten Ergebnissen führen, geben sie jedoch nur wenig Einblick in die Lösungsfindung. Um ein tieferes Verständnis der Materie zu gewinnen, wird in den folgenden Abschnitten daher ein anderer Ansatz verfolgt. Dieser besteht hauptsächlich in der Untersuchung der Abhängigkeiten zwischen der Güte der Objekterkennung auf der einen Seite und den Modell- und Verfahrensparametern auf der anderen Seite. Das Ziel der Untersuchungen ist, Regeln zu finden, die den Aufbau bzw. die Parametrisierung günstiger Modelle anleiten. Diese Erkenntnisse könnten dann auch wieder genutzt werden, um bestehende Trainingsverfahren zielgerichteter einzusetzen und damit zu beschleunigen oder zuverlässiger zu machen. Der letzte Punkt ist allerdings nicht das Thema dieser Arbeit.

6.2.2 Abhängigkeit zwischen Ortstoleranz und Schwellwert

Da für verschiedene Modellausrichtungen immer wieder die Ortstoleranz und der Schwellwert von Knoten geändert werden müssen, wird als erstes untersucht, welche Einflußmöglichkeiten hier bestehen. Dazu wird noch einmal auf die Versuchsergebnisse in Tabelle B.3 zurückgegriffen. Diese zeigt die Genauigkeit der Objekterkennung über den vollständigen plausiblen Parameterraum eines Augenbrauen-Modells für systematische Variationen der Quantisierung, der Ortstoleranz und des Schwellwerts. Durch Summation entlang der Quantisierungsachse werden die Versuchsergebnisse auf die Parameterebene aus Ortstoleranz und Schwellwert projiziert. Die Ergebnisse zeigt Abbildung 6.26. Offenbar sind Ortstoleranz und Schwellwert stark korreliert. Für hohe Schwellwerte und geringe Ortstoleranz sind die Ergebnisse jedoch zunehmend nichtlinear. Der Zusammenhang zwischen Ortstoleranz und Schwellwert läßt sich teilweise durch den Versuchsaufbau erklären, da alle Testbilder durch Skalierung eines einzelnen Originalbildes entstanden sind. Da die Bilder bis auf die Größe gleich sind, sollten im Prinzip alle Merkmale des Modells in allen Testbildern auffindbar sein, allerdings in unterschiedlichen Abständen. Bei geringer Ortstoleranz können jedoch große Skalierungen nicht berücksichtigt werden, da dann einige Merkmale bei großen Objekten außerhalb der Toleranzgrenzen liegen. Damit das Objek-

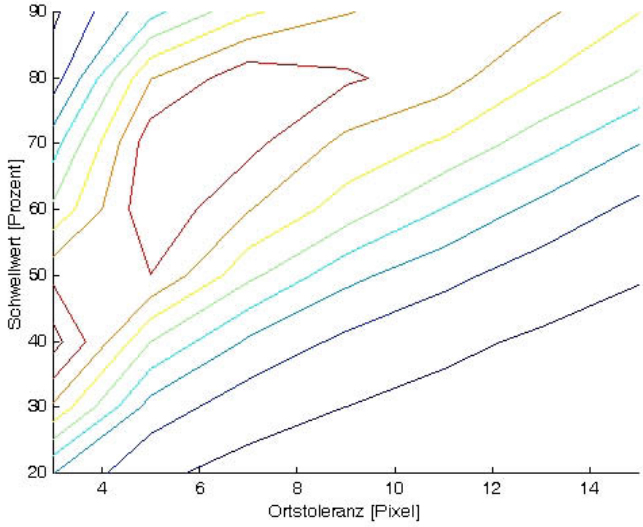


Abbildung 6.26: **Genauigkeit der Augenbrauenerkennung abhängig von der Ortstoleranz und der Schwelle.** Der Plot zeigt die Genauigkeit jeweils aufsummiert über alle verschiedenen Quantisierungen zwischen 5 und 40 Stufen in Form von Höhenlinien. In dieser Darstellung erscheinen Ortstoleranz und Glättung stark korreliert.

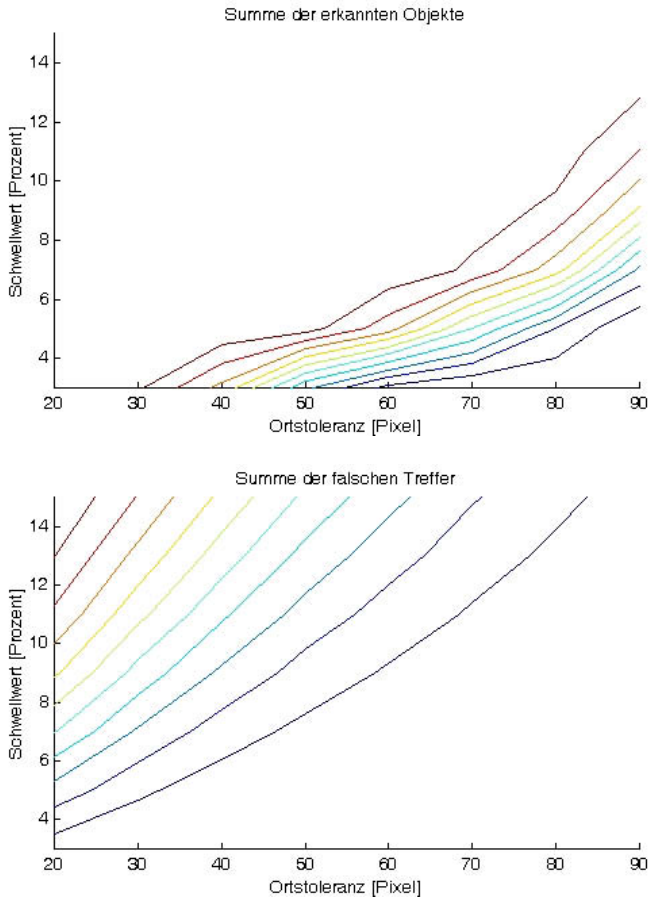


Abbildung 6.27: **Erkennung von 24 Augenbrauen über Ortstoleranz und Schwelle.** Das obere Diagramm zeigt die Anzahl der erkannten Augenbrauen in 24 Testbildern in Form von Höhenlinien. Die höchsten Werte ergeben sich für niedrige Schwellen bei hoher Ortstoleranz. Die niedrigsten Werte ergeben sich für hohe Schwellen und geringe Ortstoleranz. Die zwei getesteten Parameter hängen offenbar voneinander ab. Allerdings ist der Zusammenhang leicht nichtlinear. Das untere Diagramm zeigt die Anzahl der falschen Treffer über der Ortstoleranz und der Schwelle. Für hohe Schwellen bei geringer Ortstoleranz ergeben sich nur wenige falsche Treffer. Die Trefferzahl steigt allerdings für niedrige Schwellen und höhere Ortstoleranzen steil an. Aufgrund des starken Anstiegs spielt die leichte Nichtlinearität bei der Zahl richtiger Treffer eine Rolle, wenn die Genauigkeit als Quotient aus richtigen und falschen Treffern berechnet wird.

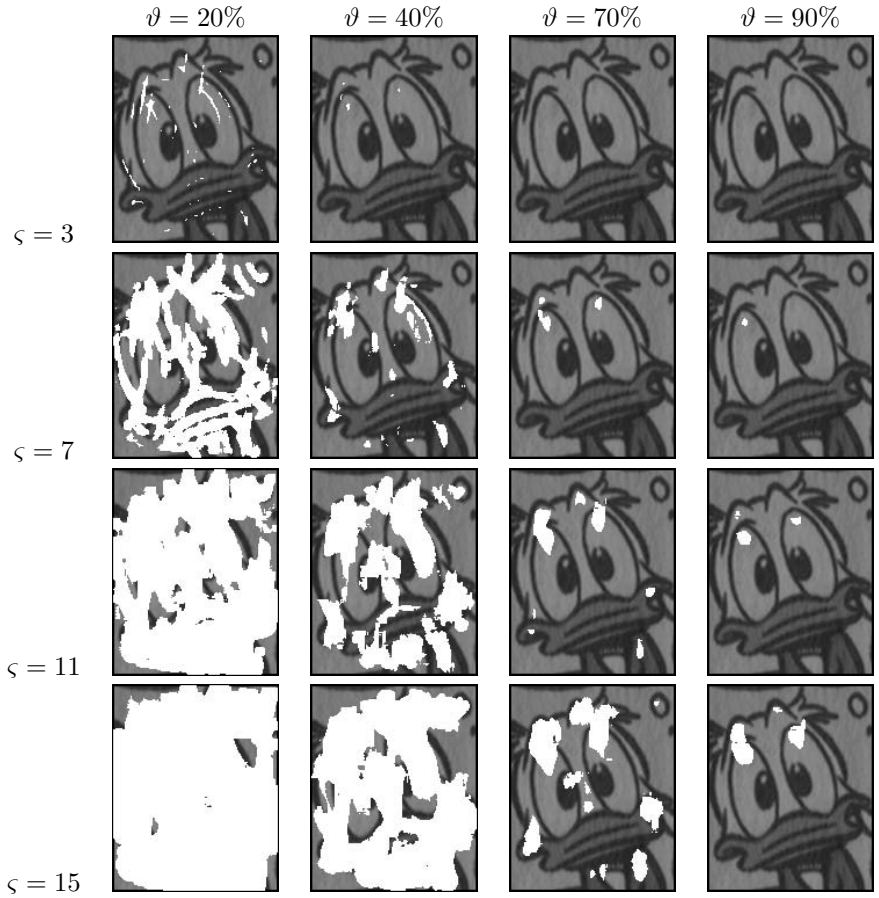


Abbildung 6.28: **Trefferbilder über Ortstoleranz und Schwelle.** Bei der Objekterkennung wurden 20 Kantenrichtungen unterschieden. Die Ortstoleranz ist wieder in Pixeln angegeben. Hohe Schwellen mit hoher Ortstoleranz ergeben kompaktere Trefferbilder.

terkennungssystem trotz einer geringeren Merkmalsanzahl noch Treffer liefert, muß die Schwelle niedriger gesetzt werden.

Ein ähnliches Verhalten ergibt sich, wenn die Objekterkennung über verschiedene Objekte hinweg durchgeführt wird. Die Ergebnisse für die Erkennung von 24 Augenbrauen zeigt Abbildung 6.27. Die gezeigten Diagramme beruhen auf den bereits in Abschnitt 6.1.4 beschriebenen Meßwerten der Tabellen B.5 und B.6. Sie stellen jeweils die Summen der richtigen und falschen Treffer über alle Quantisierungsfeinheiten dar.

Wenn Ortstoleranz und Schwellwert voneinander abhängen, kann ein Parameter frei gewählt und der andere passend eingestellt werden. Da aus Abbildung 6.26 nicht ersichtlich ist, ob eine hohe Schwelle zusammen mit einer großen Ortstoleranz besser ist als eine niedrige Schwelle mit einer geringen Ortstoleranz, muß die Entscheidung aufgrund der Trefferbilder getroffen werden.

Abbildung 6.28 zeigt die Trefferbilder für ein Objekt bei verschiedenen Parametrisierungen bezüglich Schwellwert und Ortstoleranz. Wie zu sehen ist, läßt sich durch die Auswahl einer niedrigen Ortstoleranz zusammen mit einer niedrigen Schwelle eine ähnlichen Anzahl an Treffern einstellen wie für eine hohe Ortstoleranz mit einer hohen Schwelle. Die erkannten Bereiche sind für eine geringe Ortstoleranz sehr schmal und oft punktförmig klein. Wenn gleichzeitig die Schwelle hoch eingestellt wird, verschwinden die meisten Treffer, bis das Modell so selektiv ist, daß keine Objekte mehr erkannt werden. Aufgrund der geringen Ortstoleranz ergeben sich jedoch auch für niedrige Schwellen nur wenige Treffer, wobei diese häufig nicht nur auf der erlernten Augenbraue liegen. Für hohe Ortstoleranzen werden die Trefferbereiche dagegen zunehmend dicker und verschmelzen. Damit das Modell nicht unspezifisch alle Bildkoordinaten erkennt, muß die Schwelle hoch eingestellt werden. In dem Fall ergeben sich kompakte Trefferbereiche, wobei die Anzahl an Fehlern gering ist. Offenbar repräsentiert eine möglichst vollständige Modellübereinstimmung, die durch die hohe Schwelle erzwungen wird, das gesuchte Muster besser als eine hohe geometrische Übereinstimmung, die sich aufgrund der niedrigen Schwelle dann allerdings nur auf einen Teil des Modells bezieht. Dies deckt sich auch mit den Trefferbildern in Abbildung 6.21, die für hohe Schwellen kompaktere und besser lokalisierte Trefferbereiche anzeigen.

6.2.3 Anteil der Geometrieinformation an der Erkennung von Teilen

Aufgrund der in der Literaturdurchsicht herausgestellten Bedeutung von Bag-of-feature-Modellen auf der einen Seite und Modellen mit geometrischen Teilebeziehungen auf der anderen Seite, wurde das in der vorliegenden Arbeit vorgestellte Modell so ausgelegt, daß beide Extreme repräsentiert werden können. Diese Wandlungsfähigkeit wird nun ausgenutzt, um zu prüfen, welche Bedeutung die geometrische Beziehung zwischen Merkmalen für die Objekterkennung auf der Merkmals- bzw. Teileebene hat. Bereits bei der Untersuchung der Merkmalsquantisierung hat sich ein stärkere Abhängigkeit von Geometrieparametern

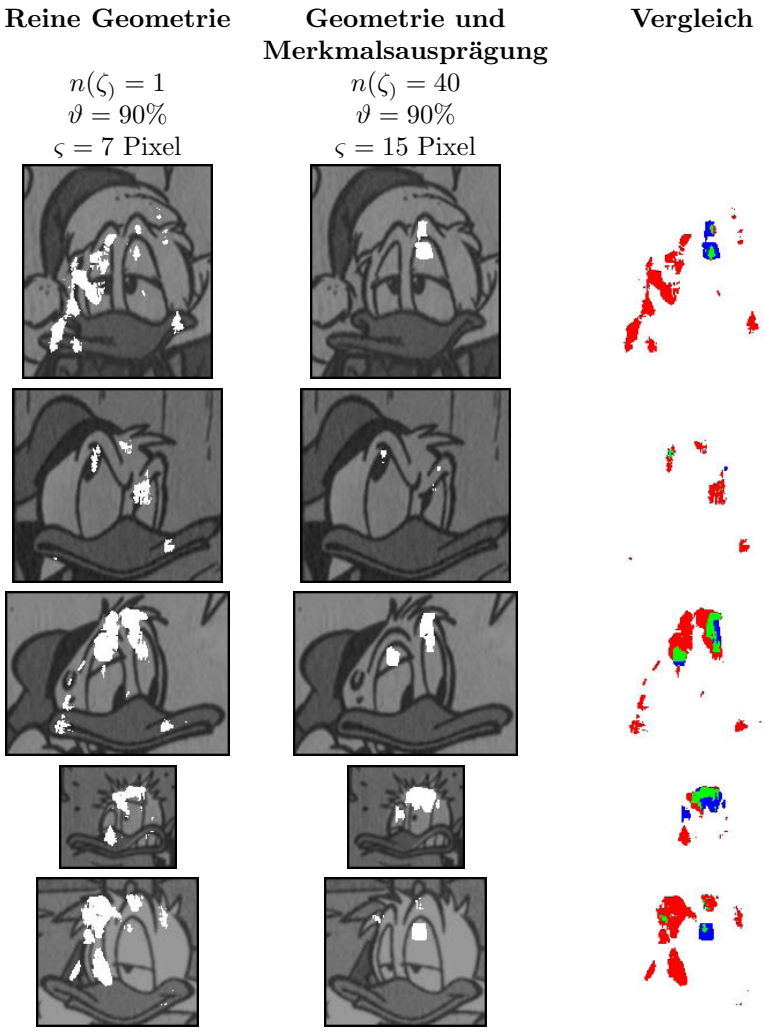


Abbildung 6.29: **Anteil von Geometrie und Merkmalsausprägung an der Erkennung von Augenbrauen.** Die Trefferbilder links beruhen nur auf der Geometrie, die Trefferbilder in der Mitte dagegen auf der Geometrie und der Merkmalsausprägung. Als Merkmale dienen ausschließlich Kantenpunkte. Die rechte Spalte zeigt einen pixelweisen Vergleich der Treffer, wobei Rot ausschließliche Treffer in der linken Spalte, Blau ausschließliche Treffer in der mittleren Spalte und Grün übereinstimmende Treffer anzeigt. Etwa 10% der Treffer in der linken Spalte kommen auch in der mittleren Spalte vor. Umgekehrt treten etwa 44% der Treffer der mittleren Spalte auch links auf.

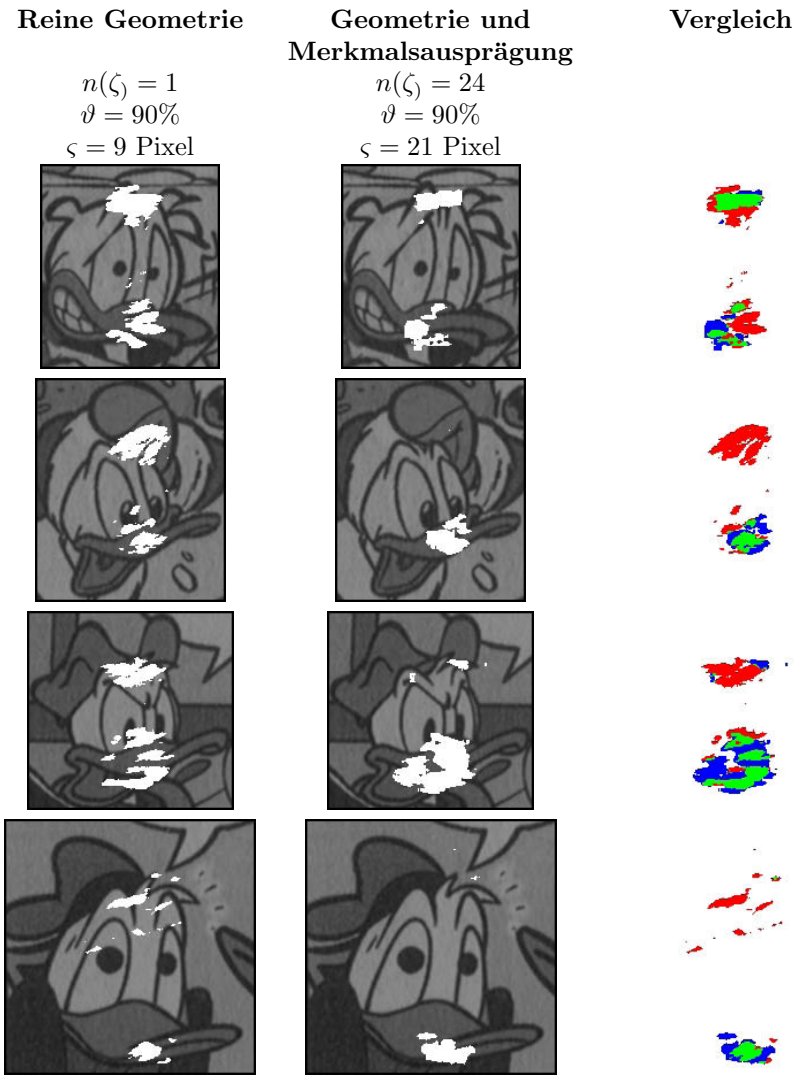


Abbildung 6.30: **Anteil von Geometrie und Merkmalsausprägung an der Erkennung von Schnäbeln.** Die Trefferbilder in der linken Spalte beruhen ausschließlich auf der geometrischen Anordnung der Merkmale innerhalb eines Schnabel-Modells. Als Merkmale dienen ausschließlich Skelettpunkte, allerdings ohne verschiedene Merkmalsausprägungen zu berücksichtigen. Für die mittlere Spalte wurde dagegen sowohl die Geometrie als auch die Merkmalsausprägung ausgewertet. Die rechte Spalte zeigt einen pixelweisen Vergleich der Treffer, wobei Rot ausschließliche Treffer in der linken Spalte, Blau ausschließliche Treffer in der mittleren Spalte und Grün übereinstimmende Treffer anzeigt. Etwa 58% der Treffer in der linken Spalte kommen auch in der mittleren Spalte vor. Umgekehrt treten etwa 65% der Treffer der mittleren Spalte auch links auf.

(Ortstoleranz und Schwelle) als von der Merkmalsausprägung angedeutet. Dieser Gesichtspunkt wird noch einmal genauer beleuchtet.

Dazu werden die Trefferbilder für eine strikt geometrische Verfahrensausrichtung mit denen einer sowohl auf die Geometrie als auch auf die Merkmalsausprägung orientierten Objekterkennung verglichen. Die Einstellung zwischen den beiden Ausrichtungen ist über die Knotenparameter ς und ϑ, sowie über die Anzahl an Quantisierungsstufen für die Merkmale möglich.

Auf der einen Seite wird ein Modell erzeugt, daß nur auf den geometrischen Beziehungen zwischen dem Merkmalen aufbaut. Um dies zu erreichen, dürfen keine verschiedenen Merkmalsausprägungen unterschieden werden, d.h. $n(\zeta) = 1$. Wenn jedoch die Merkmalsausprägung keinen Aufschluß mehr über das zu erkennende Objekt gibt, muß die geometrische Übereinstimmung umso höher sein. Die Ortstoleranz muß daher niedrig sein. Der Schwellwert sollte dagegen hoch eingestellt sein, um das Auftreten der Merkmale selbst stärker zu gewichten. Für die andere Modellausrichtung wird zusätzlich noch die Merkmalsausprägung beachtet. Die Anzahl an Quantisierungsstufen wird dazu hoch eingestellt. Um von der Geometrie nicht vollständig unabhängig zu werden, darf die Ortstoleranz nicht zu groß eingestellt werden.

Dabei ergibt sich jedoch das Problem, daß eine nur auf diesen theoretischen Erwägungen festgelegte Parametrisierung die Eigenschaften der Objekte und die Grenzen der Algorithmen ignoriert, und so zu schlechten oder schlecht vergleichbaren Trefferbildern führt. Wie schon in den Versuchen zur Merkmalsquantisierung werden daher sowohl für Parametrisierungen ohne Merkmalsunterscheidung ($n(\zeta) = 1$) und eine sehr feine Quantisierung jeweils die besten 5 Ergebnisse ausgewählt. Aus diesen wiederum werden dann Einstellungen gewählt, die den genannten Forderungen am nächsten kommen. Aufgrund der Ergebnisse des vorigen Versuchs werden bei der Abwägung von Ortstoleranz und Schwellwert bevorzugt hohe Schwellwerte ausgewählt. Da sich die Ergebnisse zudem für Kanten- und Flächenmerkmale unterscheiden können, wird hier wieder eine Erkennung von Augenbrauen an Kantenmerkmale und eine Erkennung von Schnäbeln an Flächenmerkmalen durchgeführt. Ein Teil der Ergebnisse ist in den Tabellen B.5–B.8 bereits dargestellt. Ergänzende Messungen für extreme Parametrisierungen zeigen die Tabellen B.9, B.10 und B.11 im Anhang.

Für eine auf die Geometrie ausgerichtete Einstellung ergeben sich die Parameter $n(\zeta_)) = 1$, $\vartheta = 90\%$ und $\varsigma = 7$ Pixel für Augenbrauen, sowie $n(\zeta_)) = 1$, $\vartheta = 90\%$ und $\varsigma = 9$ Pixel für Schnäbel. Diese Parametrisierungen entsprechen in jeder Hinsicht den gewünschten Kriterien. Für eine zusätzlich auf die Merkmalsausprägung ausgerichtete Erkennung wurden die Parametrisierungen $n(\zeta_)) = 40$, $\vartheta = 90\%$ und $\varsigma = 15$ Pixel für Augenbrauen, sowie $n(\zeta_)) = 24$, $\vartheta = 90\%$ und $\varsigma = 21$ Pixel ausgewählt. Da die Ortstoleranz nur etwa die Hälfte der mittleren Augenbrauengröße umfaßt, kann die Augenbrauenparametrisierung noch nicht vollständig als Bag-of-features-Modell bewertet werden. Bei der Erkennung von Schnäbeln macht die Ortstoleranz nur ein Sechstel der mittleren Schnabelgröße aus. Geometrie- und Merkmalsinformation sind hier also stark gemischt. Trotzdem ist die Ortstoleranz bei der auf die Merkmalsausprägung ausgerichteten

Parametrisierung noch etwas höher als bei der geometrieoptimierten Einstellung.

Die Abbildungen 6.29 und 6.30 zeigen einen Vergleich eines Teils der erzeugten Trefferbilder. Bei der kantenbasierten Erkennung fällt zunächst auf, daß das Geometriemodell deutlich mehr Treffer erzeugt als das bezüglich Geometrie und Merkmalsausprägung gemischte Modell, was an der Auswahl der Parametrisierungen liegt, und nicht an der Geometrieinformation. Über alle getesteten Bilder zusammengefaßt machen 10 Prozent der Treffer des Geometriemodells schon einen hohen Anteil von 44 Prozent der Treffer des gemischten Modells aus. Bei der flächenbasierten Erkennung von Schnäbeln fällt der Anteil noch höher aus. Hier machen 58 Prozent der Geometrietreffer 65 Prozent der Treffer des gemischten Modells aus. Zudem passen die Parametrisierungen bezüglich der Gesamtzahlen an Treffern besser zueinander. Zusammenfassend kann man also sagen, daß die Gemeinsamkeiten zwischen den Ergebnissen einer nur auf Geometrie ausgerichteten und einer gemischt ausgerichteten Objekterkennung bei 44–65 Prozent liegen. Aufgrund der vielen Entwurfsentscheidungen, die für diesen Vergleich getroffen werden mußten, ist dies allerdings nur ein ungefährer Wert.

6.2.4 Geometrische Beziehungen zwischen Merkmalen

Da die geometrischen Beziehungen zwischen Merkmalen offenbar von großer Bedeutung sind, stellt sich nun die Frage, welche Auswirkungen sich daraus für die Zusammenstellung von Merkmalen zu Teilen ergeben. Die hohe Bedeutung der Geometrieinformation spricht dafür, daß bestimmte Merkmale nur in bestimmten Relativpositionen zueinander vorkommen. Es wäre daher zweckmäßig, Merkmale an diesen Positionen zu Teilen zu kombinieren.

Eine erste Erklärung könnten die mit einem genetischen Algorithmus optimierten Modelle von Stommel und Kuhnert [SK05, SK06] liefern. Diese zeigen jedoch keine anschaulich erklärbaren Regelmäßigkeiten. Dabei könnte auch eine Rolle spielen, daß ausschließlich starre Objekte modelliert wurden. Dies kann dazu führen, daß die Teilegeometrie eine geringere Rolle spielt und daher einen geringeren evolutionären Druck auf die Modelloptimierung ausübt. Leider erlaubt der genetische Algorithmus keinen tieferen Einblick in die für das Training kritischen Faktoren. Ommer und Buhmann [OB06] deuten an, daß der Abstand der Merkmale von Bedeutung ist. In der zitierten Arbeit wird allerdings auf eine genauere Erläuterung verzichtet.

Um hier eine Klärung zu erreichen, wird nun das gemeinsame Auftreten von Merkmalen statistisch untersucht. Aufgrund der Komplexität der Aufgabe werden allerdings nur Paare von Merkmalen untersucht. Der Rechenaufwand ist allerdings immer noch beträchtlich. Die in den Abbildungen 6.34 und 6.35 dargestellte Berechnung dauerte beispielsweise über 300 Stunden (Rechner: AMD Athlon XP, 2GHz).

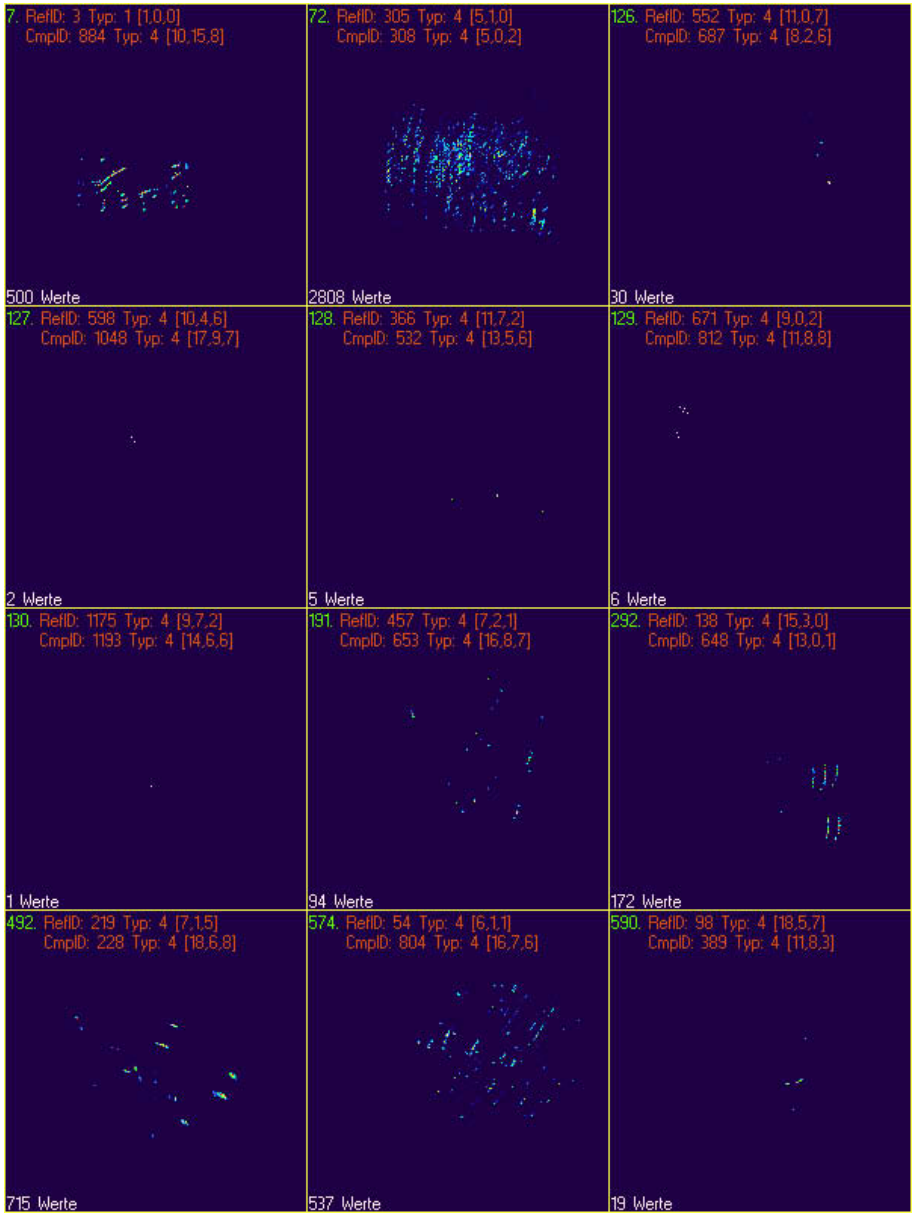


Abbildung 6.31: **Räumliche Verteilung von selten zusammen auftretenden Merkmalen.** Für jedes Diagramm wurde ein Referenz- und ein Vergleichsmerkmal in bestimmten Ausprägungen festgelegt. Die Diagramme zeigen jeweils die Verteilung des Vergleichsmerkmals über der Bildebene relativ zum Referenzmerkmal an. Die Darstellungen sind auf das jeweilige Referenzmerkmal zentriert. Die grünen Zahlen geben Diagrammnummern an. Die Häufigkeit der Merkmalspaare in der Stichprobe ist jeweils unten in weißer Schrift angegeben.

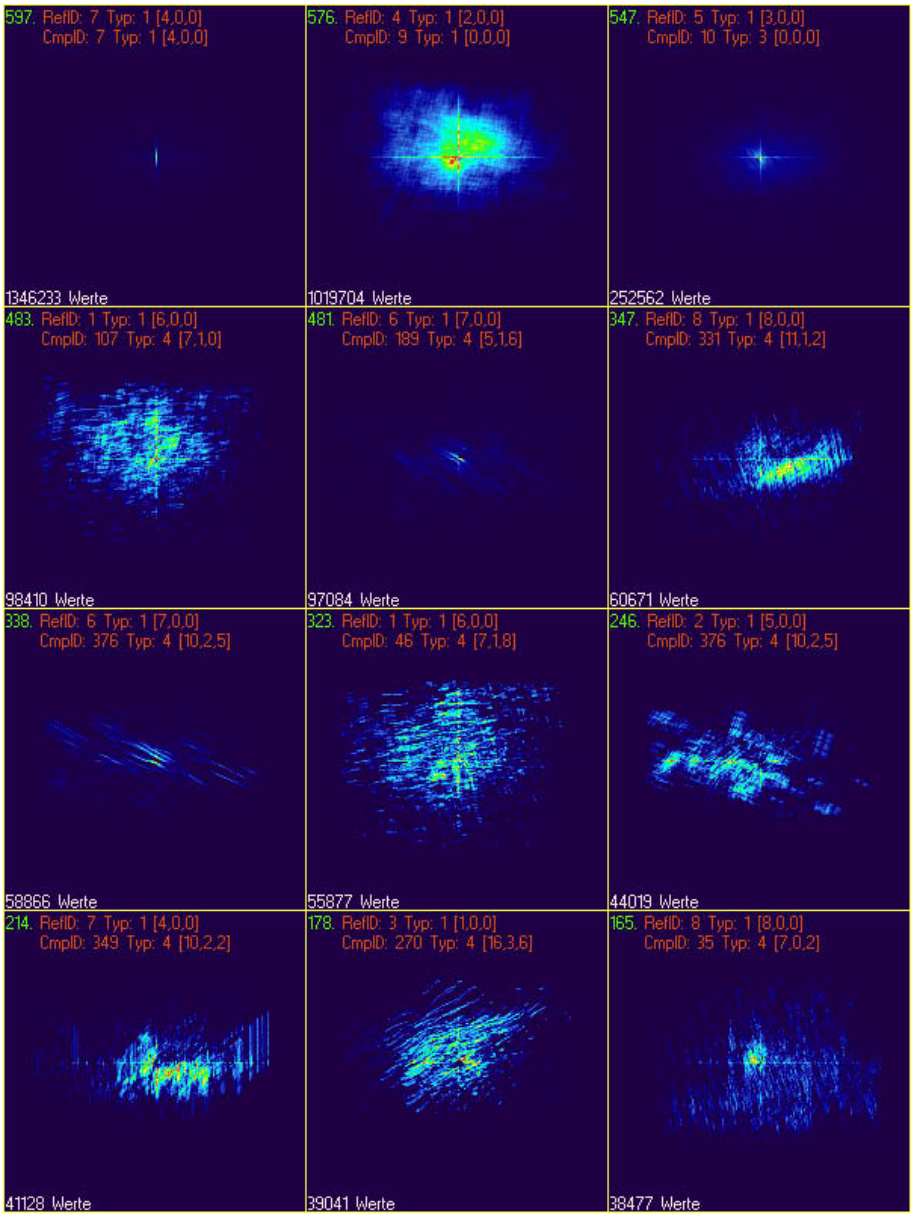


Abbildung 6.32: **Räumliche Verteilung von häufig zusammen auftretenden Merkmalen.** Die Breite der Diagramme entspricht einem Durchmesser von 600 Pixeln in einem Bild.

Räumliche Verteilung von Merkmalspaaren

Zur Untersuchung der geometrischen Abhängigkeiten innerhalb von Paaren von Merkmalen wird zunächst eine Stichprobe von 137 Bildern mit Augenbrauen aus Donald-Bildern zusammengestellt. Eine Merkmalsextraktion liefert etwa 125 000 Merkmalspunkte in über 1200 verschiedenen Ausprägungen. Die Quantisierung geschieht mit 10 Stufen für Kanten, 1 Stufe für Ecken, 20 Stufen für die Intensität an Skelettpunkten, 10 Stufen für die Richtung von Skelettlinien und 2 Pixel-Stufen für die Flächengrößen. Von den etwa 1,5 Millionen möglichen Merkmalspaaren (alle Merkmalsausprägungen kombiniert mit allen) treten ca. 500 000 tatsächlich in der Stichprobe auf. Über diese werden nun Histogramme erstellt. Dazu wird jeweils ein Merkmal eines Paares als Referenzmerkmal und das andere als Vergleichsmerkmal definiert. Für jedes Auftreten eines Paares in der Stichprobe wird nun die Position des Vergleichsmerkmals relativ zum Referenzmerkmal bestimmt. Die für ein Paar auftretenden Relativpositionen werden in einem zweidimensionalen Histogramm über die relative x- und y-Koordinate gespeichert.

Die Abbildungen 6.31 und 6.32 zeigen Beispiele der erzeugten Histogramme. Die Diagramme sind immer auf die Referenzmerkmale zentriert. Die Breite der Diagramme entspricht einem Durchmesser von 600 Pixeln in einem Bild. Die grünen Nummern geben die laufende Nummer eines Merkmalspaares an. Die rote Schrift gibt Aufschluß über die untersuchten Merkmale. Dabei bezeichnet 'RefID' das Referenzmerkmal und 'CmpID' das Vergleichsmerkmal. Der Typ legt die Art des Merkmals fest, wobei die Werte 1 für Kantenpunkte, 3 für Ecken und 4 für Skelettpunkte benutzt werden. In eckigen Klammern ist zudem der Merkmalsdeskriptor angegeben. Nicht genutzte Einträge für Ecken und Kanten sind auf Null gesetzt.

Abbildung 6.31 zeigt eine Unterauswahl von ursprünglich 600 zufällig ausgewählten Histogrammen. Wie an den Diagrammen 126–130 zu sehen ist, treten viele Merkmalskombinationen nur sehr selten auf. Die Häufigkeiten liegen oft im Zehnerbereich. Hier liegen meistens nicht genügend Daten vor, um günstige Merkmalspositionen zuverlässig zu ermitteln. Dieses Ergebnis ist typisch und resultiert aus der ungleichmäßigen Häufigkeit verschiedener Merkmalskombinationen. Es zeigt sich, daß die am häufigsten in der Stichprobe wiedergefundenen Paare bis zu etwa 1,4 Millionen mal auftreten. Allerdings gibt es nur wenige verschiedene Paare, die so häufig auftreten. Sortiert man die Histogramme nach der absoluten Auftrittshäufigkeit eines Paares, dann sinkt bereits nach den am häufigsten auftretenden 50 Merkmalskombinationen die Häufigkeit von etwa 1,4 Millionen auf nur noch 150 000. Nach den häufigsten 600 Merkmalskombinationen liegt die Häufigkeit bei unter 33 000.

Abbildung 6.32 zeigt eine Auswahl der am häufigsten auftretenden Merkmalskombinationen. Diese sind fast immer Merkmalspaare, die mindestens ein Kantenmerkmal enthalten. Dies liegt daran, daß Kantenmerkmale zum einen sehr häufig sind und zum anderen mit 10 Stufen grob quantisiert sind. Da Skelettmerkmale einen dreidimensionalen Deskriptor haben, existieren hier theoretisch 10 000 verschiedene Ausprägungen, von denen jedoch nur wenige

tatsächlich auftreten. Die häufigsten 45 Merkmalspaare sind daher Kombinationen von Kanten mit Kanten. Innerhalb dieser treten wiederum Paare aus identischen Merkmalen (Diagramm 597) und Paare mit entgegengesetzten Kantenorientierungen sehr häufig auf. Paare mit entgegengesetzten Orientierungen treten offenbar aufgrund der schwarzen Striche in Comic-Zeichnungen auf, welche sich durch zwei entgegengesetzt orientierte Kanten vom helleren Hintergrund abheben. Auf die 45 Kantenkombinationen folgen alle 10 Kombinationen des Eckenmerkmals mit einer Kante. Diagramm 547 zeigt das Histogramm einer solchen Kombination. Die Gestalt dieser häufigsten 55 Histogramme entspricht weitgehend den obersten drei Diagrammen in Abbildung 6.32. Die Histogramme zeigen einen mehr oder weniger steilen Abfall der Häufigkeit mit zunehmendem Abstand von dem Referenzmerkmal ohne eine feinere Struktur aufzuweisen. Wie die übrigen Diagramme in der Abbildung zeigen, sind die Kombinationen von Kanten und Skelettpunkten dagegen deutlich vielgestaltiger, wobei sich stärkere Untergruppierungen wie in Diagramm 246 oder Asymmetrien wie in Diagramm 347 zeigen können. Wie die Diagramme 483, 323 und 178 zeigen, sind mögliche feinere Strukturen in den Histogrammen jedoch oft nicht stark ausgeprägt. Hier ist anscheinend hauptsächlich der Abstand vom Referenzpunkt bedeutsam.

Für die Erkennung von Objekten ergeben sich zusammengefaßt folgende Resultate:

- Die Teilegeometrie spielt eine Rolle.
- Die Histogramme zeigen oft räumliche Cluster, die möglicherweise wichtige Positionen für zu modellierende Teile darstellen.
- Histogramme mit solchen Clustern eignen sich jedoch aufgrund der geringen Häufigkeit der betroffenen Merkmale nicht für die Teilemodellierung.
- Histogramme über mehr Merkmale wären aussagekräftiger, allerdings auch noch dünner besetzt. Außerdem sind sie rechentechnisch aufwendig.
- Bei sehr häufigen Merkmalskombinationen spielt vor allem der räumliche Abstand der Merkmale eine Rolle, die genaue Position dagegen weniger.

Einfluß des Abstands auf die Korrespondenz von Merkmalsgruppen

Da der räumliche Abstand offenbar die Häufigkeit von Merkmalspaaren beeinflußt, stellt sich als nächstes die Frage, ob sich konsistente Ergebnisse für größere Mengen von Merkmalen ergeben.

Abweichende Ergebnisse könnten sich beispielsweise dadurch ergeben, daß aufgrund der geringen Häufigkeit der meisten Merkmalskombinationen wichtige systematische Effekte nicht festgestellt wurden. Darüberhinaus könnten für größere Gruppen von Merkmalen andere Gesetze gelten als für Merkmalspaare. Um dies zu überprüfen bietet es sich an, den vorigen Versuch auf größere Merkmalsgruppen zu erweitern. Der Rechenaufwand steigt jedoch mit jedem weiteren Merkmal um eine Größenordnung, wobei die Histogramme gleichzeitig noch spärlicher werden.

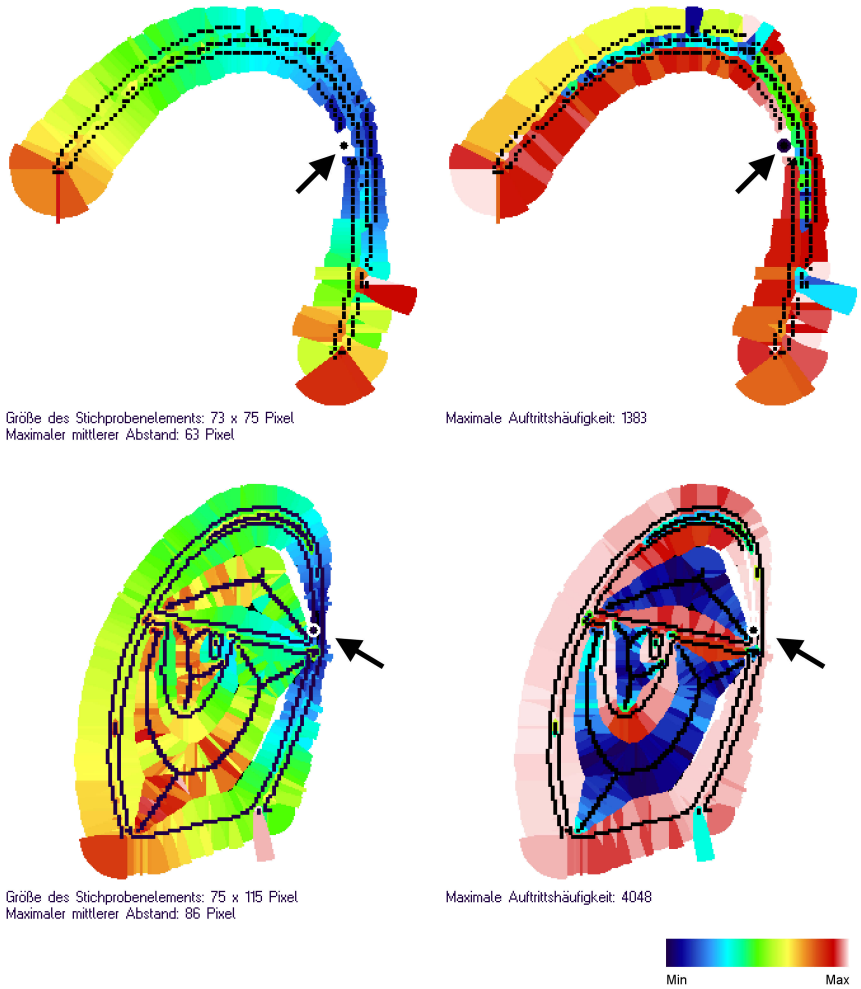


Abbildung 6.33: **Merkmalskorrespondenz und Entfernung.** Die schwarzen Punkte repräsentieren die Merkmale einer Augenbraue (obere Diagramme) und eines Auges (untere Diagramme). Die farbigen Bereiche geben an, wie gut die Merkmale in anderen Bildern wiedergefunden wurden, wenn bereits das durch den Pfeil gekennzeichnete Referenzmerkmal gefunden wurde. Dabei wird erwartet, daß ein Merkmal in anderen Bildern an der gleichen Relativposition zu dem Referenzmerkmal auftritt wie in dem ursprünglichen Bild. Tatsächlich tritt jedoch eine gewisse räumliche Abweichung auf. Deren Größe wird durch den Radius der farbigen Bereiche angegeben, aus Darstellungsgründen allerdings um den Faktor 5 verkleinert. In den linken Bildern wird der Abstand zusätzlich noch durch die Farbe kodiert. Korrespondierende Merkmale können jedoch nicht immer gefunden werden. Daher gibt in den Diagrammen auf der rechten Seite die Farbe die Häufigkeit von Korrespondenzen an.

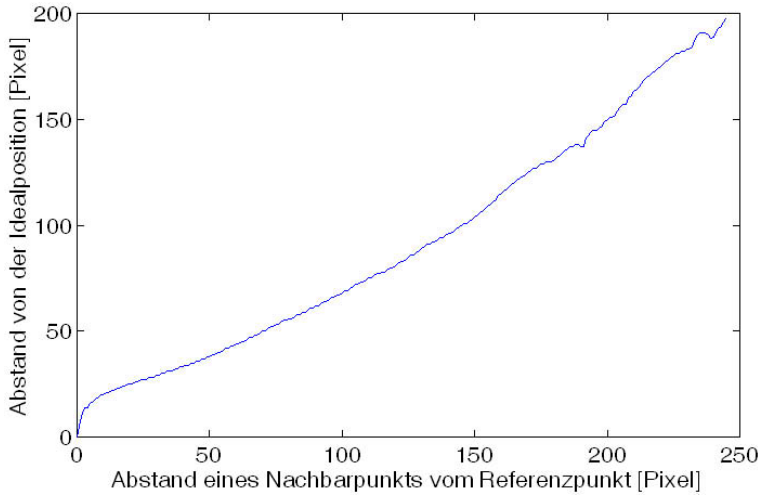


Abbildung 6.34: **Abweichung von der erwarteten Merkmalsposition abhängig vom Abstand zum Referenzpunkt.** Das Diagramm gibt auf der vertikalen Achse an, in welchem Abstand von der erwarteten Position ein Merkmal im Mittel gefunden wird, wenn die Suche mit unbegrenzter Ortstoleranz durchgeführt wird. Die Suche wird allerdings nur in Objekten durchgeführt, in denen bereits ein Referenzmerkmal gefunden wurde. Der Abstand zwischen dem Referenzmerkmal und dem zu suchenden Merkmal ist auf der horizontalen Achse aufgetragen.

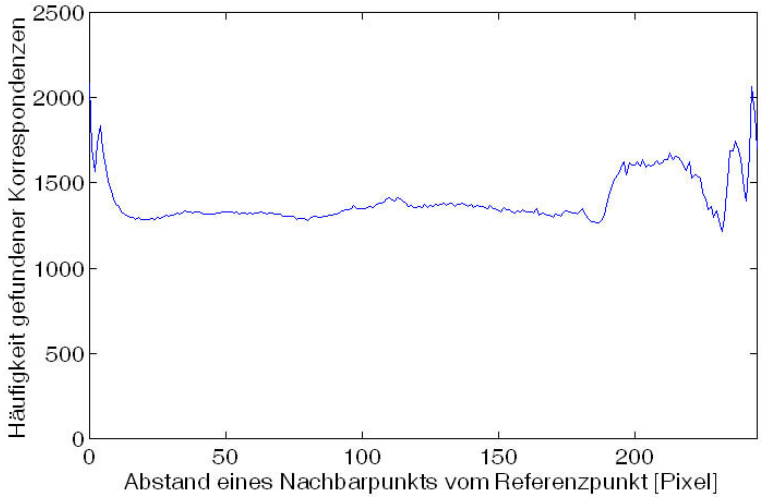


Abbildung 6.35: **Häufigkeit einer Korrespondenz abhängig vom Abstand zum Referenzpunkt.** Das Diagramm gibt an, wie häufig Korrespondenzen ermittelt werden konnten. Eine fehlgeschlagene Suche bedeutet, daß in einem untersuchten Objekt kein Merkmal in der geforderten Ausprägung vorliegt.

Aus diesem Grund wird hier ein alternativer Ansatz gewählt, der darauf beruht, alle Merkmale einzelner Referenzobjekte in einzelnen anderen Objekten wiederzufinden. Dabei wird für jedes Merkmal eines Referenzobjekts ermittelt, wie oft und in welchem räumlichen Abstand es zusammen mit einem vorher festgelegten Referenzmerkmal in anderen Objekten auftritt. Dabei sollte einer der folgenden zwei Fälle eintreten:

1. Benachbarte Merkmale eines Objekts treten bevorzugt zusammen auf. Mit steigendem Abstand werden die anderen Merkmale eines Testobjekts seltener gefunden. Für die Teileauswahl ist daher der Abstand entscheidend. Die Ergebnisse des vorhergehenden Versuchs gelten dann nicht nur für Merkmalspaare, sondern auch für größere Gruppen.

2. Entfernte Merkmale eines Objekts treten häufiger zusammen auf als benachbarte. In dem Fall hätten die Cluster der großen Anzahl dünn besetzter Histogramme den Haupteinfluß. Für Merkmalsgruppen würden sich demnach die Cluster der Histogramme von Merkmalspaaren gegenseitig verstärken.

Um hier eine Klärung zu erreichen, wird zunächst wieder eine Stichprobe mit Bildern von rechten Augenbrauen von Enten zusammengestellt. Die Merkmalsextraktion auf der Stichprobe liefert ca. 43 000 Merkmale in 273 verschiedenen Ausprägungen. Die Merkmale wurden mit 35 Kantenrichtungen, 20 Orientierungen von Skelettlinien und 10 Helligkeitswerten quantisiert. Der Abstand eines Skelettpunkts von der nächsten Kante wurde in Schritten von 5 Pixeln gemessen. Die Merkmale unterteilen sich in ca. 30 000 Kantenpunkte, ca. 11 000 Skelettpunkte und etwa 1000 Ecken.

Jedes Merkmal der Stichprobe wird nun einmal als Referenzpunkt festgelegt. Das Stichprobenbild, das den Referenzpunkt enthält, ist das Referenzbild. Als Nachbarpunkte des Referenzpunkts gelten alle übrigen Merkmale des Referenzbildes. Nun wird versucht, Korrespondenzen des Referenzpunkts und der entsprechenden Nachbarn in den übrigen Bildern der Stichprobe zu finden. Dazu wird für jedes Bild untersucht, ob es ein Merkmal enthält, das die gleiche Ausprägung hat wie der Referenzpunkt. Die Nachbarpunkte des Referenzpunktes treten in dem Stichprobenbild idealerweise ebenfalls auf und zwar in den gleichen relativen Positionen. Die Abweichung von dem Ideal wird in zwei Werten gemessen: Zum einen wird für jeden Nachbarn die Häufigkeit bestimmt, mit der er in anderen Bildern gefunden wurde. Zum anderen wird der mittlere Abstand eines korrespondierenden Punkts von der Idealposition berechnet. Dabei wird keine Obergrenze für diesen Abstand festgelegt. Wenn ein Punkt nicht zugeordnet werden kann, bedeutet dies, daß in dem betreffenden Stichprobenbild kein Merkmal mit einer übereinstimmenden Ausprägung vorliegt. Für den Fall, daß ein Stichprobenbild mehrere Merkmale einer passenden Ausprägung enthält, wird das Merkmal mit dem geringsten Abstand von der idealen Position als Korrespondenz ausgewählt. Um objektspezifische Objekte auszuschließen, wird eine zweite Stichprobe mit ca. 100 000 Merkmalen von Augen in Donald-Bildern zusammengestellt.

Bezüglich der Häufigkeit, mit der Nachbarpunkte in der Stichprobe wiedergefunden werden können, läßt sich keine Abhängigkeit vom Abstand zum Referenzpunkt erkennen. Stattdessen scheint die Häufigkeit der Korrespondenzen für die verschiedenen Merkmalstypen (Kanten, Ecken, Flächen) von lokalen Störungen abgesehen konstant über das gesamte Objekt zu sein. Skelettpunkte können dabei schlechter zugeordnet werden als Kantenpunkte. Letzteres kann durch die große Zahl unterschiedlicher Merkmalsausprägungen erklärt werden und ist konsistent mit den Histogrammen der Merkmalspaare. Für den Abstand korrespondierender Merkmale von der Idealposition, die sich aus der Relativposition eines Nachbarn zum Referenzpunkt ergibt, zeigt sich dagegen eine stetige Zunahme mit wachsendem Abstand des Nachbarn vom Referenzpunkt. Da sich für die Augenbrauenstichprobe und die Augenstichprobe die gleichen Ergebnisse zeigen, kann eine Abhängigkeit von bestimmten Objekten ausgeschlossen werden.

Abbildung 6.33 zeigt die Korrespondenzergebnisse für zwei Referenzbilder, eines für die Augenbrauenstichprobe (obere Bilder) und eins für die Augenstichprobe (untere Bilder). Die Merkmalspositionen des Referenzbildes werden jeweils durch schwarze Punkte angegeben. Der Referenzpunkt ist durch einen Pfeil markiert. Die Diagramme wurden aus Darstellungsgründen auf die gleiche Größe skaliert. Zu jedem Merkmalspunkt werden nun zwei Variablen dargestellt. Die erste Variable ist der Abstand, den Korrespondenzen zu Nachbarpunkten im Mittel von der Idealposition haben. Diese Variable wird für jedes Merkmal in den Diagrammen durch den Radius einer farbigen Fläche angegeben. Um die Darstellung übersichtlich zu halten wurde hier ein Verkleinerungsfaktor von 5 gewählt, d.h. der mittlere Abstand von Korrespondenzen ist fünfmal größer als es der Radius der farbigen Fläche angibt. Die Bedeutung der Flächengröße ist für alle gezeigten Diagramme gleich. Die zweite Variable wird durch die Farbe der Fläche dargestellt. Für die Diagramme links zeigt diese ebenfalls den mittleren Abstand, d.h. die Größe und die Farbe der Flächen sind hier gleichbedeutend. In den Diagrammen rechts kodiert die Farbe die Häufigkeit, mit der Korrespondenzen hergestellt werden konnten. Die genannte geometrische Abhängigkeit zeigt sich in den linken Diagrammen sowohl in einem kontinuierlichen Farbverlauf vom Referenzpunkt zu den entfernten Randpunkten des Objekts als auch in einer stetig zunehmenden Flächengröße. Eine solche Abhängigkeit ist in den rechten Bildern bezüglich der Häufigkeit nicht zu erkennen. Hier zeigt sich eine große Häufigkeit von Kantenzuordnungen an grünen und roten Flächen. Die niedrigere Häufigkeit von korrespondierenden Skelettpunkte zeigt sich an den blauen Flächen.

Um zu prüfen, ob die genannten Effekte systematisch auftreten, wird ein weiteres Experiment über etwas mehr als 100 000 verschiedene Referenzpunkte einer Stichprobe von Augenbildern durchgeführt. Für jedes Referenzbild eines Referenzpunkts wird wieder der Abstand aller Nachbarn von der jeweiligen Idealposition und die Häufigkeit von Korrespondenzen ermittelt. Diese werden dann jeweils über dem Abstand zwischen Nachbar und Referenzpunkt aufgetragen. Die Ergebnisse zeigen die Abbildungen 6.34 und 6.35. Offenbar steigt der mittlere Abstand wiedergefundener Merkmale von der erwarteten Idealposition

weitgehend linear mit dem Abstand vom Referenzpunkt. Nur für direkte Nachbarn in einem Radius von unter 5 Pixeln ergibt sich ein steilerer Verlauf. Auf der anderen Seite ist auch die Häufigkeit, mit der Merkmale gefunden werden können über weite Abstandsbereiche konstant.

Auf den ersten Blick scheinen die Ergebnisse weder zu der ersten noch zu der zweiten erwarteten Lösung zu passen. Die Benachbarung von Merkmalen spielt zwar eine Rolle, allerdings nicht in Bezug auf die Häufigkeit, sondern auf den Abstand von der Idealposition. Dabei muß allerdings bedacht werden, daß die Suche nach korrespondierenden Merkmalen über jeweils vollständige Objekte durchgeführt wurde. Da die Häufigkeit einigermaßen konstant ist, bedeutet dies, daß die meisten Merkmale zugeordnet werden können, wenn man nur beliebig große Bildbereiche durchsucht. Anders ausgedrückt könnte man alle Merkmale eines Modells zuordnen, wenn man die Ortstoleranz mindestens so groß wählen würde, wie das Diagramm angibt. Die gemessenen Abstände geben daher keine Grenze für die sichere Erkennung bestimmter Objekte an, sondern eine Grenze, ab der ein Klassifikator beliebig wird. Das in dieser Arbeit vorgeschlagene Objekterkennungssystem beschränkt den Suchraum daher durch die Ortstoleranz ς auf eine angemessene Umgebung um die Idealposition eines Merkmals. Dies entspricht einer Schwelle für den Abstand von der Idealposition. Betrachtet man nun die in Abbildung 6.34 dargestellten Ergebnisse, bedeutet dies, daß ab einem bestimmten Abstand vom Referenzpunkt eine einmal festgelegte Ortstoleranz überschritten wird und weiter entfernte Merkmale nicht gefunden werden. Daraus ergibt sich ein Zusammenhang zwischen der Benachbarung von Merkmalen und der Häufigkeit entdeckter Korrespondenzen. Die Ergebnisse entsprechen daher der ersten erwarteten Lösung.

6.2.5 Clusterung von Merkmalen nach räumlicher Nähe

Für die Zusammenfassung von Merkmalen zu Teilen lassen sich nun die folgenden Aussagen treffen:

- Die Benachbarung von Merkmalen ist relevant.
- Es gibt keine Hinweise, daß sich Teile aus unzusammenhängenden, weit entfernten Dingen zusammensetzen.
- Die Ortstoleranz darf mit der Teilegröße steigen, ohne daß das Modell beliebig wird.
- Die Ortstoleranz, ab der das Modell beliebig wird, steigt über weite Bereiche linear mit der Teilegröße.

Um die Benachbarung von Merkmalen systematisch zu erfassen, werden die Merkmale nach ihrem Abstand in der Bildebene gruppiert. Dazu wird ein hierarchisches Clusterungsverfahren eingesetzt, welches ein Dendrogramm als Ergebnis liefert. Das Dendrogramm ist ein Binärbaum, dessen Blätter die Merkmale der Bildebene sind. Die Knoten des Dendrogramms fassen dabei immer

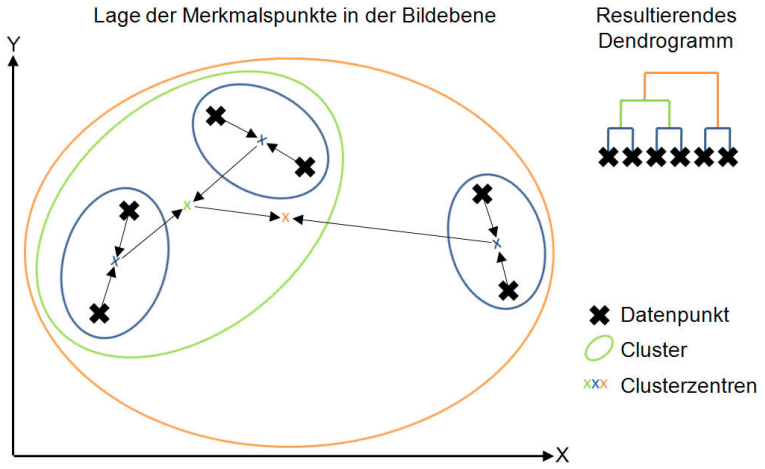


Abbildung 6.36: **Agglomerative Clusterung von Merkmalspunkten.** Die zwei am stärksten benachbarten Merkmalspunkte oder Clusterzentren werden zu einem neuen Cluster zusammengefaßt. Das Clusterzentrum wird als Mittelwert der Merkmalspunkte berechnet. Der Schritt wird wiederholt bis ein Cluster erzeugt wurde, der alle Merkmalspunkte enthält. Die Clusterabstände und die Reihenfolge, in der Cluster zusammengefaßt werden, gibt das Dendrogramm an.

möglichst eng benachbarte Merkmale oder Gruppen von Merkmalen zusammen. Zu jedem Knoten des Dendrogramms kann dann ein Teil erzeugt werden, daß die Merkmale an den Blattknoten des betreffenden Teilbaums umfaßt.

Die hiearchische Clusterung hat den großen Vorteil, daß die Anzahl der resultierenden Cluster nicht im Voraus festgelegt werden muß. Dies würde Wissen über die Häufigkeit und Parametrisierung der resultierenden Teile voraussetzen, das zu diesem Zeitpunkt des Trainings noch nicht vorliegt. Die Repräsentation der Ergebnisse als Dendrogramm erlaubt es dagegen, zuerst die Nachbarschaftsinformationen der gesamten Stichprobe auszuwerten und die Auswahl der Teile von den Gesamtergebnissen abhängig zu machen. Die Dendrogramme liefern zudem Informationen über alle Größenbereiche von Benachbarungen, was wodurch sich mehr Möglichkeiten zur Modellparametrisierung ergeben.

Es wird ein agglomeratives Verfahren eingesetzt, das schrittweise die jeweils ähnlichsten Gruppen bzw. Merkmale zusammenfaßt. Die Gruppen werden dabei durch den Mittelwert beschrieben. Abbildung 6.36 verdeutlicht die Idee des Verfahrens. Manning et al. [MRS08] bewerten Verfahren dieser Art als "beste Wahl für die meisten Anwendungen" und geben an, daß sie robust gegen degenerative Verkettungen der zu clusternden Elemente sind. Die Laufzeit für n_E Elemente beträgt laut Manning ohne weitere Optimierungen $\mathcal{O}(n_E^3)$, läßt sich aber durch den Einsatz von Vorrangwarteschlangen auf $\mathcal{O}(n_E^2 \log n_E)$ verkürzen. In der vorliegenden Arbeit werden jedoch andere Optimierungen eingesetzt, die zwar nicht die asymptotische Laufzeit verringern, aber zu sehr niedrigen kon-

stanten Laufzeitkoeffizienten führen. Aus diesem Grund ist das im Folgenden beschriebene Verfahren in der Praxis effizient genug, um mehrere 10 000 Elemente in überschaubarer Zeit zu clustern.

Die Merkmale liegen nach der Extraktion aus einem Bild der Stichprobe in Form von Instanzen von Merkmalsknoten des Modells vor. Sie besitzen daher die in Gleichung 5.2 angegebene Form. Die Clusterung wird bezüglich der Absolutposition x, y der Instanzen durchgeführt. Diese kann als Vektor $\boldsymbol{x} = (x, y)^\top$ geschrieben werden. Für ein Stichprobenbild ergibt sich so eine Menge von Vektoren $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n_E}$ mit jeweils $n_D = 2$ Dimensionen.

Clusterungsverfahren

Cluster innerhalb der Vektoren werden durch eine Menge $\beta \subseteq \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n_E}\}$ mit dem Mittelwert $\boldsymbol{v}$ über alle Elemente von β beschrieben. Da das hier beschriebene Verfahren im Rahmen dieser Arbeit auch noch in anderen Zusammenhängen genutzt wird, wird zusätzlich eine Kodierung $\psi(\boldsymbol{v})$ der Mittelwerte eingeführt, welche eine für andere Anwendungen rechentechnisch effizientere Darstellung der zu bearbeitenden Daten erlaubt. Zu Beginn der Clusterung wird für jeden Vektor $\boldsymbol{x}$ ein initialer Cluster erzeugt, der nur dieses Element enthält. Der jeweilige Mittelwert $\boldsymbol{v}$ ist daher mit dem gespeicherten Element $\boldsymbol{x}$ identisch. Zu jedem initialen Cluster wird ein entsprechender Blattknoten für das als Resultat dienende Dendrogramm konstruiert.

Die Clusterung wird auf Basis der Abstände der Clustermittelwerte $\boldsymbol{v}$ durchgeführt. Die Abstände werden in einer quadratischen Abstandsmatrix

$$I_A = \begin{bmatrix} \iota_{1,1} & & \iota_{n_E,1} \\ & \ddots & \\ \iota_{1,n_E} & & \iota_{n_E,n_E} \end{bmatrix}$$

gespeichert, deren Elemente

$$\iota_{i,j} = ||\psi(\boldsymbol{v}_{\mathfrak{m}_i}), \psi(\boldsymbol{v}_{\mathfrak{m}_j})||_{\gamma,\psi}$$

den Abstand jeweils zweier Cluster angeben. Das mit ψ gekennzeichnete Abstandsmaß

$$||\psi(\boldsymbol{v}_{\mathfrak{m}_i}), \psi(\boldsymbol{v}_{\mathfrak{m}_j})||_{\gamma,\psi} = ||\boldsymbol{v}_{\mathfrak{m}_i}, \boldsymbol{v}_{\mathfrak{m}_j}||_\gamma \tag{6.13}$$

berücksichtigt dabei die Kodierung der Clustermittelwerte. Eine Übersetzungstabelle $\mathfrak{m} = (\mathfrak{m}_1, \mathfrak{m}_2, \mathfrak{m}_3, \ldots, \mathfrak{m}_{n_E})$ dient der Entkoppelung der Indices der Abstandsmatrix von denen der Cluster. Die Tabelle ist initial mit den Werten $\mathfrak{m}_i = i$ belegt. Für eine Abstandsberechnung

$$||\boldsymbol{v}_i = (v_{i,1}, \ldots, v_{i,n_D})^\top, \boldsymbol{v}_j = (v_{j,1}, \ldots, v_{j,n_D})^\top||_\gamma = \left(\sum_{k=1}^{n_D} |v_{i,k} - v_{j,k}|^\gamma \right)^{\frac{1}{\gamma}} \tag{6.14}$$

muß ferner eine Norm γ festgelegt werden. Zur Clusterung von Merkmalen nach ihren Bildkoordinaten wird mit $\gamma = 2$ der euklidische Abstand gewählt. Andere Trainingsschritte erfordern jedoch auch kleinere Werte.

Die Abstandsmatrix wird genutzt, um möglichst ähnliche Cluster zu ermitteln. Da die Abstandsmatrix quadratisch mit der Anzahl zu clusternder Vektoren wächst, ist eine Optimierung über alle Matrizenelemente sehr aufwendig. Es wird daher eine um den Faktor n_U unterabgetastete Abstandsmatrix

$$I_U = \begin{bmatrix} \iota_{1,1} & & \iota_{n_E/n_U,1} \\ & \ddots & \\ \iota_{1,n_E} & & \iota_{n_E/n_U,n_E} \end{bmatrix}$$

eingeführt, deren Elemente höchstens so groß sind wie das Minimum an den entsprechenden Positionen der großen Abstandsmatrix I_A, d.h.

$$\iota_{i,j,I_U} \leq \min_{k=0\ldots n_U-1} \iota_{i+k,j,I_A}. \tag{6.15}$$

Nach der vollständigen Berechnung der Abstandsmatrizen I_A und I_U werden in einem sich wiederholenden Prozeß die $n_E - 1$ inneren Knoten des Dendrogramms berechnet.

Dazu wird zunächst die Zeile i_{min} und die Spalte j_{min} bestimmt, für die die Abstandsmatrix I_A das Minimum ι_{min} ergibt. Dies geschieht am einfachsten dadurch, daß das Minimum mit dem ersten Matrizenelement $\iota_{min} = \iota_{1,1,I_A}$ vorbelegt wird und dann die Abstandsmatrix I_A elementweise (von links nach rechts) durchlaufen wird, wobei ι_{min} gegebenenfalls angepaßt wird. Um die vollständige Suche über I_A abzukürzen, wird ι_{min} vor dem Vergleich mit einem bestimmten Element aus I_A zuerst mit dem entsprechenden Eintrag der unterabgetasteten Abstandsmatrix I_U verglichen. Falls der Eintrag dort größer ist, können die nächsten n_U Einträge in I_A übersprungen werden, was einen enormen Geschwindigkeitsgewinn ergibt. Falls der Eintrag in I_U jedoch kleiner ist, muß er anhand von I_A neu berechnet werden, da ein kleinerer Wert aufgrund der Ungleichung 6.15 unter dem tatsächlichen Minimum liegen kann. Das Minimum ι_{min} wird gegebenenfalls aktualisiert.

Die Cluster zu der Koordinate i_{min}, j_{min} des Minimums sind mit $\mathfrak{m}_{i_{min}}$ und $\mathfrak{m}_{j_{min}}$ indiziert. Diese werden nun zu einem neuen Cluster zusammengefaßt. Dieser enthält die Elemente beider Cluster

$$\beta_{neu} = \beta_{\mathfrak{m}_{i_{min}}} \cap \beta_{\mathfrak{m}_{j_{min}}}$$

und den möglicherweise kodierten Mittelwert aller Vektoren

$$\psi(\boldsymbol{v}_{neu}) = \psi\left(\frac{|\beta_{\mathfrak{m}_{i_{min}}}|}{|\beta_{neu}|}\boldsymbol{v}_{\mathfrak{m}_{i_{min}}} + \frac{|\beta_{\mathfrak{m}_{j_{min}}}|}{|\beta_{neu}|}\boldsymbol{v}_{\mathfrak{m}_{j_{min}}}\right),$$

wobei $|\beta|$ die Anzahl der Elemente in der Menge bezeichnet. Für den neuen Cluster wird ein neuer Knoten im Dendrogramm erzeugt.

Die Cluster der Minimum-Koordinate werden nun nicht mehr benötigt, da die weiteren Clusterungsschritte auf dem neuen Cluster arbeiten. Der neue Cluster wird daher unter dem Index $\mathfrak{m}_{i_{min}}$ gespeichert, sodaß die Abstandsmatrizen I_A und I_U nicht aufwendig vergrößert werden müssen. Um zu verhindern, daß die Abstandsmatrizen durch den nicht mehr benutzten Index j_{min} zerstückelt wird, wird die Spalte und Zeile j_{min} an den rechten Rand der Abstandsmatrizen sortiert. Dies geschieht durch Umkopieren der entsprechenden Matrizenelemente und die Anpassung der Werte in der unterabgetasteten Variante. Die Zuordnung zu den Clustern wird durch die Vertauschung der entsprechenden Einträge in der Übersetzungstabelle $\mathfrak{m}$ aufrechterhalten. Die letzte Zeile und Spalte von I_A wird fortan nicht mehr berücksichtigt. Dadurch verringert sich für jeden neuen Cluster die Größe des zu durchsuchenden Bereichs der Abstandsmatrix.

Für den neuen Cluster müssen nun die Einträge der Abstandsmatrizen aktualisiert werden. Da die genauen Abstandswerte für die Minimierung nicht vorliegen müssen, wird eine rechentechnisch weniger aufwendige Schätzung durchgeführt. Dazu wird nur die Abstandsungenauigkeit, die sich durch den Abstand des neuen Clusters von den ursprünglichen Clustern ergibt, von den Abstandswerten in der Zeile und der Spalte i_{min} subtrahiert. Die geschätzten Werte in I_A werden markiert, sodaß sie während der Abstandsminimierung bei Bedarf erkannt und neu berechnet werden können. Die Werte der unterabgetasteten Abstandsmatrix werden entsprechend angepaßt. Nun kann der nächste innere Knoten des Dendrogramms berechnet werden.

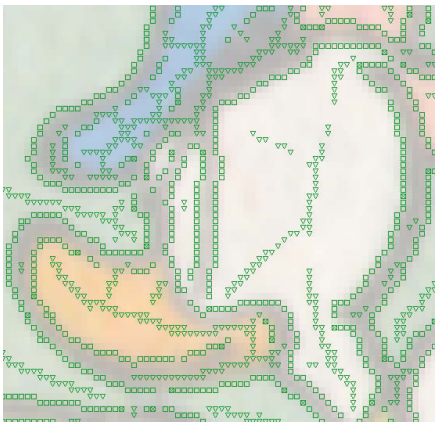
6.2.6 Bestimmung der optimalen Teilegröße

Das Clusterungsverfahren wird auf alle Merkmale jeweils eines Stichprobenbildes angewandt. Für jedes Stichprobenbild ergibt sich daher ein Dendrogramm, dessen Knoten jeweils Gruppen benachbarter Merkmale entsprechen. Ein Beispiel für ein solches Dendrogramm zeigt Abbildung 6.37. Zu jeder Gruppe von benachbarten Knoten kann nun ein Teileknoten für das Modell erzeugt werden. Jeder Knoten des Dendrogramms ist damit ein Kandidat, um in Form eines Teileknotens in das Modell aufgenommen zu werden. Dies geschieht dadurch, daß zunächst die Blattknoten des durch einen Dendrogrammknoten identifizierten Teilbaums ermittelt bestimmt werden. Diese geben die zu einer Gruppe zusammengefaßten, benachbarten Merkmale an. Um diese in das hierarchische Modell aufzunehmen zu können, wird ein neuer Teileknoten erzeugt. Die Merkmale der Gruppe werden dann als Unterknoten in den neuen Teileknoten eingefügt. Dabei stellen sich die folgenden Fragen:

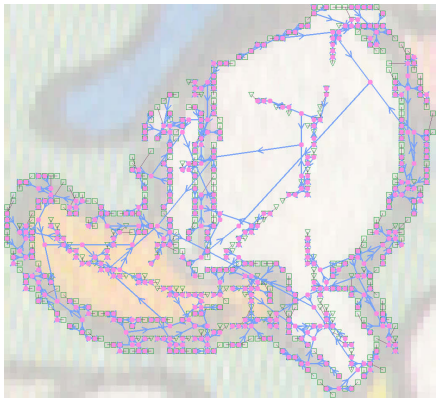
- Welche Teilekandidaten sollen in das Modell aufgenommen werden?
- Wie werden die neuen Teile bezüglich Ortstoleranz und Schwelle parametrisiert?
- Wie erzeugt man aus einzelnen Teileknoten ein vollständiges Modell?

Um die erste Frage zu klären, wird angenommen, daß sich verschiedene Teilekandidaten unterschiedlich gut für die Aufnahme in das Modell eignen. Da die

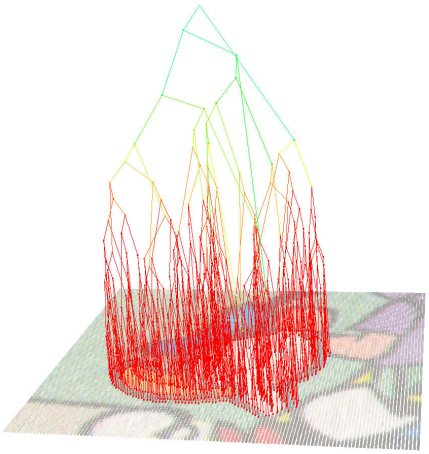
Originalbild, 117x113 Pixel



Extrahierte Merkmale



Dendrogramm, 2D-Darstellung



Dendrogramm, 3D-Darstellung

Abbildung 6.37: **Dendrogramm der räumlichen Benachbarung von Merkmalen eines Donald-Kopfes.** Oben: Vorbereitende Merkmalsextraktion. Kanten sind durch Rechtecke, Skelettlinien durch Dreiecke und Ecken durch Kreuze markiert. Unten links: 2D-Darstellung des Dendrogramms. Benachbarte Merkmale sind durch blaue Linien verbunden. Pfeile zeigen in Richtung höherer Knoten im Dendrogramm. Unten rechts: 3D-Darstellung des Dendrogramms mit der Höhe der Knoten im Dendrogramm als dritte Dimension.

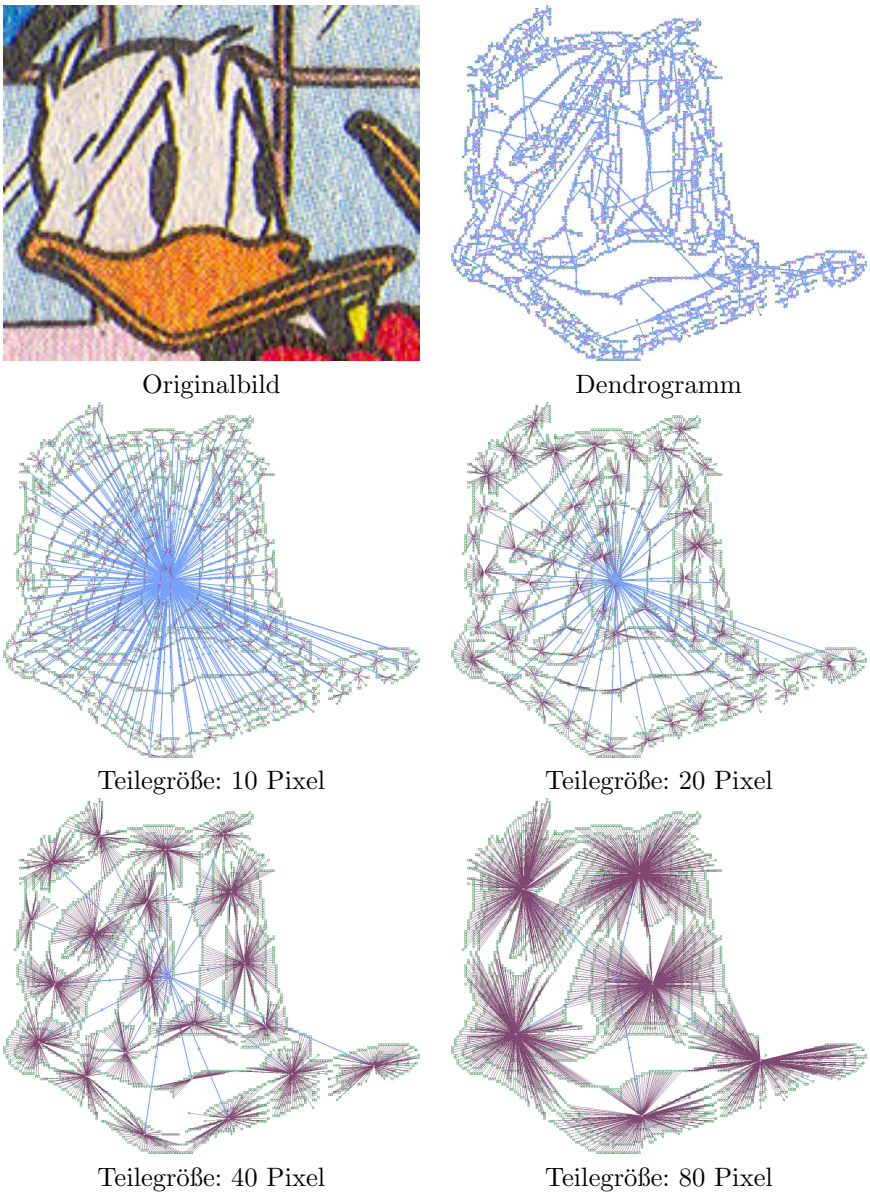


Abbildung 6.38: **Umwandlung von Dendrogrammen in provisorische Ansichtsmodelle mit unterschiedlichen Teilegrößen.** Für die Modellerzeugung werden mehrere Knoten des Dendrogramms ausgewählt. Die entsprechenden Merkmale an den Blattknoten der jeweiligen Teilbäume werden zu neuen Teileknoten zusammengefaßt (violette Verbindungslinien). Die Teileknoten werden einem gemeinsamen Knoten für das gesamte Objekt untergeordnet (hellblaue Verbindungen). Je nach Auswahl der Knoten im Dendrogramm können unterschiedlich große Teile erzeugt werden (untere 4 Bilder).

Teilekandidaten aufgrund der räumlichen Benachbarung der zugrunde liegenden Merkmale erzeugt wurden, werden Teilkandidaten bezüglich ihrer räumlichen Ausdehnung unterschieden. Diese ergibt sich aus dem Durchmesser der konvexen Hülle um die Merkmale in der Bildebene. Aus Gründen der Einfachheit wird dieser als Mittelwert der maximalen vertikalen und horizontalen Ausdehnung berechnet. Um den Einfluß dieser Variablen auf die Objekterkennung zu überprüfen, müssen verschieden parametrisierte Teileknoten getestet werden. Da die Versuche auf der Merkmalsebene des Modells ergaben, daß aufgrund der starken Abhängigkeit zwischen der Ortstoleranz und der Schwelle ein Parameter frei gewählt werden kann, wird die Schwelle für alle Teileknoten auf 90 Prozent festgelegt. Dahinter steht die Beobachtung, daß die Trefferbilder für hohe Schwellen deutlich zielgenauer aussehen. Für die Untersuchung der Teilekandidaten ergibt sich dadurch ein zweidimensionaler Parameterraum, der aus der Teilegröße und der Ortstoleranz besteht.

Die verschiedenen Teileparametrisierungen werden idealerweise bezüglich ihrer Güte zur Objekterkennung bewertet. Dazu müssen die Teile zu einem Modell für vollständige Objekte zusammengefaßt werden. Eine einfache Methode besteht darin, alle Knoten einer bestimmten Größe jeweils eines Dendrogramms auszuwählen und mit Hilfe eines neuen übergeordneten Modellknotens zusammenzufassen. Da sich hier verschiedene Teilegrößen und Ortstoleranzen wählen lassen, ergeben sich für jedes Dendrogramm mehrere Modelle. Diese eignen sich zur Erkennung der Objektansicht, die durch das dem Dendrogramm zugrunde liegende Stichprobenelement repräsentiert wird. Die Bewertung dieser verschieden parametrisierten Ansichtsmodelle geschieht über die Klassifikation der Stichprobe. Die optimale Teileparametrisierung zeigt sich in möglichst vielen erkannten positiven Stichprobenelementen und möglichst wenigen Treffern in Hintergrundbildern.

Obwohl diese Art der Erzeugung von Ansichtsmodellen sehr einfach ist und vor allem für die Untersuchung von Teileparametrisierungen entworfen wurde, bietet sie auch einen Ausblick auf die Erzeugung des endgültigen Objektmodells. Für jedes Stichprobenelement kann ein optimal parametrisiertes Ansichtsmodell erzeugt werden, das eine bestimmte Menge von Stichprobenelementen erkennt. Um diese zu bestimmen wird mit jedem Ansichtsmodell die komplette Stichprobe klassifiziert. Redundante Ansichtsmodelle können durch den Vergleich der Klassifikationsergebnisse ermittelt und verworfen werden, falls ein kompaktes Modell gewünscht wird. Die übrigen Elemente ergeben dann das endgültige Modell.

Gegen diese Art, ein alle Objektansichten überdeckendes Modell zu erzeugen, spricht jedoch, daß die Abstraktionsleistung einzelner Teile nicht ausgewertet wird. Es wird beispielsweise nicht geprüft, wie stark einzelne Teile über verschiedene Muster generalisieren, oder ob sich für Untermengen der Teile eines Ansichtsmodells noch bessere Ergebnisse erzielen lassen. Die hier erzeugten Ansichtsmodelle sind daher noch als provisorisch zu betrachten. Die Ergebnisse der Teileoptimierung können darüberhinaus noch weitere Strategien der Modellerzeugung aufzeigen.

Ortstoleranz [Pixel]	**Teilegröße [Pixel]**			
	10	**30**	**50**	**70**
5	2,55	2,27	2,27	2,27
15	*0,29*	4,97	4,91	4,91
25	*0,08*	5,91	1,89	20,42
35	*0,06*	*0,54*	0,78	0,50

Tabelle 6.5: Genauigkeit der Objekterkennung über Teilegröße und Ortstoleranz unter Berücksichtigung von Bag-of-features-Modellen. Die Genauigkeit wird hier als Summe der erkannten Stichprobenbilder durch die Anzahl an Trefferpositionen in Hintergrundbildern berechnet. Hohe Werte zeigen gute Modelle an. Bag-of-Features-Modelle sind kursiv hervorgehoben.

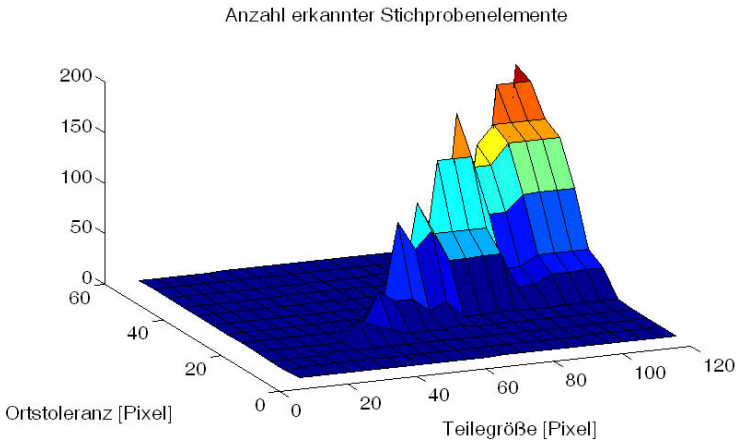


Abbildung 6.39: **Anzahl erkannter Positivbeispiele abhängig von Teilegröße und Ortstoleranz.** Die Ortstoleranz hat den Haupteinfluß auf die Zahl der erkannten Stichprobenelemente. Da Modelle mit einer Ortstoleranz über der Teilegröße nicht getestet wurden, sind die Werte auf der linken Seite des Diagramms auf Null gesetzt und daher nicht zu beachten.

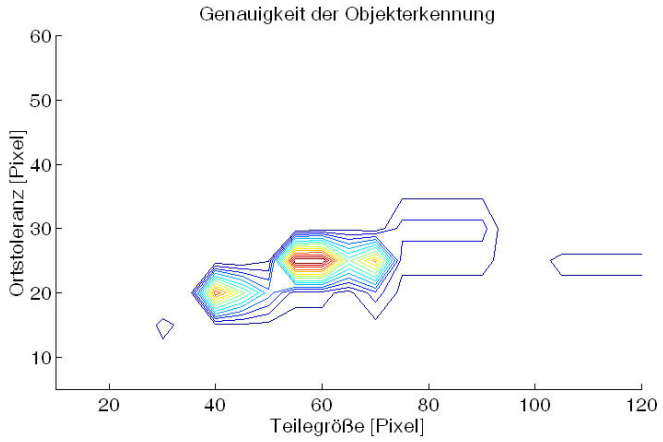


Abbildung 6.40: **Genauigkeit der Objekterkennung abhängig von Teilegröße und Ortstoleranz.** Die Höhenlinien geben die Anzahl der erkannten Stichprobenelemente dividiert durch die Anzahl falscher Pixeltreffer an. Für Teile mit 55–60 Pixeln Größe und 25 Pixeln Ortstoleranz ergibt sich ein Maximum.

Die provisorischen Ansichtsmodelle werden daher erst einmal nur eingesetzt, um den Parameterraum der Teileknoten zu erkunden. Vor der Konstruktion eines solchen Ansichtsmodells muß eine Parametrisierung bezüglich der Ortstoleranz und der Größe der Teile gewählt werden sowie ein Dendrogramm, das die gewünschte Ansicht repräsentiert. Die Erzeugung des Ansichtsmodells geschieht mit Hilfe eines rekursiven Verfahrens, das am Wurzelknoten des Dendrogramms startet. Für diesen sowie seine zwei Unterknoten wird die Ausdehnung der entsprechenden Merkmalsgruppe bestimmt. Falls die Ausdehnung für den gerade besuchten Knoten größer als die geforderte Teilegröße ist, wird geprüft, ob zu den Kindknoten abgestiegen werden kann. Wenn die Ausdehnung der Kindknoten ebenfalls über der gewünschten Teilegröße liegt, kann dies gefahrlos geschehen. Für die Kindknoten wird dann eine erneute Unterteilung überprüft. Wenn der aktuelle Knoten zwar eine zu hohe, die Kindknoten jedoch eine zu geringe Ausdehnung besitzen, kann nicht rekursiv abgestiegen werden. Hier besteht die Gefahr, sehr kleine Teilbäume abzuspalten, welche die Stabilität des Verfahrens beeinträchtigen. Da die geforderte Teilegröße ohnehin selten exakt mit der Ausdehnung der Merkmalsknoten übereinstimmt, wird die Teilegröße als Mindestgröße interpretiert. Wenn die Rekursion an einem Knoten mit ausreichend kleiner Ausdehnung terminiert, werden die Merkmale des Knotens durch einen neuen Teileknoten zusammengefaßt. Für diesen werden die Ortstoleranz und der Schwellwert wunschgemäß eingestellt. Da für den aktuellen Teilbaum keine weiteren Knoten erzeugt werden, liefert das Verfahren eine Menge von Teilen, deren Merkmale sich nicht überlappen. Die resultierende Teilemenge wird nun mit Hilfe eines weiteren neuen Knotens zusammengefaßt, welcher die komplette Ansicht repräsentiert. Abbildung 6.38 zeigt vier Beispiele für auf diese Art er-

zeugte provisorische Ansichtsmodelle. Die Merkmale sind hier grün dargestellt, Verbindungen zwischen Merkmalen und Teileknoten sind violett und die Verbindungen der Teileknoten zu den übergeordneten Ansichtsknoten sind blau markiert.

Zum Test von provisorischen Ansichtsmodellen wird eine Stichprobe aus 286 Donald-Köpfen und 7 Hintergrundbildern eingesetzt. Die Größe der Donaldköpfe liegt im Mittel bei 171×167 Pixel mit einer Standardabweichung von 73×80. Die Hintergrundbilder sind im Mittel 667×536 Pixel groß. Zur Bewertung eines Modells wird die Anzahl der erkannten positiven Stichprobenbilder bestimmt sowie die Anzahl der Pixeltreffer in den Hintergrundbildern. Die Genauigkeit der Objekterkennung wird als Anzahl der erkannten Positivbeispiele durch die Anzahl an falschen Pixeltreffern berechnet. Falls keine falschen Pixeltreffer auftreten, wird eine Minimalzahl von 1 angenommen, um eine Division durch Null zu vermeiden.

Als erstes werden verschieden parametrisierte Ansichtsmodelle zu dem in Abbildung 6.38 dargestellten Stichprobenelement erzeugt. Die Ausdehnung der Merkmale dieses Stichprobenelements liegt bei 158×135 Pixel, was verglichen mit der Standardabweichung nahe an der mittleren Größe aller Stichprobenelemente liegt. Das ausgewählte Dendrogramm kann daher als durchschnittlich betrachtet werden. Die Genauigkeit verschiedener Modelle gibt Tabelle 6.5 an. Die höchsten Werte scheinen in der Nähe der Diagonalen von links oben nach rechts unten zu liegen. Die Ergebnisse für Parametrisierungen mit großen Teilen und geringer Ortstoleranz leiden vor allem an der niedrigen Zahl erkannter Stichprobenelemente. Meistens wird hier nur das Stichprobenelement erkannt, aus dem das Modell erzeugt wurde. Für kleine Teile mit großer Ortstoleranz steigt dagegen die Zahl falscher Treffer. Dabei fällt auf, daß bei einigen Parametrisierungen die Teile kleiner sind als die Ortstoleranz. Diese sind in der Tabelle kursiv hervorgehoben. Bei diesen Parametrisierungen geht die Teilegeometrie nicht in die Klassifikation ein, sodaß man die Modelle in dieser Beziehung als Bag-of-Features bezeichnen kann. Da die Testbilder größer sein können als die Ausdehnung des Modells, führt eine über die Teilegröße steigende Ortstoleranz dazu, daß mehr Merkmale zu einer positiven Objekterkennung beitragen als das Modell enthalten kann. Die Genauigkeiten sind entsprechend niedrig. Dies ist auch in Übereinstimmung mit den bisherigen Ergebnissen, die einen positiven Beitrag der Geometrie anzeigen. Eine weitere Übereinstimmung ergibt sich bei den erkannten Stichprobenelementen. Die zweitbeste Parametrisierung mit einer Teilegröße von 30 und einer Ortstoleranz von 25 ergibt ein Modell, bei dem die Geometrie nur eine geringe Rolle spielt. Bei diesem Modell werden 71 Donaldköpfe erkannt, von denen etwa 62 Prozent wie der trainierte Donaldkopf nach rechts schauen. Bei dem Modell mit der höchsten Genauigkeit für Teile der Größe 70 und eine Ortstoleranz von 25 Pixeln geht die Geometrie stärker in die Objekterkennung ein. Von den erkannten 18 Donaldköpfen schauen hier 83 Prozent in die gleiche Richtung wie der trainierte Donaldkopf. Es werden also ähnlichere Stichprobenelemente erkannt.

Um ein vollständigeres Bild zu bekommen, werden nun weitere provisorische Ansichtsmodelle berechnet. Die Teilegröße wird dabei in Schritten von 5

Pixeln zwischen 10 und 120 Pixeln variiert. Ab einer Größe von 125 Pixeln enthält das Modell nur noch höchstens zwei Teileknoten. Da bei so wenigen Teilen keine sinnvolle Mehrheitsentscheidung mehr getroffen werden kann, verliert der Ansichtsknoten, der die Teileknoten zusammenfaßt, zunehmend an Bedeutung. Für die Ortstoleranz werden Werte von 5 bis 60 Pixeln gewählt, ebenfalls in Schritten von 5 Pixeln. Modelle, bei denen die Ortstoleranz größer ist als die Teilegröße, werden aus den genannten Gründen ausgelassen. Der überprüfte Parameterbereich deckt damit die meisten sinnvollen Parametrisierungen ab.

Die Ergebnisse sind in den Abbildungen 6.39 und 6.40 dargestellt. Wie an dem Diagramm in Abbildung 6.39 zu sehen ist, stellt die Ortstoleranz den Haupteinfluß dar, da mit steigender Ortstoleranz die Anzahl der Treffer stark steigt. Dies gilt sowohl für die in der Abbildung dargestellte Anzahl erkannter Positivbeispiele als auch für die Anzahl der Pixeltreffer in Hintergrundbildern. Mit steigender Teilegröße sinkt dabei die Anzahl der erkannten Teile leicht, wohingegen die Anzahl der in einem Bild erkannten Pixel etwas ansteigt. Die Modelle werden demnach mit steigender Teilegröße etwas trennschärfer, wobei sich für eine Größe von 60 Pixeln und eine Ortstoleranz von 25 Pixeln ein guter Kompromiß zwischen einer möglichst großen Zahl erkannter Elemente und einer möglichst hohen Unterdrückung von falschen Treffern ergibt. Dies ist in Abbildung 6.40 dargestellt. Dabei spielt möglicherweise auch eine Rolle, daß bei größeren Teilen eine höhere Ortstoleranz eingestellt werden kann, ohne daß die Geometrie vernachlässigt wird.

Nachdem nun eine optimale Teileparametrisierung bezüglich der Ortstoleranz und der Teilegröße ermittelt wurde, kann überprüft werden, ob sich diese eignet, um ein vollständig die Stichprobe abdeckendes Modell zu erzeugen. Dazu wird zu allen 286 Positivbeispielen der Stichprobe ein Ansichtsmodell mit den gerade berechneten optimalen Parametern erzeugt. Die resultierenden 286 Modelle werden anschließend eingesetzt, um die Stichprobe zu reklassifizieren.

Idealerweise werden von jedem Ansichtsmodell mehrere Stichprobenelemente einschließlich des trainierten erkannt. In dem Fall weiß man, daß eine gewisse Generalisierbarkeit vorliegt, die zu einer hohen Stichprobenabdeckung führt und kompakte Modelle erlaubt. Außerdem sollten alle Ansichtsmodelle zu etwa ähnlich vielen erkannten Stichprobenelementen führen. Die Treffer verschiedener Modelle sollten sich dabei möglichst wenig überschneiden. Wie Abbildung 6.41 zeigt, weichen die erzeugten provisorischen Ansichtsmodelle von diesem Ideal ab. Zum einen eignen sich die meisten Ansichtsmodelle nur zur Erkennung des trainierten Stichprobenelements. Die Modelle verallgemeinern offenbar schlecht. Zum anderen liefern verschiedene Modelle unterschiedliche viele Treffer, sofern mehrere Stichprobenelemente erkannt werden. Dies läßt darauf schließen, daß eine für alle Ansichtsmodelle gleich gewählte Parametrisierung zu stark vereinfacht ist. Dabei fällt auf, daß einige Modelle sehr viele Elemente erkennen, was eine geringe Trennschärfe anzeigt. Zum anderen überschneiden sich viele Treffer, d.h. einige Stichprobenelemente lassen sich durch viele Modelle darstellen, wohingegen bei anderen Stichprobenelementen kaum ein Modell Treffer anzeigt. Es gibt also leicht und schwer zu modellierende Stichprobenelemente.

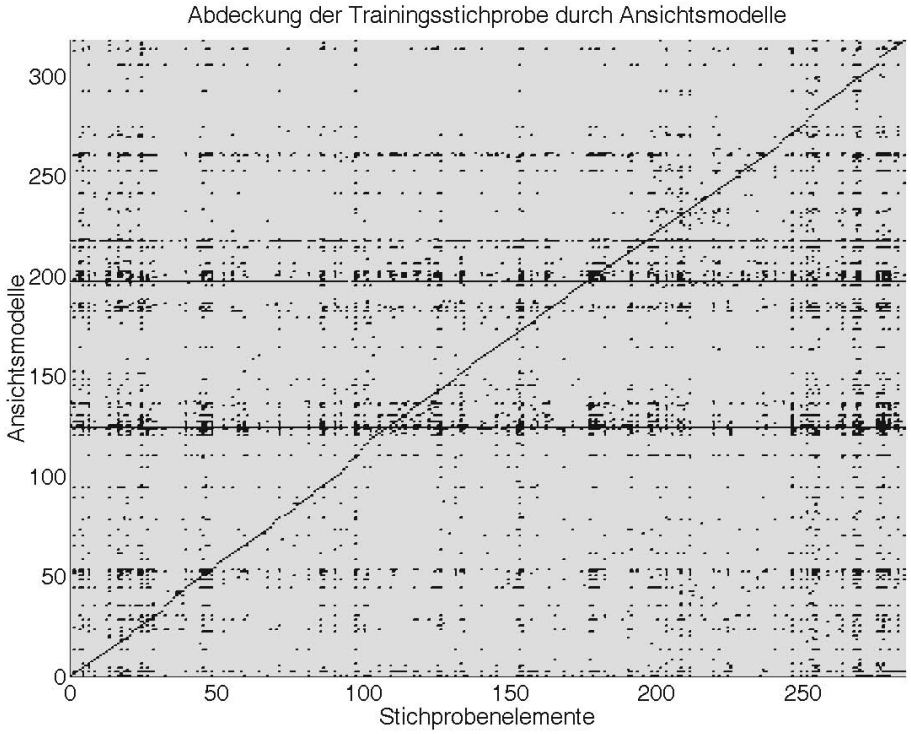


Abbildung 6.41: **Stichprobenabdeckung durch provisorische Ansichtsmodelle.** 286 Ansichtsmodelle wurden zur Erkennung der trainierten sowie 34 zusätzlichen untrainierten positven Stichprobenelemente eingesetzt. Schwarze Punkte zeigen an, daß ein Stichprobenelement durch ein Ansichtsmodell erkannt wurde.

Zusammenfassend läßt sich also sagen, daß eine bestimmte Parametrisierung der Teile bezüglich Ortstoleranz und Schwelle nicht für alle Ansichtsmodelle optimal ist. Eine mögliche Lösung besteht darin, die Ansichtsmodelle individuell für jedes Stichprobenelement zu optimieren. Dies ist aufgrund des hohen Rechenaufwandes vorerst unpraktikabel. Eine Hochrechnung der Rechenzeiten der beschriebenen Versuche auf die komplette Stichprobe ergibt eine Prozessor-Zeit von etwa zwei Jahren für das Training. Die Schätzung basiert auf einem AMD Athlon Xp mit 2GHz Takt und 2GB RAM als Versuchsrechner.

Daß das für ein Stichprobenelement gefundene Optimum im Parameterraum offenbar bei verschiedenen Stichprobenelementen zu unterschiedlichen Ergebnissen führt, deutet auf weitere wichtige Einflußfaktoren hin, die bisher vernachlässigt wurden. Ein solcher Einflußfaktor wird im folgenden identifiziert und näher untersucht.

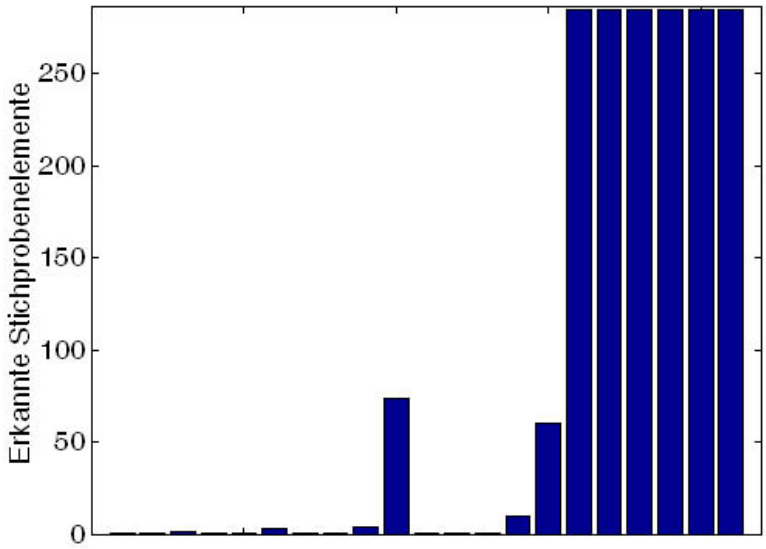
6.2.7 Zusammenhang zwischen der Objektgröße und der Stichprobenabdeckung

Um herauszufinden, warum eine gemeinsame Parametrisierung nicht für alle Stichprobenelemente gute Modelle ergibt, werden verschiedene Ergebnisse der letzten Messung genauer untersucht.

Bei den Ansichtsmodellen, die auffällig viele Stichprobenelemente erkennen, zeigt sich, daß diese besonders klein sind. Für die eingestellte Teilegröße von 60 Pixeln ergibt sich jeweils ein einziges Teil, welches das gesamte Objekt umfaßt. Um herauszufinden, wie sich die Stichprobenabdeckung über dem Parameterraum verhält, werden verschiedene Ansichtsmodelle für ein kleines Objekt berechnet. Die Ergebnisse zeigt Abbildung 6.6. Ab einer Teilegröße von 60 Pixeln erkennen die Ansichtsmodelle unabhängig von der Ortstoleranz fast die vollständige Stichprobe. Dies fällt mit der Reduktion der Teileanzahl auf ein einzelnes Element zusammen, die sich für eine Teilegröße von 60 Pixeln zeigt. Bei Teilegrößen von 50, 40 und 30 Pixeln ergeben sich noch 3, 4 und 5 Teile. Anscheinend werden bei dem einzelnen großen Teil fehlende Merkmale ausgemittelt, die bei kleineren Teilen dazu führen, daß eines nicht erkannt wird. Aufgrund der geringen Teileanzahl führt dies dazu, daß das gesamte Objekt nicht erkannt wird. Dies ist ein Spezialfall, der bei größeren Objekten aufgrund der höheren Teilezahl und der im Vergleich zur Objektgröße kleineren Teile nicht auftritt. Die Teilegröße ist daher bei kleineren Objekten kritischer als bei großen.

Abbildung 6.7 zeigt die Ergebnisse für ein Stichprobenelement, daß für die gemeinsame Parametrisierung eine mittelhohe Anzahl an Stichprobenelementen erkennt. Das modellierte Objekt liegt von der Größe eine Standardabweichung unter dem Durchschnitt, ist aber noch doppelt so groß wie das zuvor beschriebene Beispielobjekt (Abbildung 6.6). Die Anzahl der erkannten Stichprobenelemente hängt hier wieder hauptsächlich von der Ortstoleranz ab, wobei sich im Gegensatz zu dem vorigen Beispiel jede beliebige Anzahl an erkannten Elementen einstellen läßt.

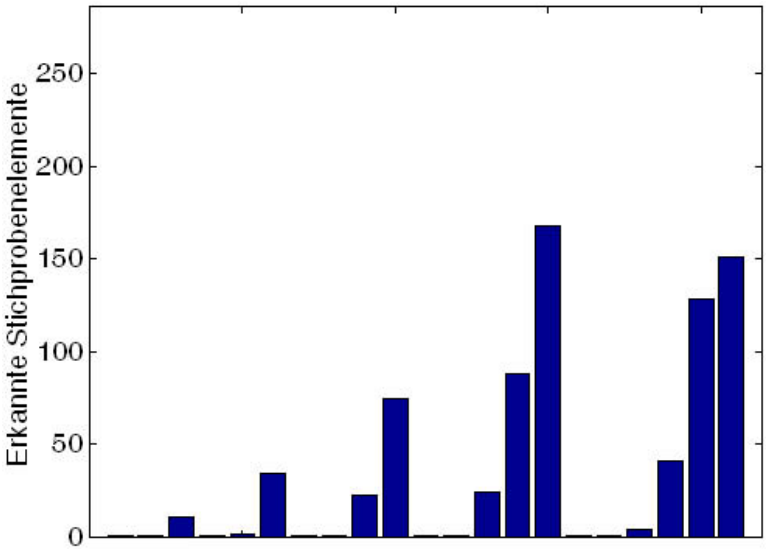
Eine Durchsicht der erkannten Stichprobenelementen ergibt jedoch, daß bei einer mittleren bis hohen Stichprobenabdeckung viele Objekte erkannt werden,



Erkannte Elemente	1	1	2	1	1	3	1	1	4	74	1	1	1	10	60	284	284	284	284	284	284
Teilegröße [Pixel]	10	20	20	30	30	30	40	40	40	40	50	50	50	50	50	60	60	60	60	60	60
Ortstoleranz [Pixel]	5	5	10	5	10	15	5	10	15	20	5	10	15	20	25	5	10	15	20	25	30

Trainiertes Element  Bildgröße 75 × 77 Pixel, Objektgröße: 52 × 47 Pixel

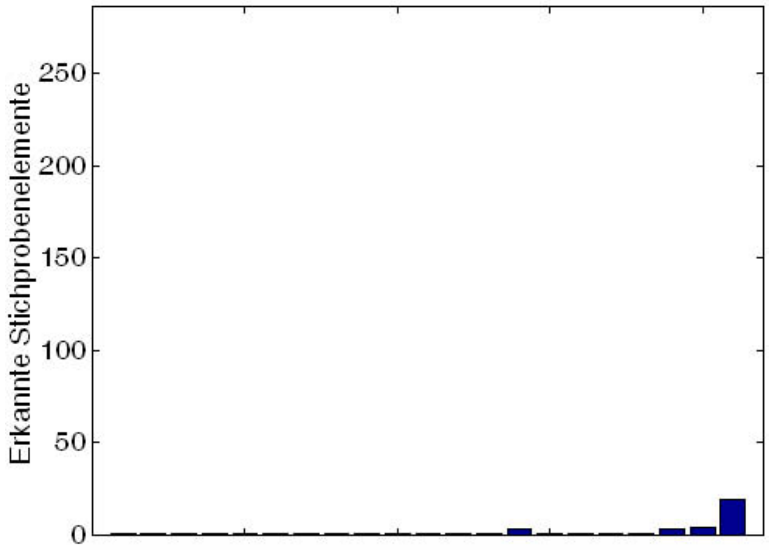
Tabelle 6.6: Stichprobenabdeckung über der Teileparametrisierung. Das Modell wird auf ein Stichprobenelement trainiert, für das sich mit der ermittelten optimalen Parametrisierung besonders viele Treffer ergaben (vgl. Abb. 6.41). Das Objekt ist mit 52 × 47 Pixeln besonders klein. Die Balkengraphik gibt für verschiedene Parametrisierungen von Ansichtsmodellen die Anzahl der erkannten Elemente einer Stichprobe mit 286 Donald-Bildern an. Bis auf zwei Parametrisierungen ergeben sich entweder sehr viele oder sehr wenige Treffer. Dies erschwert die Wahl guter Parameter.



Erkannte Elemente	1	1	11	1	2	34	1	1	23	75	1	1	24	88	168	1	1	4	41	128	151
Teilegröße [Pixel]	20	30	30	40	40	40	50	50	50	50	60	60	60	60	60	70	70	70	70	70	70
Ortstoleranz [Pixel]	10	10	15	10	15	20	10	15	20	25	10	15	20	25	30	10	15	20	25	30	35

Trainiertes Element  Bildgröße: 120 × 113 Pixel, Objektgröße: 97 × 102 Pixel

Tabelle 6.7: Stichprobenabdeckung über der Teileparametrisierung. Das Modell wird auf ein Stichprobenelement trainiert, für das sich mit der ermittelten optimalen Parametrisierung eine mittlere Anzahl Treffer ergab (vgl. Abb. 6.41). Das Objekt ist mit 97×102 Pixeln etwa eine Standardabweichung kleiner als die mittlere Objektgröße in der Stichprobe. Die Balkengraphik gibt für verschiedene Parametrisierungen von Ansichtsmodellen die Anzahl der erkannten Elemente einer Stichprobe mit 286 Donald-Bildern an. Die verschiedenen Teileparametrisierungen ergeben fein abgestufte Trefferzahlen. Der Haupteinfluß ist die Ortstoleranz.



Erkannte Elemente	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	3	4	19
Teilegröße [Pixel]	10	20	20	30	30	30	50	50	50	60	60	60	60	80	80	80	80	80	80	80
Ortstoleranz [Pixel]	5	5	10	5	10	15	5	15	25	5	10	20	25	30	5	10	20	25	30	35	40

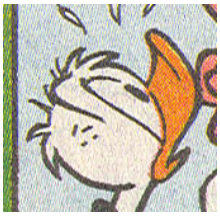
Trainiertes Element  Bildgröße: 169 × 164 Pixel, Objektgröße: 142 × 130 Pixel

Tabelle 6.8: Stichprobenabdeckung über der Teileparametrisierung. Das Modell wird auf ein Stichprobenelement trainiert, für das mit der ermittelten optimalen Parametrisierung nur das trainierte Stichprobenelement selbst erkannt wurde (vgl. Abb. 6.41). Das Objekt hat eine mittlere Größe. Nur für eine sehr große Ortstoleranz ergeben sich Treffer. Insgesamt werden nur sehr wenige Stichprobenelemente erkannt. Dies liegt vermutlich an der seltenen Objektpose mit einem nach oben gerichteten Gesicht.

die dem trainierten Element nicht ähnlich sehen. Bei einer Teilegröße von 40 Pixeln und einer Ortstoleranz von 20 Pixeln ergeben sich beispielsweise 34 Treffer, was einer Stichprobenabdeckung von ungefähr 12 Prozent entspricht. Die erkannten Stichprobenelemente haben alle ungefähr die gleiche Größe. Das mit Abstand größte Objekt hat eine Breite von etwa 200 Pixeln. Subjektiv betrachtet, stellen sie auch eher moderat deformierte Donald-Köpfe mit neutralen Gesichtsausdrücken dar. Damit ähneln sie bis zu einem gewissen Grad dem trainierten Stichprobenelement. Objektiv sind jedoch nur 22 Donaldköpfe wie das Trainingselement nach rechts ausgerichtet, 11 dagegen nach links.

Eine Parametrisierung mit einer Teilegröße von 70 Pixeln und einer Ortstoleranz von 35 Pixeln ergibt eine besonders hohe Stichprobenabdeckung von etwa 53 Prozent. Die erkannten Objekte unterscheiden sich bei dieser Parametrisierung deutlicher. Die Größen reichen häufiger bis etwa 250 Pixel. Von den erkannten 151 Donaldköpfen sind 63 nach links und 77 wie das Trainingselement nach rechts ausgerichtet. Die geometrische Übereinstimmung ist damit noch einmal deutlich geringer. Die erkannten Objekte sind trotz der geringen Ähnlichkeit allerdings noch nicht vollkommen beliebig. Dies zeigt sich daran, daß die nicht erkannten Objekte auch größte Durchmesser bis 400 Pixel aufweisen. Bei den nicht erkannten Objekten tritt zudem die umgekehrte Ausrichtung des Kopfes etwa um die Hälfte häufiger auf als die mit dem Trainingselement übereinstimmende.

Die Ergebnisse für ein überselektives Modell, das für die gewählte einheitliche Parametrisierung ausschließlich das trainierte Objekt erkennt, zeigt Abbildung 6.8. Da die Objektgröße nahe am Mittelwert der Stichprobe liegt, macht sich hier offenbar ein anderer Einfluß bemerkbar. Ein visueller Vergleich des trainierten Objekts mit den übrigen Stichprobenelementen legt nahe, daß hier das Aussehen des Objekts selbst für die geringe Generalisierbarkeit verantwortlich ist. Nur etwa 4 Prozent der Stichprobenbilder zeigen eine ähnliche Pose, in der Hinsicht, daß der Kopf nach rechts ausgerichtet ist und die Schnabelspitze höher als der Kopf reicht. Selbst in diesen seltenen Fällen weichen die Stichprobenelemente oft dadurch ab, daß die hohe Schnabelspitze nicht durch eine Drehung des Kopfes, sondern einen weit aufgerissenen Schnabel mit tief heruntergezogenem Unterkiefer hervorgerufen wird. Die geringe Stichprobenabdeckung rührt also daher, daß eine seltene Objektansicht modelliert wurde. Die Ergebnisse zeigen eine Verallgemeinerung des trainierten Musters auf weitere Objekte daher nur für Parametrisierungen mit hoher Ortstoleranz.

Als Zwischenfazit können daher folgende Ergebnisse festgehalten werden:

- Die Größe der trainierten Objekte spielt eine Rolle.

- Je höher die Stichprobenabdeckung ist, desto stärker weichen die erkannten Objekte von dem trainierten Objekt ab. Das vorgeschlagene Objekterkennungssystem verhält sich in dieser Hinsicht vorhersehbar und plausibel.

- Schon bei einer mäßigen Stichprobenabdeckung von 12 Prozent werden subjektiv und objektiv unterschiedliche Objekte erkannt.

- Die Repräsentativität der trainierten Objekte für die Gesamtmenge hat einen Einfluß auf die Stichprobenabdeckung. Statistische Ausreißer werden am besten individuell behandelt.

Da die Objektgröße das konkreteste Merkmal ist, wird dieser Einfluß als nächstes untersucht. Dazu wird die in Abbildung 6.41 gezeigte Abdeckungsmatrix für die Ansichtsmodelle mit gemeinsamer Parametrisierung spalten- und zeilenweise nach der Objektgröße sortiert. Das Ergebnis zeigt Abbildung 6.42. Die Zeilen sind jetzt bezüglich der Größe der trainierten Objekte und damit auch nach der Ausdehnung der Ansichtsmodelle sortiert. Hier zeigt sich eine deutliche Abhängigkeit. Für kleinere Modelle werden offenbar mehr Stichprobenelemente erkannt als für größere. Für die Spalten des Diagramms, die jetzt nach der Größe der zu erkennenden Stichprobenelemente sortiert sind, ist keine Abhängigkeit erkennbar. Um die aufgetretene Abhängigkeit deutlicher zu machen, wird nun für jedes Ansichtsmodell die Anzahl der erkannten Stichprobenelemente aufsummiert. Die in Abbildung 6.42 dargestellte Rangordnung der Elemente wird nun auf eine (einigermaßen) lineare Achse für die Größe des jeweils trainierten Objekts aufgespreizt. Fehlende Größenbereiche werden dabei durch die Ergebnisse für das nächstkleinere Objekt ergänzt. Abbildung 6.43 zeigt das Ergebnis. Dieses bestätigt den durch die Sortierung der Abdeckungsmatrix gewonnenen Eindruck. Ab einer Größe von 168 Pixeln werden nur noch einstellige Mengen an Stichprobenelementen erkannt. Klar unterscheidbare Größenbereiche, die auf Besonderheiten des Klassifikators oder des Modells schließen lassen, sind nicht erkennbar. Ansichtsmodelle von großen Objekten sind demnach wohl vor allem deswegen selektiver, da sie mehr Teile und mehr Merkmale enthalten.

Als nächstes wird überprüft, ob die Vergrößerung des durch die Teilegröße und die Ortstoleranz gebildeten Parameterraums um die Objektgröße als dritte Dimension genügt, um brauchbare Modelle zu erzeugen.

Dazu werden systematisch verschieden parametrisierte Ansichtsmodelle zu unterschiedlich großen Objekten berechnet. Für jede Parametrisierung wird die Stichprobenabdeckung berechnet. Dabei kommt die bereits beschriebene Stichprobe aus 286 Positivbeispielen und 7 Hintergrundbildern zum Einsatz. Die Berechnung der Stichprobenabdeckung geschieht mit Modellen zu 51 verschiedenen Objekten für die Vordergrundbilder und mit Modellen von 62 verschiedenen Objekten für die Hintergrundbeispiele. Für jedes trainierte Objekt werden jeweils Ansichtsmodelle für Teilegrößen von 10 bis 80 Pixeln bei einer Schrittweite von 10 Pixeln und für Ortstoleranzen von 5 bis 40 Pixeln mit einer Schrittweite von 5 Pixeln berechnet. Die dritte Dimension des Parameterraums ergibt sich aus der Größe der trainierten Objekte. Parametrisierungen mit einer Ortstoleranz über der Teilegröße werden weggelassen. Für jedes trainierte Objekt und jede Parametrisierung wird die Anzahl der positiven bzw. der negativen Beispiele der Stichprobe ermittelt. Anhand dieser Berechnungen wird nun entschieden, welche Parametrisierungen gut sind und welche schlecht sind.

Dazu wird für jede Kombination aus einer Teilegröße und einer Objektgröße die minimale Ortstoleranz ermittelt, ab der sich mindestens zwei Treffer in den Positivbeispielen der Stichprobe ergeben. Ein Treffer in der Stichprobe bedeutet

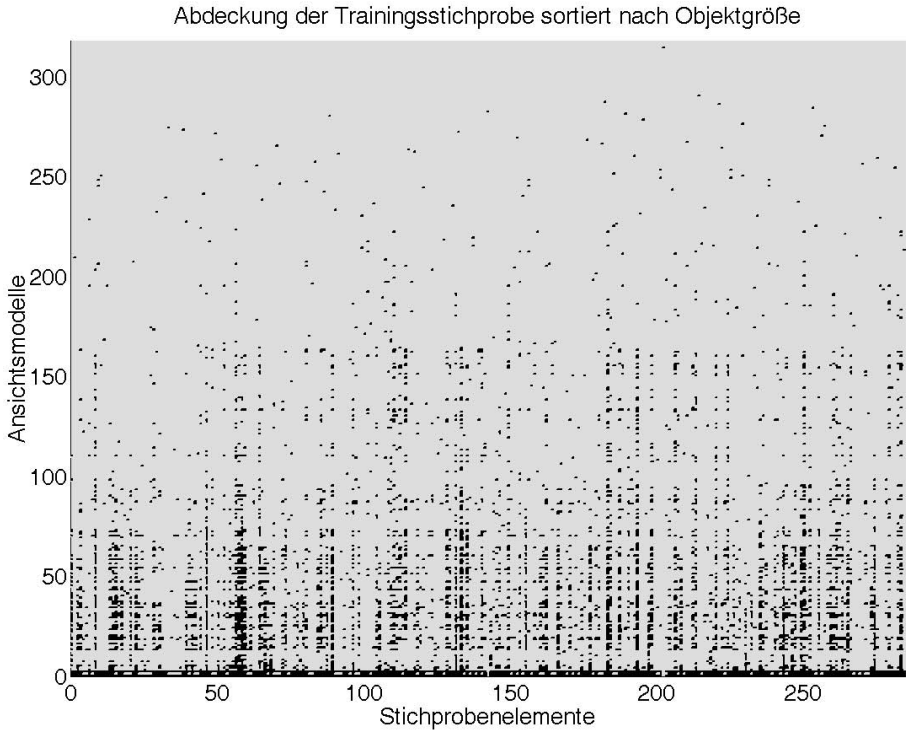


Abbildung 6.42: **Stichprobenabdeckung abhängig von der Objektgröße.** Das Diagramm stellt die bereits dargestellten Klassifikationsergebnisse für die gemeinsame Parametrisierung mit einer Teilegröße von 60 Pixeln und einer Ortstoleranz von 25 Pixeln für eine andere Reihenfolge von Stichprobenelementen dar. Dazu wurde die in Abbildung 6.41 gezeigte Matrix spaltenweise nach der Größe der zu erkennenden und zeilenweise nach der Größe der trainierten Objekte sortiert. Der Durchmesser der Objekte liegt zwischen 49 und 447 Pixeln. Die Objektgrößen steigen sowohl von unten nach oben als auch von links nach rechts an. Die Stichprobenabdeckung hängt offenbar von der Größe der Trainingsobjekte ab, nicht aber von den zu erkennenden Objekten.

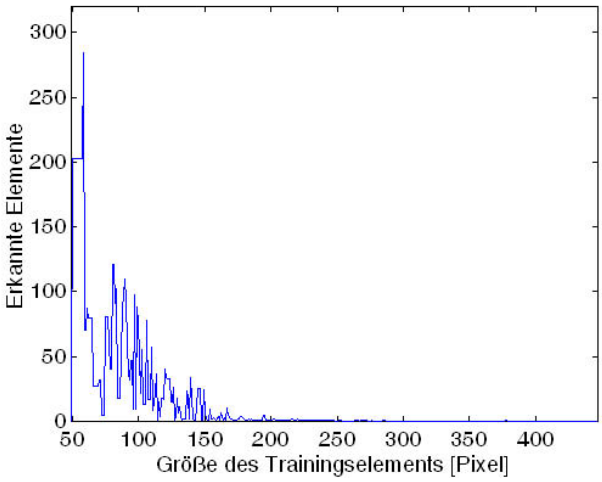


Abbildung 6.43: **Stichprobenabdeckung abhängig von der Größe der trainierten Objekte.** Ansichtsmodelle zu kleinen Objekten erkennen mehr Elemente der Stichprobe als Modelle zu großen Objekten.

in der Regel, daß genau das trainierte Objekt erkannt wird. Zwei Treffer markieren dagegen die Grenze, ab der das Modell nicht mehr nur selektiv auf das trainierte Muster reagiert, sondern beginnt, über mehrere Stichprobenelemente zu generalisieren. Höhere Ortstoleranzen bewirken, wie bereits festgestellt, eine weitere Erhöhung der Trefferzahlen. Ab einer gewissen Höhe der Ortstoleranz werden nicht mehr nur das trainierte und ähnliche Elemente erkannt, sondern auch Hintergrundbilder. Die ideale Modellparametrisierung liegt zwischen dem Beginn einer Generalisierung und dem Beginn der unspezifischen Reaktion auf Hintergrundmustern. Die minimale Ortstoleranz, ab der mindestens ein Hintergrundbild erkannt wird, wird daher ebenfalls bestimmt.

Es zeigt sich, daß die minimal generalisierenden Parametrisierungen und die beginnend unspezifischen Parametrisierungen zwei weitgehend getrennte Gruppen in dem dreidimensionalen Parameterraum bilden. Um nun die besten Parametrisierungen zu finden, wird per linearer Diskriminanzanalyse eine optimale Trennebene zwischen die beiden Gruppen eingepaßt. Die ermittelte Ebenengleichung lautet

$$-0{,}1202 \cdot \varsigma + 0{,}0117 \cdot r_t + 0{,}0113 \cdot r_a + 1{,}0306 = 0, \tag{6.16}$$

wobei r_t die Teilegröße und r_a die Größe einer Ansicht bzw. eines Beispielobjekts bezeichnet. Die Parametergruppen und die Übereinstimmung mit der Trennebene zeigt Abbildung 6.44. Minimal generalisierende Parametrisierungen sind als rote Rauten dargestellt, unspezifische Parametrisierungen dagegen als blaue Rauten. Kleiner und heller dargestellt sind Parametrisierungen, die durch

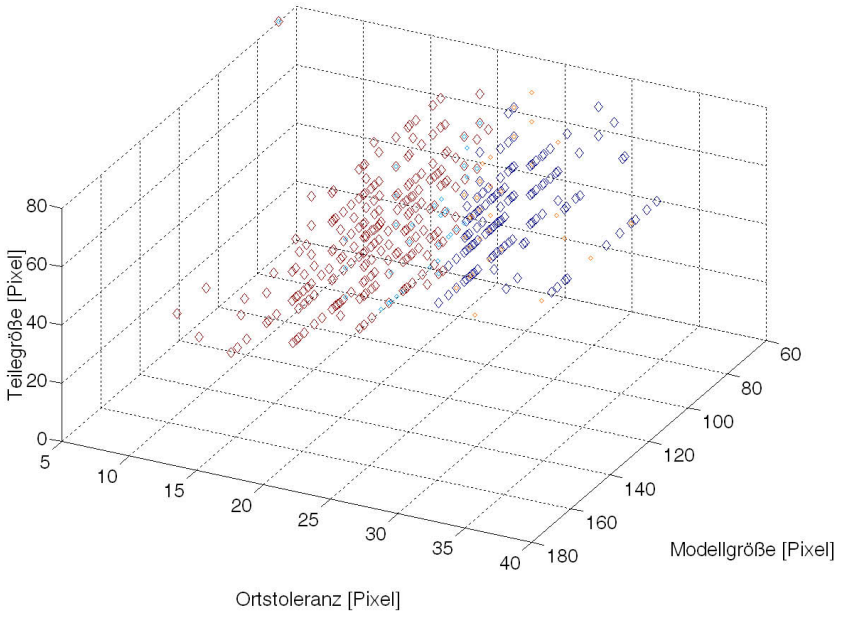


Abbildung 6.44: **Parameterraum aus Ortstoleranz, Teilegröße und Objektgröße.** Rote Rauten kennzeichnen Parametersätze, die nur korrekte Treffer liefern. Blaue Rauten liefern dagegen auch auf Hintergrundbildern Treffer. Zur Trennung der beiden Gruppen von Parametern wird ein linearer Klassifikator eingesetzt. Korrekt eingeordnete Parametrisierungen sind als große Rauten, falsche Einordnungen dagegen als kleine Rauten dargestellt.

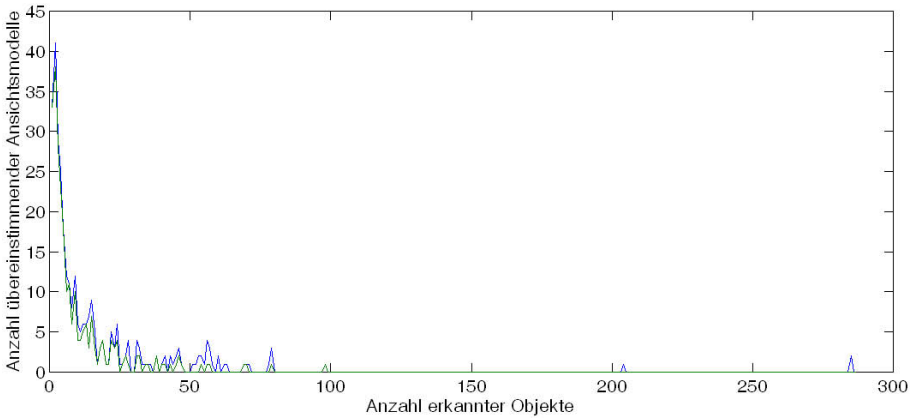


Abbildung 6.45: **Verteilung übereinstimmender Ansichtsmodelle über der Anzahl erkannter Objekte.** Die meisten Objekte werden nur von wenigen Modellen erkannt.

die Trennebene der falschen Gruppe zugeordnet werden. Etwa 84 Prozent der Parametrisierungen werden durch die Trennebene korrekt zugeordnet.

Da die ermittelte Trennebene einen guten Kompromiß zwischen Generalisierbarkeit und Fehlerunterdrückung darstellt, werden nun Ansichtsklassifikatoren für Parametrisierungen auf der Trennebene berechnet. Das resultierende Modell wird anhand der Reklassifikation der Trainingsstichprobe überprüft.

Dazu wird zu 286 Donaldköpfen jeweils ein Ansichtsmodell erzeugt, dessen Parametrisierung anhand der Objektgröße und einer Teilegröße von 60 Pixeln anhand von Gleichung 6.16 berechnet wird. Die gewählte Teilegröße entspricht dem Optimum aus der Ansichtsmodellierung in einem zweidimensionalen Parameterraum (s. Abb. 6.40). Da die Berechnungen ziemlich aufwendig sind, wird auf die Überprüfung weiterer Teilegrößen verzichtet. Die erzeugten Ansichtsmodelle werden nun zur Reklassifikation der 286 trainierten und etwa 30 weiterer Donaldköpfe eingesetzt.

Wie sich herausstellt, erkennen 64 von 286 Ansichtsmodelle ebenfalls die Hintergrundbilder. Werden diese ausgelassen, decken die übrigen Ansichtsmodelle immer noch 95 Prozent der Stichprobe ab. Die von verschiedenen Ansichtsmodellen erkannten Bilder überschneiden sich teilweise. Läßt man Ansichtsmodelle weg, die nur Stichprobenelemente erkennen, die bereits von anderen Ansichtsmodellen abgedeckt werden, ergibt sich ein kleineres Modell aus maximal 77 Ansichtsmodellen. Diese Reduktion deutet auf eine hohe Generalisierbarkeit des Modells hin. Tatsächlich abstrahieren jedoch nur 25 Ansichtsmodelle über mehrere Stichprobenelemente. Zwei Drittel der Ansichtsmodelle erkennen dagegen nur das jeweils eine trainierte Stichprobenelement. Diese Ansichtsmodelle werden in das Gesamtmodell aufgenommen, da die jeweils erkannten Stichprobenelemente von keinem anderen Ansichtsmodell abgedeckt werden. Die geringe

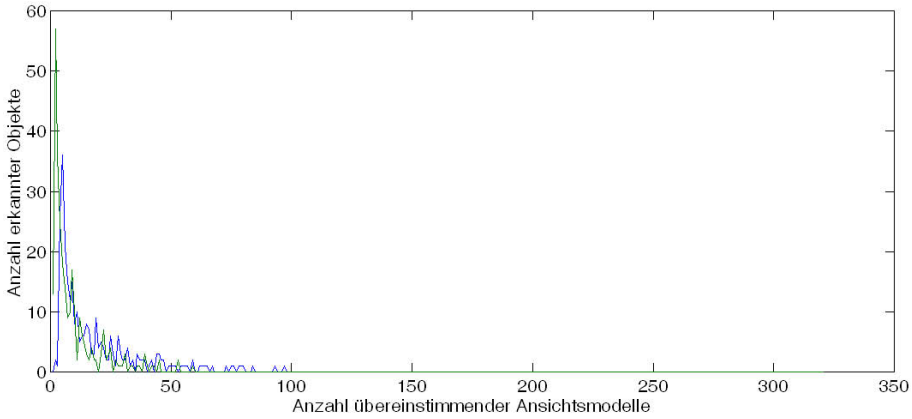


Abbildung 6.46: **Verteilung erkannter Stichprobenelemente über der Anzahl übereinstimmender Ansichtsmodelle.** Blau: Alle Ansichtsmodelle, grün: nur Ansichtsmodelle ohne Hintergrundtreffer. Die meisten Ansichtsmodelle erkennen nur wenige Objekte.

Abstraktionsleistung der Ansichtsmodelle verdeutlichen auch die Histogramme in den Abbildungen 6.45 und 6.46.

Abbildung 6.45 zeigt, durch wieviele Ansichtsmodelle ein bestimmtes Stichprobenelement erkannt wird. Das Diagramm gibt also an, wie stark sich die Ansichtsmodelle in ihren Ergebnissen bei der Objekterkennung überschneiden. Die blaue Linie bezieht sich wieder auf alle Ansichtsmodelle, die grüne wieder nur auf Ansichtsmodelle ohne falsche Treffer. Hier zeigt sich, daß die Hälfte der Stichprobenelemente von weniger als elf Ansichtsmodellen erkannt wird. Für eine Erkennung durch 4 Ansichtsmodellen ergibt sich die maximale Anzahl an Stichprobenelementen. Das bedeutet, daß für die meisten Stichprobenelemente nur eine geringe Überlappung zwischen verschiedenen Ansichtsmodellen auftritt. Allerdings existiert eine kleine Menge von Objekten, die durch sehr viele Modelle erkannt wird.

Die Abbildung 6.46 gibt an, wie die Anzahl der durch ein Ansichtsmodell erkannten Stichprobenelemente verteilt ist. Die blaue Linie bezieht sich dabei auf alle Ansichtsmodelle, die grüne dagegen nur auf Ansichtsmodelle ohne falsche Treffer in Hintergrundbildern. Die Hälfte der Ansichtsmodelle erkennt höchstens 6 Stichprobenelemente. Bis auf eine kleine Gruppe von stark verallgemeinernden Ansichtsmodellen erkennen die meisten Ansichtsmodelle daher nur wenige Objekte.

Für eine zuverlässige Objekterkennung sind die Ansichtsmodelle daher noch zu selektiv. Eine stärkere Generalisierbarkeit ergibt sich durch das Verschieben der Trennebene in Richtung einer höheren Ortstoleranz. Messungen für eine um 5 und um 10 Pixel erhöhte Ortstoleranz zeigen jedoch eine starke Zunahme von falschen Hintergrundtreffern. Eine Ursache dafür kann eine schlechte Lage oder Form der Trennfläche sein. Dies trifft insbesondere für sehr kleine Teile zu. Für

eine Teilegröße von 5 Pixeln ergibt sich beispielsweise bei einem kleinen Objekt von 93 Pixeln, was der mittleren Objektgröße minus der Standardabweichung entspricht, eine Ortstoleranz von etwa 18 Pixeln. Für ein großes Objekt mit einem Durchmesser von 245 Pixeln, was dem Mittelwert plus der Standardabweichung entspricht, ergibt sich sogar eine Ortstoleranz von 62 Pixeln. Da in beiden Fällen die Ortstoleranz deutlich größer ist als das Teil, spielt das Aussehen der Teile kaum noch eine Rolle. Die Ergebnisse der Gleichung 6.16 sind daher für kleine Teile nicht plausibel. In dem hier durchgeführten Versuch war die Teilegröße jedoch auf 60 Pixel eingestellt. In diesem Größenbereich liefert die Gleichung einleuchtende Werte. Die im dreidimensionalen Parameterraum markierten Parametrisierungen (siehe Abbildung 6.44) deuten auch keine bestimmte Form einer besseren Trenngrenze an. Die Resultate des untersuchten Ansatzes lassen sich daher wie folgt zusammenfassen:

- Gute Ansichtsmodelle lassen sich nicht in einem dreidimensionalen Parameterraum aus Ortstoleranz, Teile- und Objektgröße erzeugen.
- Das Modell ist zu selektiv. Allein durch eine Vergrößerung der Ortstoleranz ist das Problem jedoch nicht zu lösen.

Daher wird als nächstes die bereits als entscheidender Einflußfaktor identifizierte Repräsentativität der trainierten Objekte für die Stichprobe in die Modellerzeugung mit einbezogen.

6.3 Erzeugung eines visuellen Alphabets

Es wird nun eine Teiledarstellung in Form eines visuellen Alphabets eingeführt, welche die in den vorangegangenen Unterkapiteln identifizierten Einflußgrößen berücksichtigt. Das visuelle Alphabet speichert ein Repertoire an Teilen, die für bestimmte Objektansichten typisch sind. Zu spezielle oder untypische Teile können weggelassen werden, um ein abstrakteres und damit stärker generalisierendes Modell zu erzeugen. Damit kann auch die unterschiedliche Repräsentativität der modellierten Trainingsobjekte für die Stichprobe angemessen berücksichtigt werden.

Um Effekte der Stichprobe zu vermeiden, wird diese auf 800 Bilder von Donaldköpfen und 800 Hintergrundbilder vergrößert. Aus Gründen der Vergleichbarkeit werden Hintergrundbilder in der gleichen Größe wie die Vordergrundbilder verwendet. Die Trefferzahlen lassen sich so gemeinsam in der Anzahl erkannter Bilder, d.h. der Anzahl von Bildern mit mindestens einem Treffer, angeben.

Um die Repräsentativität eines Teils zu bestimmen, muß untersucht werden, in wievielen Beispielbildern zu einer bestimmten Objektpose das Teil enthalten ist. Da die Dendrogramme der Vordergrundbilder insgesamt etwa 3 Millionen Knoten und damit mögliche Teile enthalten, ist eine vollständige Auswertung derzeit rechentechnisch nicht durchführbar. Bevor die Repräsentativität der Teile untersucht werden kann, muß daher eine behandelbare Untermenge an Teilen

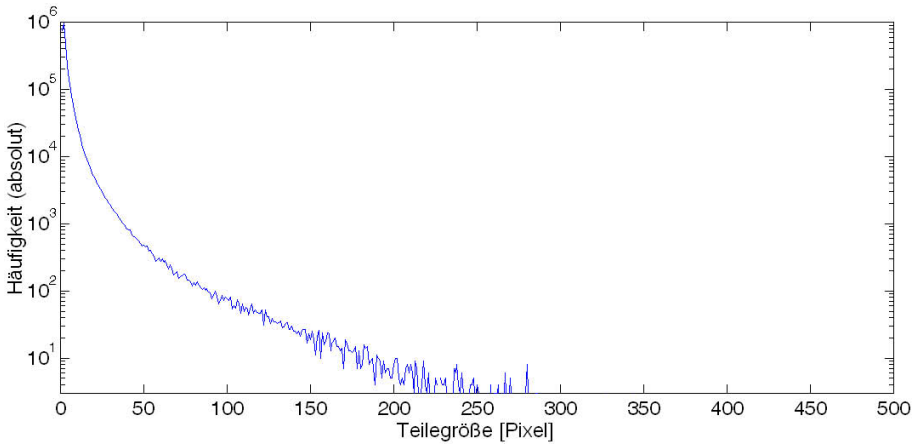


Abbildung 6.47: **Histogramm über die Größe aller Teilekandidaten.** Kleine Teile sind überproportional stark vertreten. Ca. 93% der Teile sind kleiner als 10 Pixel.

ermittelt werden. Da in nachfolgenden Schritten Bilder nur noch mit Hilfe dieser Teile dargestellt werden, sollten diese möglichst vielfältig sein, d.h. insbesondere keine Wiederholungen zeigen. Da die Teilemenge zur Darstellung aller Bilder verwendet wird, sind einzelne Teile allgemeingültig, d.h. sie beschreiben keine bestimmten Bilder mehr. Damit stellen die ausgewählten Teile eine Parallele zu dem visuellen Alphabet dar, das Tanaka [Tan96] im visuellen Kortex beschreibt. Die Bestimmung des Teilealphabets geschieht über die Clusterung der Teilekandidaten aus den Dendrogrammen.

Ein geeignetes hierarchisches Clusterungsverfahren wird im nächsten Abschnitt vorgeschlagen. Der Vergleich von Teilekandidaten über ihre gegenseitige Darstellbarkeit und Klassifizierbarkeit stellt sich als visuell plausibles und mathematisch stabiles Ähnlichkeitsmaß heraus. Für die Einstellung der Teileparameter werden lineare Regeln gefunden. Es zeigt sich, daß an der Clusterstruktur verschiedene Arten visueller Ähnlichkeit zwischen Teilen erkennbar sind. Zudem sind mehrere Unterteilungen in Cluster möglich. Das Training des Teilealphabets schließt daher mit einem Algorithmus, der eine konkrete Menge besonders homogener Cluster ermittelt.

Die Auswahl von typischen Teilen geschieht bei der Erzeugung der nächsthöheren Ebene des gesamten Objektmodells. Hier werden die für bestimmte Objektansichten typischen Teile ermittelt und zu jeweils neuen Modellknoten zusammengefaßt.

6.3.1 Clusterungsverfahren

Um Teilekandidaten zu clustern, müssen mehrere Fragen beantwortet werden:

- Nach welchem Ähnlichkeitsmaß werden Teile geclustert?

- Auf welchen Teilekandidaten wird die Clusterung durchgeführt?
- Welche Ergebnisse sind zu erwarten?

Zunächst muß der Begriff *Teilekandidat* erweitert werden: Im folgenden wird mit diesem Begriff vor allem ein Modell bezeichnet, daß die Merkmale eines Teilbaums in einem Dendrogramm enthält, und diese mit Hilfe eines übergeordneten Knotens zusammenfaßt.

Da das in dieser Arbeit eingesetzte Modell eine Graphstruktur besitzt, bieten sich Ähnlichkeitsmaße aus der Graphentheorie an. Tests auf exakte Graphenübereinstimmung sind aufgrund des Bildrauschens uninteressant. Eine alternative Methode besteht darin, die Ähnlichkeit in der Anzahl von Manipulationsschritten zu messen, die nötig sind, um einen Graphen in einen anderen zu überführen. Als Manipulationsschritte zählt beispielsweise das Einfügen oder Entfernen von Knoten. Da die in den Blattknoten gespeicherten Merkmale jedoch eine visuelle Bedeutung haben, ist die Bewertung von Graphmanipulationen anspruchsvoll. Darüberhinaus können Graphmanipulationen zu inkonsistenten Modellen führen, z.B. mit Kanten- und Flächenmerkmalen an der gleichen Position. Die Berücksichtigung entsprechender Bedingungen ist modellierungs- und rechenaufwendig.

Aus diesem Grund werden Teilekandidaten hier aufgrund ihrer visuellen Erscheinung verglichen. Dazu wird ausgenutzt, daß ein Teil sowohl als Muster als auch als Klassifikator betrachtet werden kann. Wenn ein Teil als Klassifikator betrachtet wird, läßt es sich dadurch charakterisieren, welche Muster durch den Klassifikator erkannt werden. Um eine Menge von passenden Testmustern zu erhalten, werden Teilekandidaten wiederum als Muster betrachtet. Auf diese Weise wird zu jedem Teilekandidat ein Vektor erzeugt, der zu jedem anderen Teilekandidaten angibt, ob das dort gespeicherte Muster erkannt wird oder nicht. Der Vektor hat genauso viele Dimensionen wie Teilekandidaten vorliegen. Die einzelnen Vektorkomponenten enthalten die Werte 0 oder 1, je nach Übereinstimmung mit dem entsprechenden Muster. Die Strategie, Teilekandidaten über ihre Ähnlichkeit zu einer Menge anderer Teile zu beschreiben, stellt eine Analogie dar zu der von Murase und Nayar [NMN96] eingesetzten Bilddarstellung in einem Vektorraum, dessen Basisvektoren wiederum Bilder sind.

Aufgrund der hohen Anzahl an Knoten in den Dendrogrammen, kann die Clusterung nicht auf allen Teilekandidaten durchgeführt werden. Aus Geschwindigkeitsgründen müssen während der Clusterung alle Vektoren, die Teilekandidaten repräsentieren, gleichzeitig im Speicher gehalten werden. Der Speicherbedarf ergibt sich aus der Anzahl der zu clusternden Teilekandidaten multipliziert mit der Anzahl der Muster. Die Speicherausstattung des zur Clusterung eingesetzten Versuchsrechners (Intel E6850, 3GHz) von 8GB RAM begrenzt die Anzahl zu clusternder Teilekandidaten auf ca. 30 000–40 000. Da während der Clusterung selbst zusätzlicher temporärer Speicher benötigt wird, wird eine Teileanzahl von 30 000 gewählt, was eine Darstellungsgröße von etwa 4GB ergibt.

Die 30 000 Teilekandidaten werden durch stochastische Unterabtastung der Kandidatenmenge zusammengestellt. Aus der Baumstruktur der Dendrogramme ergibt sich jedoch die Komplikation, daß kleine Teile im Vergleich zu größeren

überproportional stark vertreten sind (siehe Abbildung 6.47). Dies ist problematisch, da in der unterabgetasteten Menge bei einer rein stochastischen Abtastung nicht genügend vergleichbare Teile zur Clusterung größerer Teile vorliegen. Aus diesem Grund wird die Unterabtastung so reguliert, daß die Teilegröße über den 30 000 Teilekandidaten gleichverteilt ist. Als weitere Erleichterung wird die Teilegröße auf den Bereich von 10 bis 60 Pixeln eingeschränkt. Zum einen reduziert sich dadurch die Kandidatenmenge auf etwa 200 000. Der Hauptvorteil liegt jedoch darin, das System von Teilen zu entlasten, die aufgrund einer geringen Zahl von Merkmalen keine stabile Erkennung erlauben, oder die so groß sind, daß die Teilehierarchie zu einer Kette degeneriert.

Zur Clusterung wird das bereits bei der Gruppierung von Merkmalen eingesetzte Verfahren genutzt. Das Ergebnis der Clusterung ist dementsprechend wieder ein Dendrogramm, daß hierarchisch Gruppen von ähnlichen Teilen angibt. Höhere Knoten im Dendrogramm entsprechen größeren Clustern mit zunehmend unähnlichen Teilen. Aufgrund der Informationen zur Clusterungshierarchie bietet das resultierende Dendrogramm mehrere Möglichkeiten zur Einteilung der Teilekandidaten in homogene Gruppen. Nach der Gruppierung von Teilekandidaten bezüglich ihrer gegenseitigen Ähnlichkeit müssen die eigentlichen Cluster daher noch identifiziert werden. Zu jedem Cluster wird dann ein prototypischer Teilekandidat ermittelt, der in das visuelle Alphabet aufgenommen wird.

Bei der Clusterung hat sich die Abstandsberechnung zwischen den mit 30 000 Komponenten recht großen Vektoren als laufzeitentscheidend herausgestellt. Die Laufzeit kann jedoch über eine kompaktere Kodierung ψ (vgl. Gl. 6.13) verbessert werden. Diese basiert auf der Beobachtung, daß die Vektoren zu Beginn der Clusterung hauptsächlich den Wert Null enthalten, der die Inkompatibilität zu einem Muster anzeigt. Die Kodierung speichert daher anstelle längerer Folgen von Nullen nur deren Anzahl. Die Abstandsberechnung wird so an die kompaktere Kodierung angepaßt, daß sich die gleichen Resultate ergeben wie bei der unkodierten Darstellung. Die Kompaktierung der Vektoren reduziert die Rechendauer der Clusterung von etwa vier Tagen auf nur wenige Stunden.

Die Clusterung hochdimensionaler Daten unterliegt oft dem sogenannten *Fluch der Dimensionalität*, d.h. einem starken Anwachsen des Rechenaufwands und zunehmend instabilen Ergebnissen mit steigender Dimensionalität. Als *hochdimensional* gelten nach den Erfahrungen von Beyer et al. [BGRS99] bereits Vektoren mit 10 bis 15 Dimensionen. Was die Geschwindigkeit angeht, ist das vorliegende Verfahren für die Clusterung des Teilealphabets performant genug. Bezüglich der Stabilität zeigen Beyer et al., daß der Begriff des *nächsten Nachbars* in hochdimensionalen Vektorräumen bedeutungslos wird, falls der minimale und der maximale Abstand zwischen zwei Vektoren mit steigender Dimensionalität gegeneinander konvergieren. Dies ist im vorliegenden Fall zu Beginn der Clusterung nicht der Fall, da die Vektoren dünn besetzt sind und immer auf den Achsen des Koordinatensystems liegen. Die mit 1 besetzten Komponenten überschneiden sich bei Vektoren verschiedener Muster zudem kaum. Im Verlauf der Clusterung werden jedoch immer mehr kleine Cluster zu großen Clustern auf höheren Hierarchieebenen zusammengefaßt. Große Cluster werden dabei durch einen Mittelwertvektor aus den kleinen Clustern beschrieben.

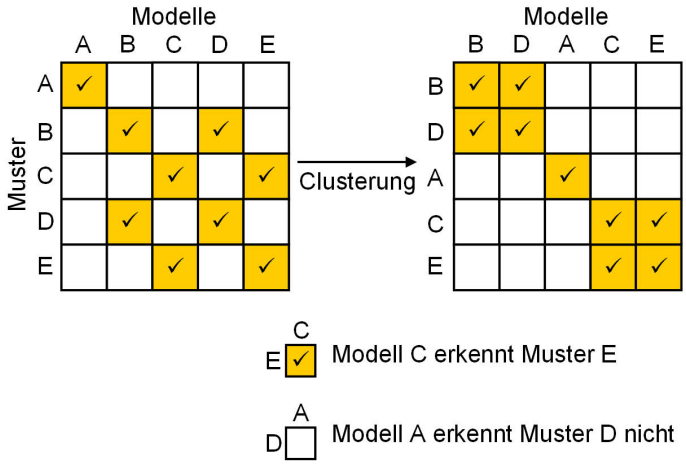


Abbildung 6.48: **Eingabe und Ergebnis der Clusterung von Teilekandidaten.** Als Eingabe dient eine Matrix, die für jeden Teilekandidaten angibt, welche Muster dieser modelliert. Durch die Clusterung rücken ähnliche Spalten und Zeilen zusammen. Cluster zeigen sich nun durch kompakte Blöcke mit Markierungen erkannter Muster.

Diese Mittelwertvektoren liegen nicht mehr auf den Koordinatenachsen. Durch das Mitteln über zunehmend unähnliche Cluster nehmen auch die mit Null besetzten Vektorkomponenten ab. Die Vektoren nähern sich dem Mittelwert der gesamten Vektormenge an. Dadurch sinkt die Varianz der Vektoren und das Ähnlichkeitsmaß bei der Clusterung verliert zunehmend an Bedeutung. Als Fazit läßt sich daher sagen, daß kleine Cluster vom Fluch der Dimensionalität verschont bleiben. Cluster auf höheren Stufen des Dendrogramms enthalten dagegen nicht nur aufgrund der größeren Menge an Teilen, sondern auch aufgrund des Fluchs der Dimensionalität zunehmend unähnliche Teile. Hinneburg und Aggarwal [AHK01, HAK00] zeigen, daß auch das gewählte Abstandsmaß einen Einfluß hat. Sie stellen fest, daß kleine Normen für hohe Dimensionalitäten stabilere Ergebnisse liefern und schlagen daher gebrochene Normen vor. Die im Rahmen der vorliegenden Arbeit durchgeführten Versuche bestätigen dies. Zur Teileclusterung wird daher eine Norm von $\gamma = 0{,}125$ (vgl. Gl. 6.14) gewählt, die hinsichtlich der Rechendauer einen erträglichen Kompromiß darstellt.

Die Eingabe der Clusterung kann als *Kandidatenmatrix* I_K dargestellt werden. Die Spalten der Matrix entsprechen den zu clusternden Vektoren. Die Matrix hat daher die Form

$$I_K = \begin{bmatrix} \iota_{1,1} & & \iota_{30000,1} \\ & \ddots & \\ \iota_{1,30000} & & \iota_{30000,30000} \end{bmatrix},$$

wobei ein Matrizenelement $\iota_{i,j}$ mit einem Wert von 1 einen Treffer, d.h. eine Erkennung des Musters mit dem Index j durch den Teilekandidaten i, angibt. Andernfalls wird der Wert auf Null gesetzt. Die Matrix wird ferner so angelegt, daß die Reihenfolge der Teilekandidaten mit der Reihenfolge der Muster übereinstimmt, d.h. der Teilekandidat i speichert das Muster j, falls i und j gleich sind. Da aufgrund der zufälligen Unterabtastung der Menge an Teilekandidaten keine Beziehung zwischen benachbarten Teilen i und $i+1$ besteht, kann aus I_K keine Clusterung direkt abgelesen werden. Die Matrix ist zunächst ungeordnet. Dies illustriert das linke Diagramm in Abbildung 6.48.

Durch die Clusterung werden nun ähnliche Muster zusammengefaßt. Gruppen ähnlicher Muster und Teilekandidaten werden dadurch sichtbar gemacht, daß die hierarchische Clusterstruktur, die in dem Dendrogramm der Teileähnlichkeit gespeichert ist, auf die Sortierung der Muster und Teilekandidaten übertragen wird. Cluster zeigen sich nun in Blöcken von Treffern in der Kandidatenmatrix. Dabei spiegelt sich die Homogenität der Cluster darin wider, wie dicht die Blöcke mit Treffern gefüllt sind. Dies illustriert der rechte Teil von Abbildung 6.48. Um das visuelle Alphabet zusammenzustellen, werden daher die Knoten des Dendrogramms ausgewählt, für die sich besonders deutliche Trefferblöcke ergeben.

Dabei ist jedoch zu beachten, daß die Kandidatenmatrix sowohl disjunkte Cluster als auch ein Kontinuum ähnlicher Teilekandidaten anzeigen kann. Wenn sich die übereinstimmenden Muster und Teilekandidaten wie in der Abbildung zu deutlich getrennten, quadratischen Blöcken von Treffern gruppieren, lassen sich die Teilekandidaten tatsächlich in separate Cluster einteilen. Wenn sich die Eigenschaften der Teilekandidaten jedoch überschneiden, ist auch eine Überschneidung der Trefferblöcke in der sortierten Matrix zu erwarten. Der Extremfall bildet ein Kontinuum von Mustern und Teilekandidaten. In der sortierten Matrix sollte sich dieses in einem Streifen von Treffern entlang der Hauptdiagonalen zeigen. Die Breite des Streifens spiegelt die Selektivität der Teilemodelle wider. Welcher Fall vorliegt, wird im folgenden experimentell ermittelt.

6.3.2 Berechnung der Kandidatenmatrix

Bevor die Clusterung durchgeführt werden kann, muß die Kandidatenmatrix I_K berechnet werden. Dies geschieht in folgenden Schritten:

- Auswahl von 30 000 Knoten in Dendrogrammen der Stichprobe
- Bestimmung der zugehörigen Merkmalsgruppen (Muster)
- Erzeugung von Modellen zu allen Merkmalsgruppen (Teilekandidaten)
- Sinnvolle Parametrisierung der Teilekandidaten
- Durchführung der Objekterkennung für jede Kombination aus Muster und Teilekandidat

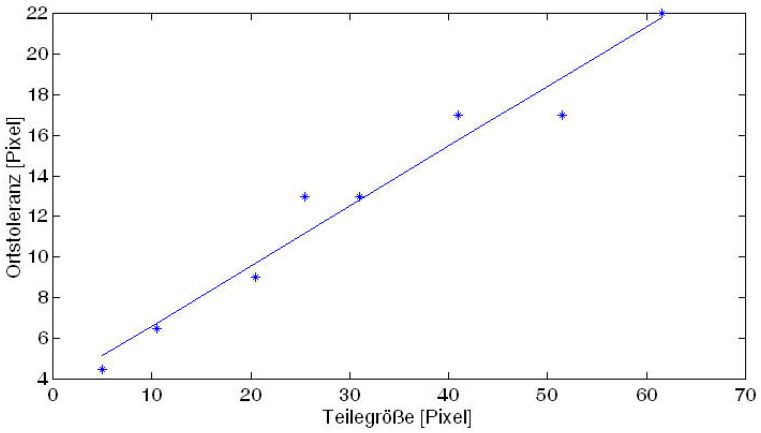


Abbildung 6.49: **Lineare Näherungslösung für die Teilegröße.** Die Sterne markieren die Parametrisierungen, für die sich die besten Probeclusterungen ergeben (siehe Tabelle 6.9). Die Gerade zeigt die lineare Näherung aus Gleichung 6.17.

Die ersten Punkte wurden bereits beschrieben. Es muß allerdings noch geklärt werden, wie die Modell- und Verfahrensparameter gewählt werden. Als Verfahrensparameter muß zunächst die Anzahl der Quantisierungsstufen der Merkmale festgelegt werden. Aufgrund der vorangegangenen Versuche wird die Gradientenrichtung in 10 Stufen und die Richtung von Skelettlinien in 12 Stufen eingeteilt. Da die Clusterungsergebnisse auch von der Genauigkeit der Merkmalsausprägung abhängen, werden probeweise Clusterungen auf kleineren Datenmengen durchgeführt. Aufgrund von visuellen Begutachtungen der zu Clustern zusammengefaßten Muster wird für die weiteren Merkmale eine Merkmalseinteilung in jeweils 10 Stufen gewählt. Für diese Parameter ergeben sich nicht zu viele Treffer in der Kandidatenmatrix und homogene Gruppen von Mustern. Die visuelle Prüfung ist zwar subjektiv, aber nicht unsinnig. Zum einen ist aufgrund des großen Parameterraums und des hohen rechnerischen Aufwands eine automatische Optimierung der Clusterungsparameter schwer durchführbar. Zum anderen ist das menschliche Sehsystem die Referenz in der Objekterkennung. Darüberhinaus sind günstige und ungünstige Parametrisierungen leicht visuell voneinander zu unterscheiden.

Neben der Merkmalsquantisierung müssen auch die Teilekandidaten parametrisiert werden. Für jeden Teilekandidaten muß die Ortstoleranz und der Schwellwert des Wurzelknotens festgelegt werden. Aufgrund der vorangegangenen Versuche wird der Schwellwert wieder auf 90 Prozent festgelegt. Bei der Bestimmung der Ortstoleranz muß die Abhängigkeit von der Teilegröße berücksichtigt werden. Dazu werden für verschiedene Teilegrößen Probeclusterungen mit unterschiedlichen Ortstoleranzen durchgeführt. Mangels besserer Methoden werden die Ergebnisse wieder visuell begutachtet. Für die Probeclusterungen werden 300 Teilekandidaten mit ihren 300 Mustern verglichen. Ta-

Teilegröße:	Ortstoleranz [Pixel]					
5 Pixel	2	3	4	5	10	
Anzahl Treffer		3000	4577	6068	>8000	
Anzahl Cluster	241	181	153	133	121	
Kommentar			gut	gut		
Teilegröße:	Ortstoleranz [Pixel]					
10–11 Pixel	2	5	8	11		
Anzahl Treffer	388	2440	7866	9232		
Anzahl Cluster	294	228	161	136		
Kommentar		gut	gut			
Teilegröße:	Ortstoleranz [Pixel]					
20–21 Pixel	5	8	10	15	20	
Anzahl Treffer	405	940	2358	8627	15 349	
Anzahl Cluster	297	274	251	168	131	
Kommentar		gut	gut			
Teilegröße:	Ortstoleranz [Pixel]					
25–26 Pixel	5	10	13	15		
Anzahl Treffer	318	1102	3081	5967		
Anzahl Cluster	300	288	243	194		
Kommentar		ok	gut			
Teilegröße:	Ortstoleranz [Pixel]					
30–32 Pixel	5	10	13	15	20	25
Anzahl Treffer		614	1601	2687	11 457	19 651
Anzahl Cluster	299	295	272	247	165	123
Kommentar			gut			
Teilegröße:	Ortstoleranz [Pixel]					
40–42 Pixel	10	15	17	20		
Anzahl Treffer	354	1097	2139	6402		
Anzahl Cluster	298	283	252	195		
Kommentar			gut			
Teilegröße:	Ortstoleranz [Pixel]					
50–53 Pixel	10	15	18	20		
Anzahl Treffer	307	601	1510	3525		
Anzahl Cluster	300	296	265	235		
Kommentar		gut	gut			
Teilegröße:	Ortstoleranz [Pixel]					
60–63 Pixel	10	15	20	22	25	30
Anzahl Treffer	307	384	1557	3515	7595	16 653
Anzahl Cluster	300	300	269	248	201	154
Kommentar				gut		

Tabelle 6.9: Clusterung der Kandidatenmatrix für verschiedene Teilegrößen und Ortstoleranzen. Zu kleine Ortstoleranzen ergeben viele kleine Cluster. Zu hohe Ortstoleranzen gruppieren dagegen viele, unähnliche Muster. Die besten Parametrisierungen sind mit dem Kommentar "gut" gekennzeichnet.

belle 6.9 gibt die Bewertungen an. Für jeden Größenbereich wird die jeweils günstigste Ortstoleranz bestimmt. Um auch für die Zwischengrößen eine Ortstoleranz bestimmen zu können, wird per Regression eine Gerade an die besten Meßwerte angepaßt. Es ergibt sich die Geradengleichung

$$\varsigma = 0{,}29 \cdot r_t + 3{,}66, \tag{6.17}$$

welche im Folgenden zur Berechnung der Ortstoleranz ς aus einer Teilegröße r_t eingesetzt wird. Abbildung 6.49 zeigt die Lage der Geraden in den Meßwerten. Die Anpassung einer komplexeren Funktion erscheint nicht notwendig.

6.3.3 Ergebnisse der Clusterung von Teilekandidaten

Die zur Clusterung vorgesehenen Teilekandidaten werden gemäß Gleichung 6.17 parametrisiert. Nun wird die Kandidatenmatrix I_K berechnet. Das Ergebnis zeigt Abbildung 6.50. Schwach zu erkennen sind die Treffer auf der Diagonalen von links oben nach rechts unten. Diese bestätigen insofern die erfolgreiche Parametrisierung der Teilekandidaten, als daß die Teilekandidaten die in ihnen gespeicherten Muster wiedererkennen. Andernfalls wären die Modelle zu selektiv eingestellt. Ansonsten ist, wie bereits durch die Schemazeichnung in Abbildung 6.48 angedeutet, eine Clusterung nicht direkt erkennbar. Die Regellosigkeit der Trefferanordnung in der Kandidatenmatrix belegt stattdessen die Randomisierung der Stichprobe. Eine leichte Tendenz zu vermehrten Treffern im rechten oberen Bereich stammt aus der Methode zur Unterabtastung der vollständigen Menge an Teilekandidaten. Die Tendenz kommt dadurch zustande, daß bei der Auswahl der Teilekandidaten die Größe der Teile berücksichtigt wird. So ergibt sich eine gewisse Größenabhängigkeit bei der Teileanordnung in der Kandidatenmatrix. Die ungleichen Trefferhäufigkeiten ergeben sich nun daraus, daß große Teile mehr erkennbare Untermuster enthalten, als Klassifikator aber selbst selektiver arbeiten.

Ober- und Untermengen

Durch die Clusterung werden die Teilekandidaten bezüglich der Ähnlichkeit der Spalten der Kandidatenmatrix zu möglichst homogenen Gruppen zusammengestellt. Das Resultat liegt wie bei der Clusterung von Merkmalen in Form eines Graphen als Dendrogramm vor. Durch das systematische Durchlaufen dieses Graphen ergibt sich eine bestimmte Reihenfolge von Teilekandidaten. Diese wird wieder auf die Kandidatenmatrix übertragen. Die Reihenfolge der Muster wird entsprechend angepaßt. Abbildung 6.51 zeigt einen Ausschnitt einer geclusterten Kandidatenmatrix. Die Trefferblöcke, die aus der Anordnung ähnlicher Spalten und der entsprechenden Umsortierung der Zeilen ergeben, sind klar erkennbar.

An der höheren Dichte an Treffern im rechten oberen Teil der initialen Kandidatenmatrix war bereits zu sehen, daß die tatsächliche Kandidatenmatrix anders als die vereinfachte Schemazeichnung in Abbildung 6.48 nicht symmetrisch zur Hauptdiagonalen ist. Nach der Clusterung zeigen sich diese Asymmetrien in Trefferblöcken auf einer Seite der Hauptdiagonalen, ohne daß ein entsprechender

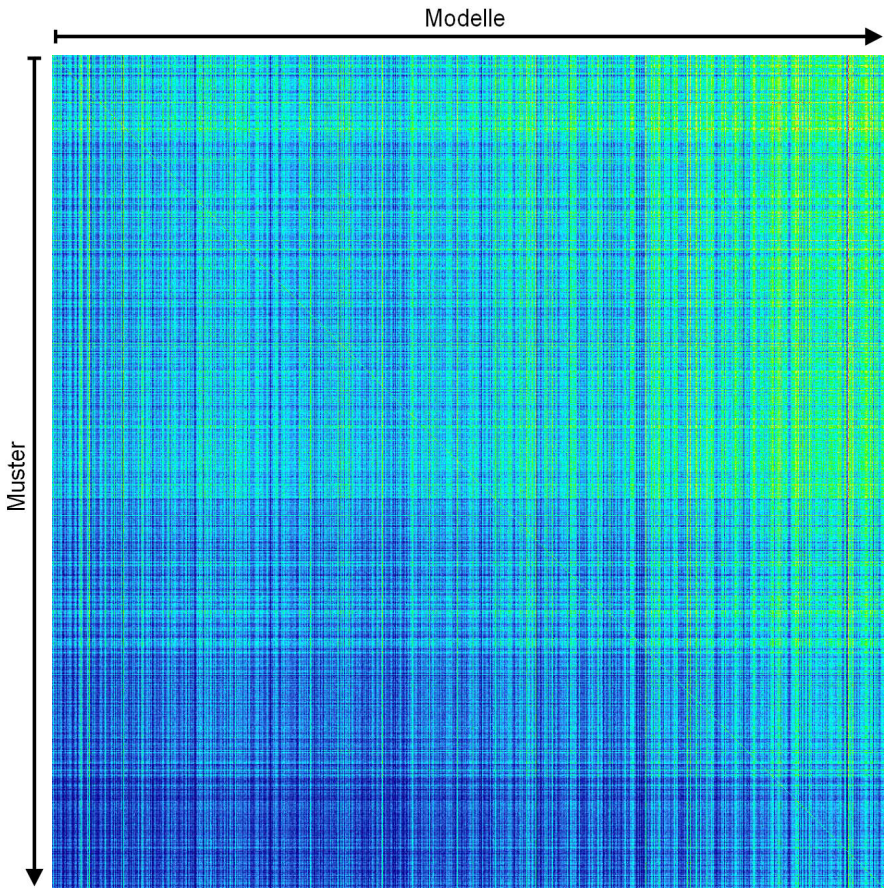


Abbildung 6.50: **Initiale Kandidatenmatrix.** Natürlicherweise sollte die Kandidatenmatrix als Binärbild dargestellt werden, wobei jedes Element einem weißen oder schwarzen Pixel entspricht. Hier ist jedoch eine unterabgetastete Version dargestellt, da die originale Matrix zu groß für den Druck ist. Dunkelblaue Bereiche zeigen wenige Treffer in der Kandidatenmatrix an, Gelb und Rot zeigen dagegen Bereiche mit vielen Treffern an. Entlang der Horizontalen werden die Teilekandidaten variiert. Entlang der Vertikalen werden die Muster variiert.

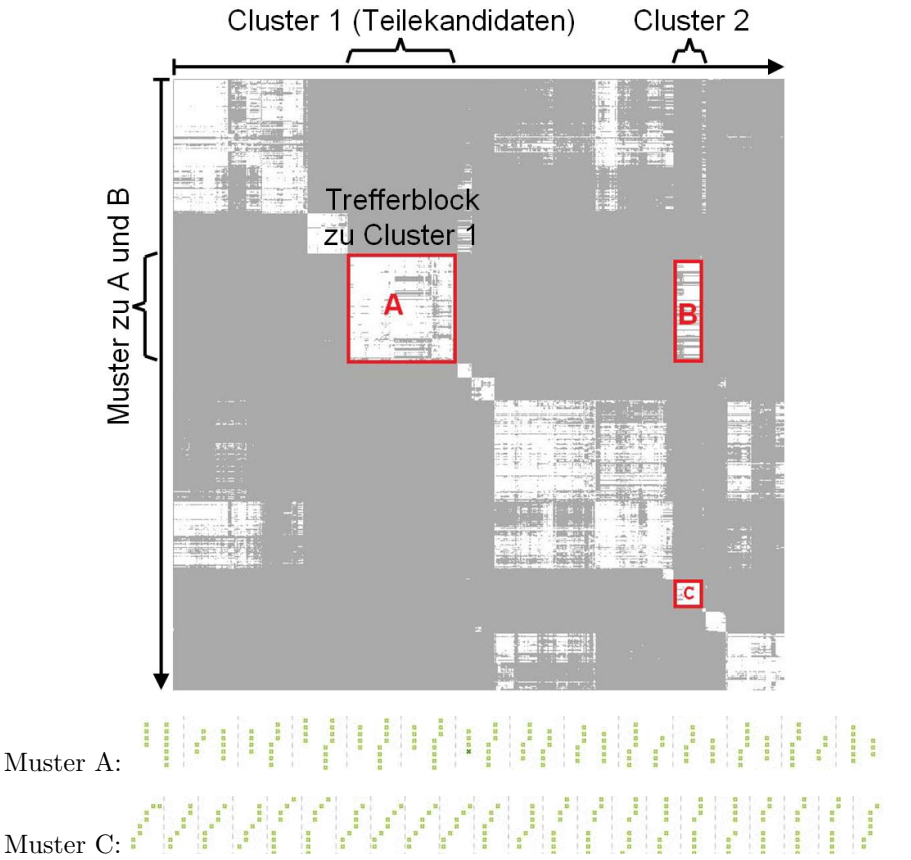


Abbildung 6.51: **Ober- und Untergruppen von Mustern.** Nach der Clusterung der Teilekandidaten zeigen sich Ober- und Unterklassen von Mustern. Der Trefferblock B zeigt an, daß Merkmale aus den Teilekandidaten von Cluster 2 als Unterknoten in den Teilekandidaten von Cluster 1 enthalten sein müssen. Genauso müssen die Muster zu Trefferblock C als Teilmuster in den Mustern zu den Blöcken A und B enthalten sein. Unter der Kandidatenmatrix sind einige Muster zu den Blöcken A und C angegeben. Die hellen Quadrate markieren Kantenpunkte. Verschiedene Muster sind durch gestrichelte Linien getrennt. Die Muster zu den Blöcken A und B bestehen aus zwei senkrechten Kantenzügen. Die Muster zu Block C bestehen dagegen nur aus einem Kantenzug.

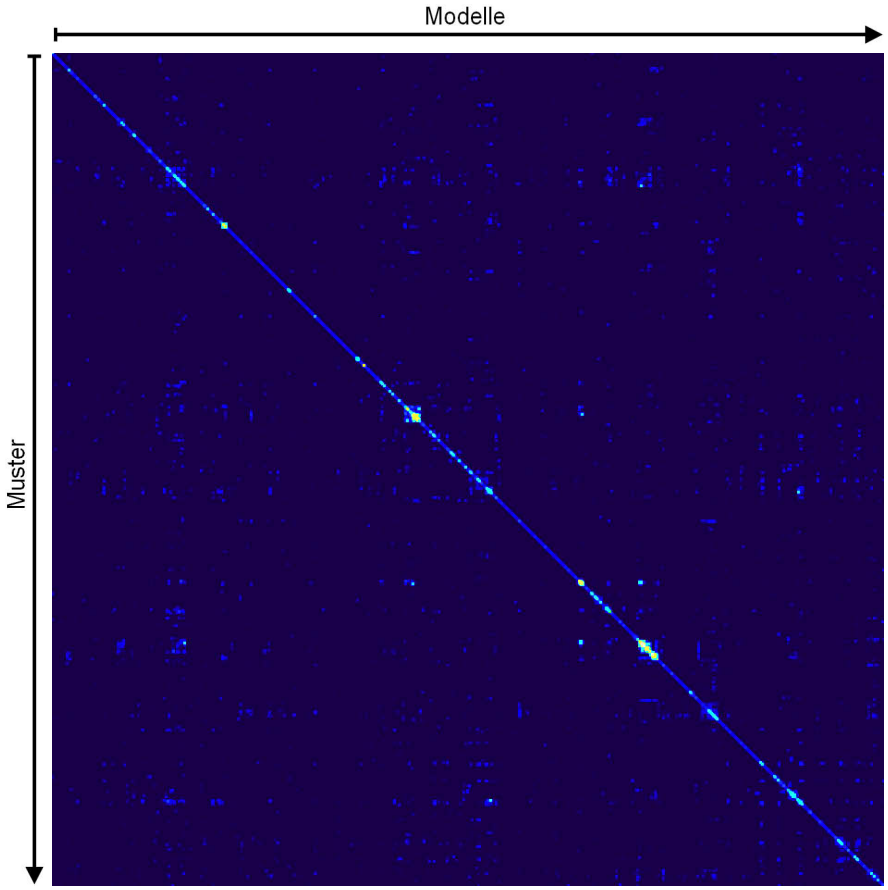


Abbildung 6.52: **Geclusterte Kandidatenmatrix.** Durch eine logische Verundung der Treffer an relativ zur Hauptdiagonale gespiegelten Positionen wurden die Beziehungen zwischen Unter- und Obergruppen von Mustern aus der Kandidatenmatrix gelöscht. Anschließend wurden ähnliche Muster und Teile durch eine Clusterung bezüglich der Spaltenähnlichkeit gruppiert.

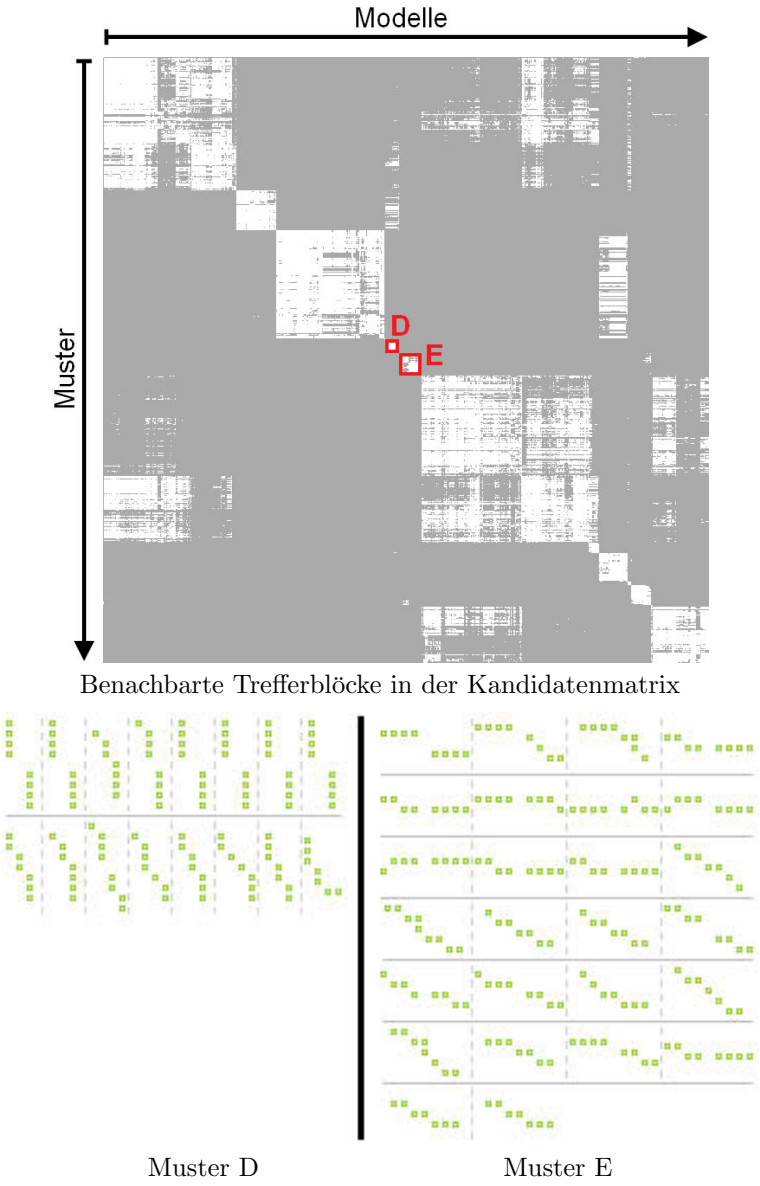


Abbildung 6.53: **Benachbarte Trefferblöcke in der Kandidatenmatrix.** Die Teilekandidaten zu den Trefferblöcken D und E haben keine gemeinsamen Treffer. Obwohl sie in der Treffermatrix direkt benachbart sind, sind die gespeicherten Muster stark verschieden. Hier liegt offenbar eine Grenze zwischen zwei großen Teilbäumen des Dendrogramms.

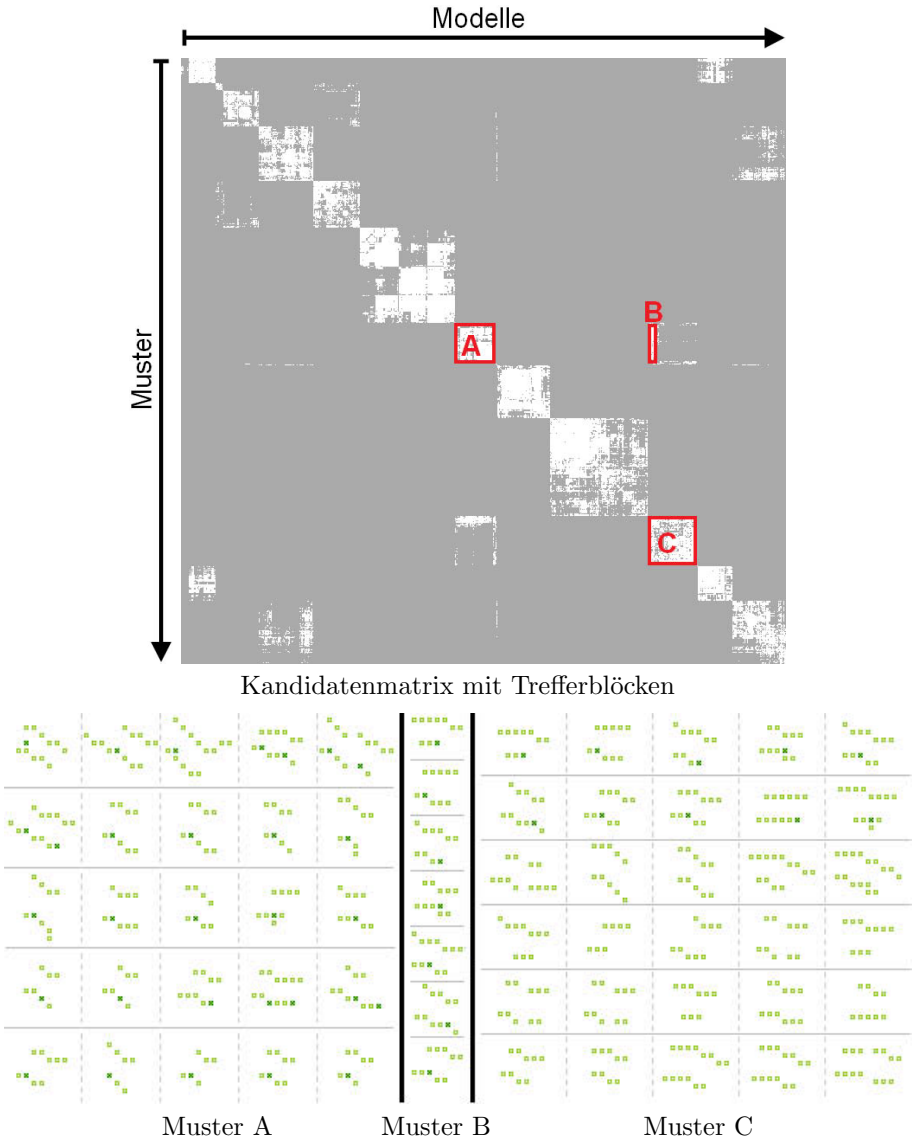


Abbildung 6.54: **Sich berührende Cluster in der Kandidatenmatrix.** Die Cluster zu den Trefferblöcken A und B müssen einen Berührungspunkt, d.h. einige ähnliche Teilekandidaten, besitzen. Die Gemeinsamkeit zeigt sich durch den kleinen Trefferblock B. Die Unterschiede zwischen A und C liegen in den durch ein dunkelgrünes Kreuz gekennzeichneten Eckenmerkmalen und einem geringfügig steileren Winkel der Kanten in A (nicht alle Muster sind angegeben).

Abbildung 6.55: **Kontinuum in der Kandidatenmatrix.** Die Cluster zu den Trefferblöcken A, B und C bilden ein Kontinuum von Mustern ab. Die Blöcke A und C sind über E verbunden. C und B sind über F verbunden. A und B sind aufgrund der wenigen Treffer in D nur schwach verbunden. Unten sind einige der in den Teilkandidaten gespeicherten Muster angegeben. Diese bestehen nur aus Skelettmerkmalen, d.h. die Teile stellen aufrechte, leicht nach links geneigte, längliche Flächen dar.

Block auf der gespiegelten Position vorliegt. In dem in Abbildung 6.51 gezeigten Beispiel werden die Muster zu Trefferblock A von den Teilekandidaten aus den Clustern 1 und 2 erkannt. Die Muster aus Block C werden dagegen nur durch Cluster 2 erkannt. Die Teilekandidaten in Cluster 2 sind daher allgemeiner als die in Cluster 1. Aufgrund der Übereinstimmungen in Trefferblock B müssen zudem Merkmale aus den Teilekandidaten von Cluster 2 als Unterknoten in den Teilekandidaten von Cluster 1 enthalten sein. Mit einer analogen Begründung müssen die Muster zu Block C als Teilmuster in den Mustern zu Block A bzw. B enthalten sein. Die Überprüfung der tatsächlich zu den Trefferblöcken gespeicherten Muster bestätigt dies. Die Muster zu Block C stellen eine einzelne senkrechte oder leicht nach rechts geneigte Kante dar. Die Muster zu Block A stellen zwei parallele, geringfügig kürzere senkrechte Kanten dar. Asymmetrische Trefferblöcke zeigen also Ober- und Untermengen von Teilen und Mustern an.

Bezüglich der Clusterung stellt sich nun die Frage, ob Ober- und Untergruppen von Teilekandidaten in den gleichen Cluster oder in getrennte Cluster eingeordnet werden sollen. Für eine Kombination von Ober- und Untergruppen in einen Cluster bietet sich eine logische Veroderung von Treffern an Positionen symmetrisch zur Hauptdiagonalen der Kandidatenmatrix an. Auf diese Weise werden nicht besetzte Trefferpositionen aufgefüllt, wenn auf der an der Diagonalen spiegelbildlichen Position bereits ein Treffer vorhanden ist. Dadurch wird eine symmetrische Kandidatenmatrix erzeugt. Fehlende Treffer in den Trefferblöcken werden ergänzt. Auf diese Weise sollten sich klarere Trefferblöcke ergeben, welche die Identifikation von Clustern und damit die Auswahl von Knoten des Dendrogramm bei der Zusammenstellung des visuellen Alphabets erleichtern. In der Praxis ist dies keine gute Strategie. Aufgrund der hohen Zahl an unsymmetrischen Trefferpositionen ergibt eine Veroderung der Treffer eine dicht besetzte Matrix und infolgedessen große und undeutliche Trefferblöcke. Eine alternative Methode ist, Ober- und Untergruppen von Teilekandidaten grundsätzlich in getrennte Cluster einzuordnen. Dies kann dadurch erreicht werden, daß statt der Veroderung ein logisches UND auf spiegelbildliche Trefferpositionen angewandt wird. In der Praxis reduziert dies die Anzahl der Treffer stark. Die resultierenden Cluster sind kleiner und deutlicher ausgeprägt als in der unbehandelten Kandidatenmatrix. Aus diesem Grund wird das Verfahren sowohl für die Probeclusterungen zur Ermittlung der Parametrisierungen von Teilekandidaten eingesetzt als auch zur Clusterung der Kandidatenmatrix. Abbildung 6.52 zeigt die geclusterte Kandidatenmatrix für eine vorherige Verundung der Treffer. Im folgenden werden weitere Effekte erläutert, die sich aus der Kandidatenmatrix ablesen lassen.

Benachbarung von Clustern

Wie die Abbildung 6.51 zeigt, sind die Teilekandidaten innerhalb der durch die Trefferblöcke angedeuteten Gruppen homogen. Da benachbarte Trefferblöcke jedoch aus stark verschiedenen Teilbäumen des Dendrogramms stammen können, gilt die Ähnlichkeit nicht generell über benachbarte Trefferblöcke. Ein Beispiel

zeigt Abbildung 6.53. Die Teilekandidaten zu Block D stellen senkrechte Kanten dar. Die Teilekandidaten zu Block E repräsentieren dagegen horizontal verlaufende Kanten.

Berührungspunkte zwischen Clustern

Die Muster verschiedener Trefferblöcke können jedoch auch sehr ähnlich sein. Die Ähnlichkeit der entsprechenden Muster zeigt sich in teilweise dünn besetzten Trefferblöcken neben der Hauptdiagonalen. Diese geben gemeinsame Berührungspunkte zwischen den Clustern zu dichter besetzten Trefferblocken in den gleichen Zeilen bzw. Spalten auf der Hauptdiagonalen an. Abbildung 6.54 zeigt ein Beispiel. Die Trefferblöcke A und B liegen auf der Hauptdiagonalen. Der Trefferblock B gibt an, daß einige der Teilekandidaten zu Block C auch Muster aus A erkennen. Da B im Vergleich zu A und C sehr klein ist, enthält C nicht in erster Linie Teilmuster von A, sondern ist als eigenständiger Cluster anzusehen. Die in den Teilekandidaten zu A gespeicherten Muster sind zwei kurze, einigermaßen parallele Kantenstücke, die zusätzlich ein Kantenmerkmal enthalten. Bei den Mustern zu C handelt es sich ebenfalls um schräge, parallele Kanten. Abgesehen von den ersten Mustern enthalten diese jedoch keine Ecken. Zudem liegen die Kanten minimal waagerechter. Aufgrund der hohen Schwelle von 90% und der geringen Zahl von nur etwa 15 Merkmalen wirkt sich das Fehlen des Eckenmerkmals in C stark genug aus, um einen eigenen Cluster zu erzeugen. Die Muster zu B liegen zwischen A und C. Sie enthalten eine Ecke, liegen jedoch meistens waagerechter als die Muster aus A.

Verkettungen von Clustern

Wenn die Ähnlichkeiten zwischen den Clustern noch größer werden, können sich Ketten von Clustern bilden, die jeweils paarweise Ähnlichkeiten zeigen. Solche Ketten können kontinuierliche Abbildungen visueller Charakteristika von Mustern abbilden. Ein Beispiel wird in Abbildung 6.55 gezeigt. Hier zeigen die Trefferblöcke A, B und C drei Cluster an. Da in Bock D kaum Treffer auftreten, bestehen nur schwache Verbindungen zwischen den Clustern A und B. Die Blöcke E und F sind dagegen relativ dicht besetzt. Aus dem Trefferblock E läßt sich schließen, daß sich die Cluster A und C stark überlappen. Aus Block F ergibt sich weiterhin, daß B und C sehr ähnlich sind. Daraus ergibt sich eine Kette von A über C nach B. Die gespeicherten Muster bestehen ausschließlich aus kettenförmig angeordneten Skelettmerkmalen. Die Muster stellen also längliche Flächen dar. Über die Cluster A, C und B wird die Orientierung der Fläche kontinuierlich variiert. Die Dichte der Treffer innerhalb der Blöcke schwankt jedoch stark, sodaß sich kaum klare Entscheidungen über die Verkettung von Clustern treffen lassen.

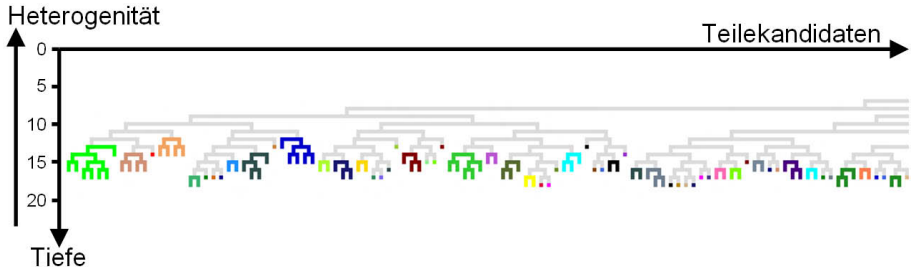


Abbildung 6.56: **Dendrogramm der Teilekandidaten.** Die Blattknoten des Graphen stellen die Teilekandidaten dar. Kanten verbinden ähnliche Gruppen von Teilekandidaten. Die Ähnlichkeit der zusammengefaßten Gruppen nimmt nach oben hin ab. Der hier dargestellte Ausschnitt zeigt 0,37% aller Teilekandidaten. Das Dendrogramm ist an der rechten Seite abgeschnitten. Die ermittelten Cluster sind verschiedenfarbig markiert.

6.3.4 Erzeugung eines visuellen Alphabets

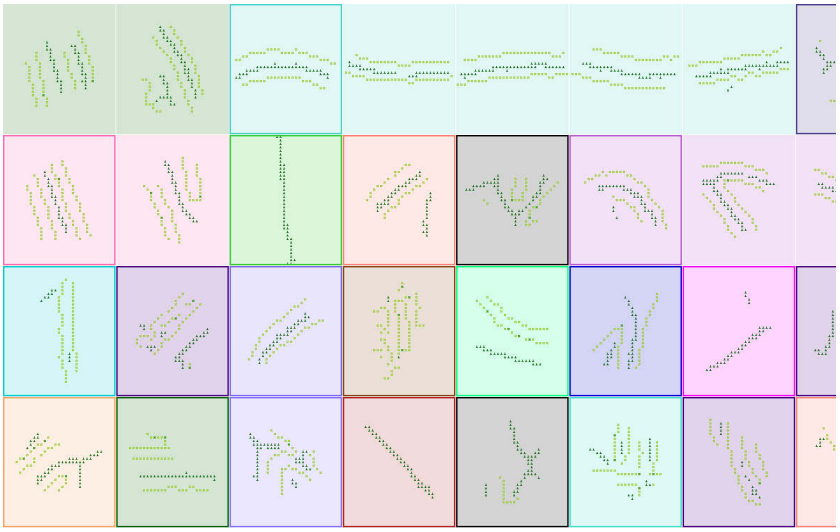
Nachdem nun ein Dendrogramm berechnet wurde, das in hierarchischer Form ähnliche Gruppen von Teilekandidaten zusammenfaßt, kann das visuelle Alphabet bestimmt werden. Dazu werden die Knoten des Dendrogramms ermittelt, die besonders homogene Gruppen von Teilekandidaten beschreiben. Aus jeder Gruppe wird ein Teilekandidat, der diese besonders gut repräsentiert, als Prototyp ausgewählt. Die Gesamtmenge der Prototypen bildet dann das visuelle Alphabet.

Dabei stellt sich die Frage, nach welchem Kriterium entschieden wird, ob eine Gruppe von Teilekandidaten homogen genug ist, um einen eigenen Cluster zu formen. Wie vorangehend dargestellt, spielt hier die Verteilung der Treffer in der Kandidatenmatrix eine Rolle. Eine Gruppe von Teilekandidaten wird daher dann als homogen angesehen, wenn es korrespondierende Trefferblöcke in der geclusterten Kandidatenmatrix gibt. Da der Übergang von einzelnen Treffern zu Blöcken von Treffern oft nur graduell ist, können Trefferblöcke jedoch nicht immer eindeutig identifiziert werden.

Die zweite Frage ist, wann ein Teilekandidat als typisch für einen ganzen Cluster angesehen wird. Da ein Teile-Prototyp den ganzen Cluster repräsentieren soll, wird gefordert, daß dieser als Modell für alle in dem Cluster gespeicherten Muster dienen soll, d.h. alle Muster des Clusters erkennt. Dadurch ergibt sich direkt ein Kriterium für die erste Frage: Eine Gruppe von Teilekandidaten bildet dann einen Cluster, wenn es einen Prototypen gibt, der alle Elemente des Clusters erkennt. Dies kann anhand der Kandidatenmatrix entschieden werden.

Der Algorithmus zur Ermittlung des visuellen Alphabets sieht daher wie folgt aus:

- Ausgehend von der Wurzel wird rekursiv im Dendrogramm abgestiegen.



Teilegröße: 20–21 Pixel, Ortstoleranz 10 Pixel

Teilegröße: 40–41 Pixel, Ortstoleranz 17 Pixel

Abbildung 6.57: **Clusterung von Teilekandidaten.** Die Abbildung zeigt die in einer Untermenge von Teilekandidaten gespeicherten Muster. Die Teilekandidaten eines Clusters liegen direkt nebeneinander und sind einfarbig hinterlegt. Die Clusterprototypen sind dick umrandet. Aus Darstellungsgründen zeigt die Abbildung Ergebnisse aus den Probeclusterungen.

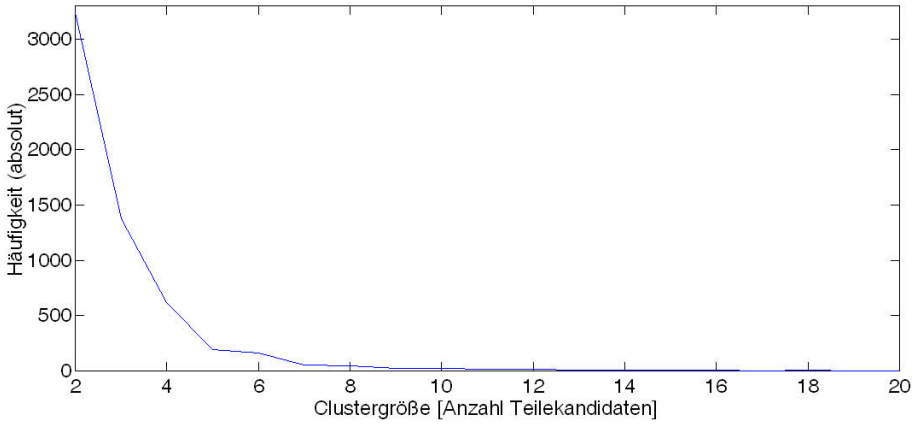


Abbildung 6.58: **Histogramm über die Clustergröße.** Das Histogramm gibt an, wieviele Cluster mindestens zwei Teilekandidaten umfassen. Aus Darstellungsgründen sind nur Clustergrößen bis 20 Teilekandidaten angegeben, obwohl selten auch größere Cluster auftreten.

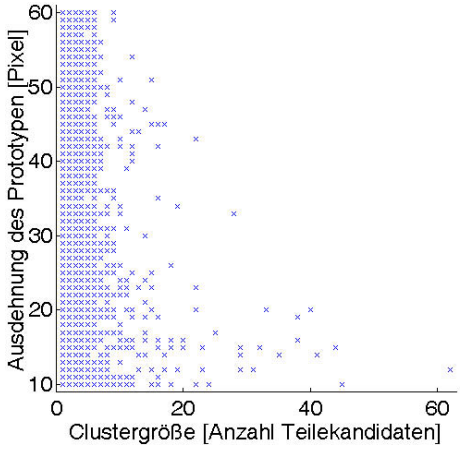


Abbildung 6.59: **Clustergröße und Teileausdehnung.** Das Streudiagramm gibt die Ausdehnung der in den Clusterprototypen gespeicherten Muster an. Alle Teilegrößen treten auf. Die Prototypen großer Cluster sind bevorzugt klein.

- An jedem Knoten wird die entsprechende Menge an Teilekandidaten bestimmt.
- Für diese Menge wird geprüft, ob es einen Prototypen gibt, der alle in der Menge gespeicherten Muster erkennt.
- Dies wird anhand der geclusterten Kandidatenmatrix entschieden.
- Falls es einen solchen Prototypen gibt, wird dieser in das visuelle Alphabet aufgenommen. Die Rekursion terminiert für den aktuellen Teilbaum.
- Andernfalls werden die Kindknoten rekursiv untersucht.

Da die Rekursion von der Wurzel des Dendrogramms ausgeht, werden immer maximal große Cluster erzeugt. Da die Rekursion zudem für jeden Teilbaum mit der Identifikation eines Prototypen endet, wird die Menge der Teilekandidaten in disjunkte Gruppen unterteilt. Abbildung 6.56 zeigt einen Teil des Dendrogramm über die 30 000 Teilekandidaten. Die Teilbäume, die einem Cluster mit Prototyp entsprechen, sind verschiedenfarbig markiert. Einen Eindruck sowohl von der Vielfalt der Teile innerhalb einzelner Cluster als auch innerhalb des visuellen Alphabets gibt Abbildung 6.57.

Wie das Histogramm über die Größen der resultierenden Cluster in Abbildung 6.58 zeigt, sind die meisten Cluster eher klein und enthalten in der Regel 5 Elemente oder weniger. Um das Modell kompakt zu halten, werden alle Prototypen verworfen, deren Cluster nur 1 Element enthalten und daher seltene Muster repräsentieren. Das resultierende visuelle Alphabet enthält $n_T = 5843$ Prototypen. Wie das Streudiagramm in Abbildung 6.59 zeigt, hängt die Clustergröße auch von der Teilegröße ab. Die größten Cluster bestehen nur aus wenigen, räumlich begrenzten Merkmalen, da diese weniger selektiv sind. Insgesamt enthält das visuelle Alphabet jedoch Teile aller Größen zwischen 10 und 60 Pixeln.

Als Zwischenfazit für die Teilemodellierung wird folgendes festgehalten:

- Als kritische Einflüsse wurden die Abhängigkeit zwischen Ortstoleranz und Schwellwert, die räumliche Nähe von Merkmalen, die Teilegröße, die Objektgröße und die Repräsentativität der Stichprobenelemente identifiziert.
- Um alle diese Einflüsse zu berücksichtigen, wird ein visuelles Alphabet aus typischen Mustern in verschiedenen Größen erzeugt.
- Das visuelle Alphabet ist allgemeingültig, d.h. nicht auf einzelne Stichprobenelemente optimiert.
- Das visuelle Alphabet enthält keine sich wiederholenden und keine überselektiven Teile.
- Wie typisch einzelne Teile für bestimmte Objektansichten sind, wird auf der nächsthöheren Modellebene untersucht.

6.4 Modellierung von Ansichten

Mit Hilfe des neu erzeugten visuellen Alphabets sollen nun Objekte modelliert werden. Dazu wird die Stichprobe in Gruppen einheitlicher Objektansichten unterteilt. Die Teile des visuellen Alphabets, die für eine bestimmte Gruppe von Objektansichten typisch sind, werden jeweils zu einem Ansichtsmodell kombiniert.

Die weitere Vorgehensweise sieht nun wie folgt aus: Zunächst wird die Kombination von Teilen untersucht. Das Ziel ist dabei, Teile mit einer hohen Repräsentativität für bestimmte Ansichten zu finden. Der gewählte Ansatz liefert direkt eine Methode zur Einteilung der Stichprobe in verschiedene Ansichten. Im Gegensatz zu den allgemeingültigen Modellen der Teile sollten Ansichtsmodelle die Treffer auf positiven Stichprobenbildern maximieren und Treffer auf Hintergrundbildern minimieren. Diese Eigenschaft wird im Anschluß an die Erzeugung von Ansichtsmodellen untersucht. Dabei werden Abhängigkeiten zwischen Modellparametern und der Einteilung der Stichprobe in Objektansichten identifiziert. Die auf Basis dieser Abhängigkeiten parametrisierten Ansichtsmodelle definieren wieder ein visuelles Alphabet, allerdings auf einer höheren Abstraktionsstufe als das Teile-Alphabet. Das abstraktere Alphabet wird im Anschluß genutzt, um Kategorieknoten auf der nächsthöheren Hierarchieebene des Modells zu erzeugen.

6.4.1 Ermittlung typischer Teile und typischer Ansichten

Die Erzeugung von Ansichtsmodellen findet in einem komplexen Raum von Modellkonfigurationen statt: Zum einen gibt es viele Möglichkeiten, die Stichprobe in verschiedene Ansichten zu unterteilen, zum anderen sind die Elemente des visuellen Alphabets für diese Ansichten unterschiedlich repräsentativ, wodurch sich viele Möglichkeiten der Kombination von Teilen ergeben. Um die wichtigsten Kriterien zu ermitteln, die bei der Erzeugung guter Modelle berücksichtigt werden müssen, wird vereinfacht angenommen, daß ein Teil dann repräsentativ für eine Gruppe von Stichprobenelementen ist, wenn es in allen Bildern der Gruppe erkannt wird. Die Hauptauswirkung dieser Vereinfachung ist, daß selten auftretende Teile bei der Ansichtsmodellierung nicht berücksichtigt werden. Dies stellt sicher, daß die Teile, die in Ansichtsmodelle aufgenommen werden, gut über die Stichprobe generalisieren. Es werden jedoch auch Teile verworfen, die zwar nicht in allen Stichprobenelementen, aber in der Mehrheit der Bilder auftreten und innerhalb der durch die Modellparameter gesetzten Grenzen immer noch zu einer Erkennung führen würden. Die Menge der Teile, die zu Ansichtsmodellen kombiniert werden können, ist daher vielleicht geringfügig kleiner als für optimale Ergebnisse möglich wäre. In der Praxis liegt die Teilezahl für Ansichtsmodelle bei mehreren Tausend bis mehreren Zehntausend. Damit liegt die Teilezahl an der Grenze dessen, was der Versuchsrechner (Intel E6850, 3GHz, 8GB RAM) überhaupt verarbeiten kann, was den möglichen Nachteil relativiert. Wie im folgenden noch dargestellt wird, hängt die Repräsentativität der Ansichtsmodelle im Übrigen stärker von den Wahrscheinlichkeiten ab, mit

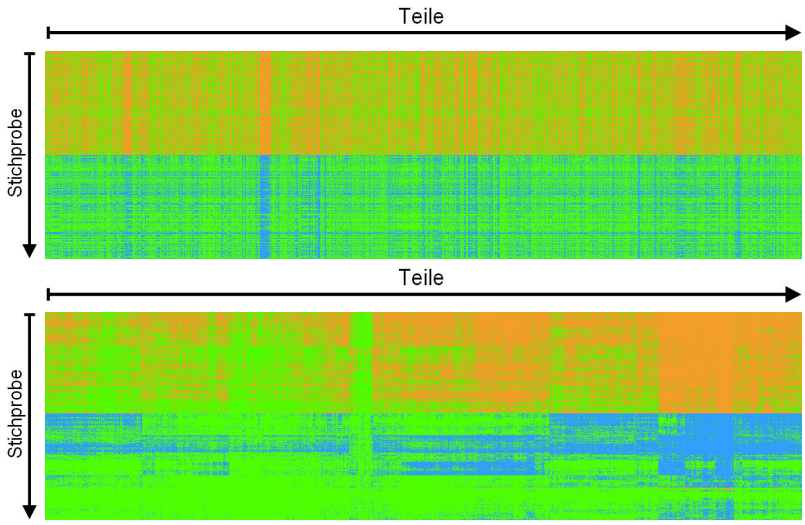


Abbildung 6.60: **Teilematrix.** Das obere Diagramm zeigt die Übereinstimmungen zwischen der Stichprobe und dem visuellen Teile-Alphabet als *Teilematrix*. Übereinstimmungen mit Vordergrundbildern sind orange markiert, Treffer auf Hintergrundbildern dagegen blau. Grüne Bereiche zeigen an, daß Teile nicht detektiert wurden. Das untere Diagramm zeigt eine umsortierte Variante der Teilematrix. Hier wurden aus Gründen der Übersichtlichkeit die Zeilen und Spalten unabhängig voneinander nach ihrer Ähnlichkeit gruppiert.

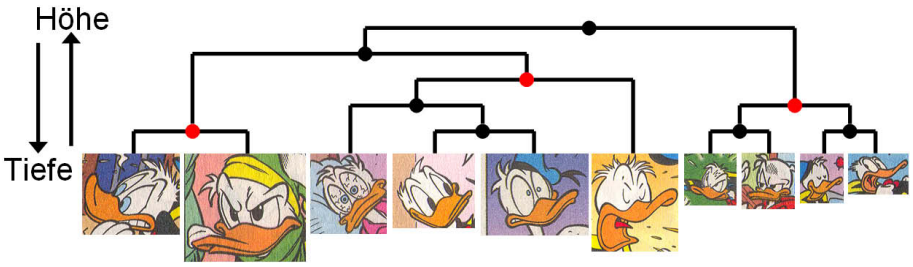


Abbildung 6.61: **Dendrogramm über Stichprobenbilder.** Die Ähnlichkeit von Stichprobenbildern wird anhand der Teile berechnet, die in den Bildern erkannt werden. Die Clusterung der Stichprobe nach den erkannten Bildern ergibt ein Dendrogramm. Wie die Illustration zeigt, entsprechen einzelne Knoten des Dendrogramms bestimmten Gruppen von Bildern. Für ausgewählte Knoten werden Ansichtsmodelle erzeugt. In diesem Beispiel werden die rot markierten Knoten ausgewählt, um die komplette Stichprobe zu modellieren.

denen Teile in Vorder- und Hintergrundbildern detektiert werden, als von der reinen Teile-Anzahl im Modell.

Um Teile zu finden, die zur Modellierung einer Objektansicht in Frage kommen, wird eine *Teilematrix* I_T berechnet. Diese gibt in der Form

$$I_T = \begin{bmatrix} \iota_{1,1} & & \iota_{n_T,1} \\ & \ddots & \\ \iota_{1,n_N+n_P} & & \iota_{n_T,n_N+n_P} \end{bmatrix},$$

an, ob ein Teil mit dem Index i in einem Stichprobenelement mit dem Index j erkannt wird. Das Matrizenelement $\iota_{i,j}$ nimmt dabei den Wert 1 an, falls ein Teil in einem Positivbeispiel der Stichprobe erkannt wird. Der Wert -1 zeigt eine Übereinstimmung mit einem Negativbeispiel an. Ansonsten ist der Wert Null. Die Variablen $n_T = 5843, n_N = 800$ und $n_P = 800$ bezeichnen die Anzahl der Teile, die Anzahl der Positivbeispiele in der Stichprobe und die Anzahl der Hintergrundbilder. Die berechnete Teilematrix zeigt das obere Diagramm in Abbildung 6.60.

Bezüglich der Objekterkennung fällt positiv auf, daß die Teile des visuellen Alphabets in der Stichprobe der Positivbeispiele häufiger gefunden werden als in den Hintergrundbildern. Im Durchschnitt wird ein Teil in 468 Vordergrundbildern, aber nur in 257 Hintergrundbildern erkannt. Dabei ist zu berücksichtigen, daß die Teile ausschließlich positivistisch, d.h. nur aus den positiven Elementen der Trainingsstichprobe erzeugt wurden.

Werden nun Teile zu Ansichtsmodellen kombiniert, läßt sich die Qualität der Objekterkennung bis zu einem gewissen Grad auch aus der Teilematrix ablesen. Eine Kombination mit Hilfe eines hohen Schwellwertes an dem übergeordneten Knoten entspricht einer logischen Und-Verknüpfung auf ausgewählten Spalten der Teilematrix. Für die Positivbeispiele der Stichprobe läuft dies auf eine Anhäufung von Treffern hinaus. Dies resultiert daraus, daß die Stichprobenelemente, die eine Ansicht definieren, ähnlich sind und daher auch ähnliche Teile enthalten. Zudem werden überhaupt nur die Teile in ein Ansichtsmodell aufgenommen, die in allen Bildern zu einer Ansicht auftreten. Da zwischen den Hintergrundbildern keine Abhängigkeiten bestehen, sollte hier keine Häufung auftreten. Dies wird durch die in Abbildung 6.60 dargestellte Teilematrix bestätigt. Aufgrund des oberen Diagramms entsteht hauptsächlich der Eindruck, daß sich die Häufigkeit von Treffern in den Vordergrundbildern in den Negativbeispielen widerspiegelt. Die untere Darstellung, bei der Zeilen und Spalten gemäß ihrer Ähnlichkeit umsortiert wurden, zeigt jedoch über Spalten mit ähnlicher Trefferhäufigkeit eine deutliche Blockstruktur in den Negativbeispielen. Dies läßt darauf schließen, daß sich Teile tatsächlich so kombinieren lassen, daß sich gezielt Treffer für bestimmte Gruppen von Vordergrundbildern anhäufen lassen, wohingegen Treffer auf Hintergrundbildern zufällig verteilt sind. Der Grad der Anhäufung hängt jedoch nicht nur davon ab, ob ein Teil in einem Stichprobenbild erkannt wird, sondern auch, an welchen Bildpositionen es gefunden wird. Dazu kommt, daß ein Ansichtsknoten im Modell wieder eine gewisse

Ortstoleranz zuläßt, d.h. die erkannten Bildpositionen müssen nicht vollständig übereinstimmen. Der Punkt wird später noch genauer untersucht.

Zur Modellierung durch ein gemeinsames Ansichtsmodell werden jeweils Gruppen von Stichprobenelementen ausgewählt, die sich bezüglich der detektierten Teile möglichst stark ähneln. Auf diese Weise ergeben sich die größten Schnittmengen von Teilen und damit die präzisesten Modelle. Mögliche Gruppen von Bildern werden durch den dritten Clusterungsschritt des Trainings identifiziert. Dazu wird wieder das bereits beschriebene Verfahren auf die Spaltenvektoren der Teilematrix angewandt. Das Ergebnis liegt wieder als Dendrogramm vor, wie Abbildung 6.61 illustriert. Das Dendrogramm gibt die Ähnlichkeit von Stichprobenelementen in einer Baumstruktur an, wobei die Blätter den einzelnen Stichprobenelementen und die inneren Knoten bestimmten Gruppen von Bildern entsprechen. Dabei nimmt die Ähnlichkeit der Bilder innerhalb der Gruppen mit zunehmender Tiefe im Baum zu. Zu jedem Knoten des Stichprobendendrogramms kann nun ein Ansichtsknoten im Modell erzeugt werden, der eine bestimmte Gruppe von Stichprobenbildern repräsentiert. Um die gesamte Stichprobe zu modellieren, muß zu jedem Pfad von der Wurzel bis zu einem Blatt mindestens ein Ansichtsmodell erzeugt werden. Eine mögliche Einteilung der Stichprobe in verschiedene Ansichten zeigen die rot markierten Knoten in der Abbildung.

Für das Training des Modells auf der Ansichtsebene ergeben sich daher die folgenden Detailfragen:

- An welchen Bildpositionen werden Teile erkannt?
- In welcher Form überlagern sich die Trefferpositionen in Hintergrundbildern?
- Wovon hängt eine gute Parametrisierung der Ansichtsmodelle ab?
- Welche Knoten des Stichprobendendrogramms sollen modelliert werden?

Diese Fragen werden im Folgenden untersucht.

6.4.2 Vordergrunderkennung durch Teile des visuellen Alphabets

Um zu untersuchen, wie Teile kombiniert werden können, wird für jedes Teil berechnet, an welchen Positionen es in den Bildern der Stichprobe erkannt wird. Das Ergebnis wird in Form von *Trefferbildern* gespeichert. Ein Trefferbild repräsentiert alle detektierten Bildpositionen in Form eines Binärbildes.

Da jedes Teil einem bestimmten Ausschnitt eines Donald-Kopfes entspricht, werden Teile möglicherweise bevorzugt an den Positionen der Stichprobenelemente erkannt, die dem Ausschnitt im jeweiligen Ursprungsbild entsprechen. In diesem Fall kann ein Ansichtsmodell mittels dieser Vorzugspositionen erzeugt werden. Dazu wird in jede Vorzugsposition ein Teileknoten gelegt. Die Teileknoten werden dann über einen gemeinsamen Ansichtsknoten zusammengefaßt.

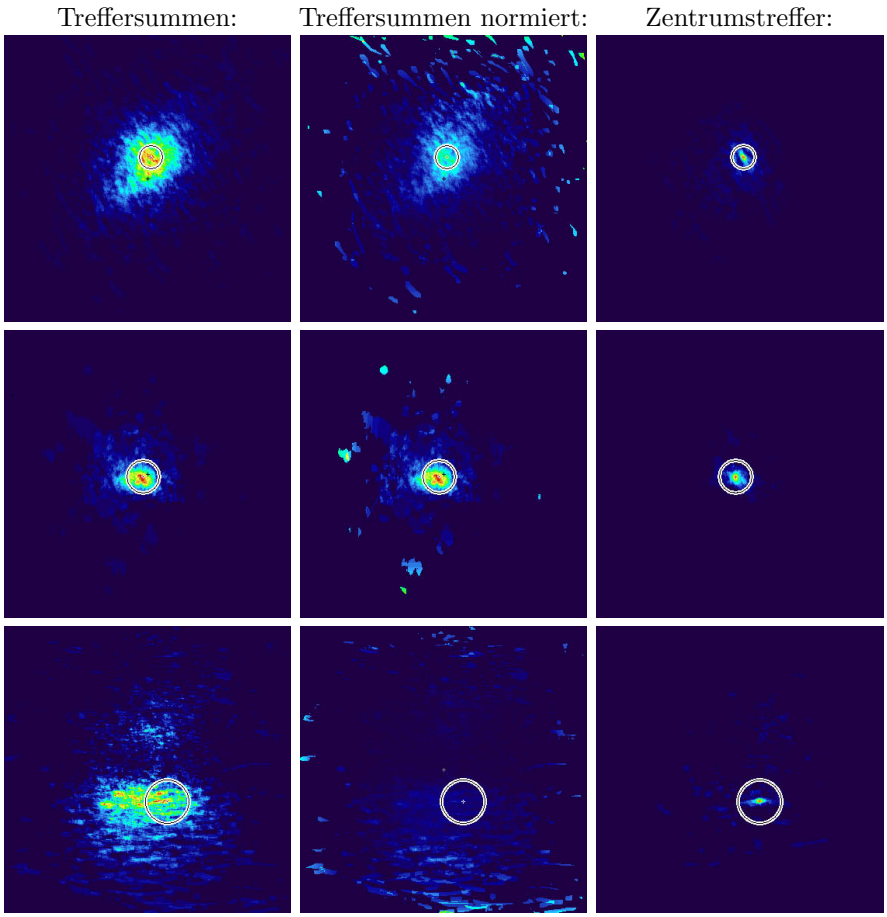


Abbildung 6.62: **Summen von Treffern durch Teile über der Bildebene.** Die Reihen zeigen aufsummierte Trefferbilder von drei verschiedenen Teilen. Die Spalten zeigen verschiedene Darstellungen der Summenbilder. Die erste Spalte zeigt die absoluten Treffersummen. Das Maximum ist durch einen Kreis eingerahmt. Der Radius des Kreises entspricht dem Median der minimalen Abstände zwischen den Treffern und der Position des Maximums für ein Teil und alle positiven Stichprobenelemente. Das bedeutet, daß die Hälfte aller Trefferbilder Treffer innerhalb des Kreises enthält. Ein dunkles Pluszeichen markiert die Mitte der Donaldköpfe. Die zweite Spalte zeigt die Treffersummen der ersten Spalte dividiert durch die Häufigkeit der Treffer aller Teile. Da diese in der Objektmitte maximal ist, ist das Maximum der Summe weniger stark ausgeprägt. Meistens liegt das Maximum an der gleichen Position wie in der nicht normierten Darstellung. Die dritte Spalte gibt die Summe aller Trefferbilder an, die einen Treffer an der Position des Maximums enthalten.

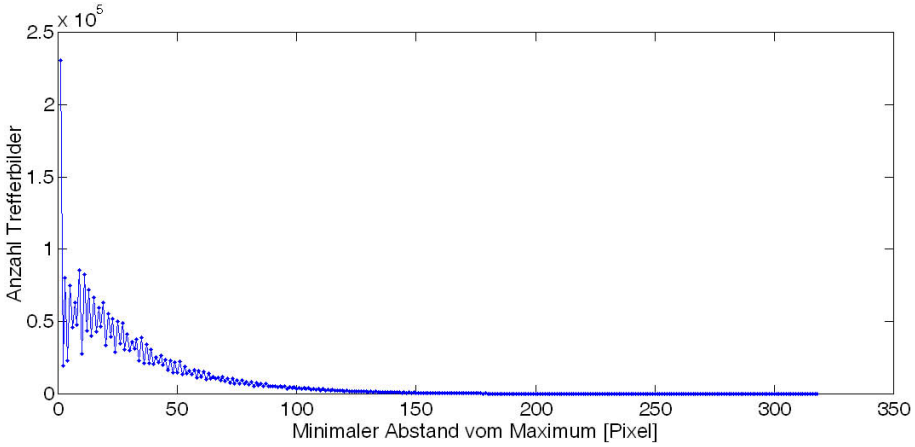


Abbildung 6.63: **Histogramm über den minimalen Abstand zwischen Treffern und einer möglichen Vorzugsposition von Teilen.** Für den Abstand Null ergibt sich ein deutlich abgegrenztes Maximum. Die übrigen Abstände treten seltener auf, jedoch ohne weitere grobe Abstufungen.

Die Ortstoleranz des Ansichtsknotens richtet sich dann nach der Streuung der Teilepositionen um die Vorzugsposition.

Um die Existenz von Vorzugspositionen zu prüfen, werden nun die Trefferbilder eines Teils über alle positiven Stichprobenelemente summiert. Eine mögliche Vorzugspositionen sollte sich als deutlich Erkennbares Maximum zeigen. Da die Bilder unterschiedlich groß sind, wird für die Summation ein gemeinsamer Bezugspunkt auf die Bildmitte festgelegt. Einen kleinen Teil der Ergebnisse zeigen die Diagramme der linken Spalte in Abbildung 6.62. Normalerweise ergibt sich ein deutliches Maximum in der Nähe des Objektzentrums. Die Position des Maximums ist dabei für verschiedene Teile unterschiedlich. Da die Objekte jedoch unterschiedlich groß sind, ist in der Bildmitte generell die absolute Trefferhäufigkeit höher. Die mittlere Spalte der Abbildung zeigt daher eine auf die generelle Trefferhäufigkeit normierte Darstellung. In dieser Darstellung bleibt das Maximum meistens erhalten. Die Höhe des Maximum, d.h. die Anzahl der Teile mit Treffern im Maximum, liegt deutlich unter der Stichprobengröße. Es besteht daher die Möglichkeit, daß das Maximum nur einen kleinen Anteil der Stichprobenbilder repräsentiert. Wenn der Anteil zu klein ist, hat es jedoch keinen Zweck, Teileknoten einer Ansicht in die Maxima zu legen, da dann in der Praxis viele Trefferpositionen außerhalb der Ortstoleranz liegen.

Um diesen Punkt zu klären, werden für jedes Teil alle Trefferbilder aufsummiert, die einen Treffer an der Position des Maximum enthalten. In Abbildung 6.62 sind diese in der rechten Spalte angegeben. Es zeigt sich, daß die Häufungen am Maximum nun sehr viel schmaler sind. Um die Streuung der Teile einschätzen zu können, wird für jedes Trefferbild der Treffer mit dem minimalen Abstand vom Maximum berechnet. Der Median der minimalen Abstände

für jeweils ein Teil ist in der Abbildung als weißer Kreis dargestellt. Der Median gibt den Radius an, bei dem die Hälfte aller Trefferbilder eines Teils mindestens einen Treffer innerhalb des Kreises hat. Er ist damit ein Maß für die Streuung der Treffer um das Maximum. Die Summen der Trefferbilder mit Treffern im Maximum bilden im Vergleich zu dem Median nur sehr schmale Häufungen. Dies kann so interpretiert werden, daß ein Treffer, der nicht genau auf dem Maximum liegt, meistens relativ weit entfernt liegt.

Um diese Aussage für alle Teile zusammen zu überprüfen wird ein Histogramm über die minimalen Abstände zwischen den Treffern eines Trefferbildes und der Position des Maximums berechnet. Dieses zeigt Abbildung 6.63. Es fällt auf, daß der Abstand Null, d.h. ein Treffer genau im Maximum, am häufigsten auftritt. Der zweithäufigste Abstand tritt bereits um den Faktor drei seltener auf. Für Abstände über einem Pixel verläuft das Histogramm relativ eben. Der deutliche Unterschied zwischen den Treffern im Maximum und den übrigen Treffern bestätigt die Beobachtung, daß Treffer abseits des Maximums oft relativ weit entfernt liegen. Das bedeutet, daß die Vorzugsposition, die durch das Maximum angedeutet wird, weniger bedeutsam ist, als anfangs angenommen. Dies zeigt sich darin, daß die Anzahl der Treffer in den Positivbildern der Stichprobe von ca. 2,7 Millionen auf etwa 230 000 sinkt, wenn nur Trefferbilder mit Treffern im Maximum zugelassen werden. Da die meisten Treffer offenbar nicht in der Nähe des Maximums liegen, sondern über die Bildebene verteilt sind, müssen bei der Ansichtsmodellierung alle Trefferpositionen berücksichtigt werden und nicht nur die vermuteten Vorzugspositionen.

Zur Modellierung von Ansichtsknoten wird daher ein Ansatz gewählt, der auf den morphologischen Operationen Dilatation und Erosion beruht. Dazu werden zunächst anhand der Teilematrix I_T alle Teile bestimmt, die in einer Bildermenge, die eine bestimmte Objektansicht darstellt, erkannt werden. Für eine Beschreibung der Objektansicht durch n_A Bilder und n_R Teile, die in allen Bildern erkannt werden, ergibt sich eine Menge von $n_A n_R$ Trefferbildern

$$I_{M,1,1}, \ldots, I_{M,n_A,n_R}.$$

Die Ortstoleranz ϑ, die für den neu zu erzeugenden Ansichtsknoten festgelegt wird, läßt sich durch eine Dilatation auf den Trefferbildern simulieren. Ein Trefferbild

$$I_M = \begin{bmatrix} \iota_{M,1,1} & & \iota_{M,\ldots,1} \\ & \ddots & \\ \iota_{M,1,\ldots} & & \iota_{M,\ldots,\ldots} \end{bmatrix}$$

wird so über die Formel

$$\iota_{D,\boldsymbol{x}} = \begin{cases} 1, & \text{falls} \sum\limits_{|\boldsymbol{x}-\boldsymbol{y}|<\vartheta} \iota_{M,\boldsymbol{y}} > 0 \\ 0, & \text{sonst} \end{cases} \tag{6.18}$$

in das dilatierte Trefferbild

$$I_D = \begin{bmatrix} \iota_{D,1,1} & & \iota_{D,\ldots,1} \\ & \ddots & \\ \iota_{D,1,\ldots} & & \iota_{D,\ldots,\ldots} \end{bmatrix}$$

überführt, wobei $\boldsymbol{x}$ und $\boldsymbol{y}$ Positionen in I_D und I_M sind. Als nächstes werden für jedes Teil die gemeinsamen Trefferpositionen über alle n_A Bilder der Objektansicht bestimmt. Dazu werden für jedes Teil alle entsprechenden dilatierten Trefferbilder aufsummiert. Für jedes Teil j entsteht so ein Summenbild

$$I_{S,j} = \begin{bmatrix} \iota_{S,1,1} & & \iota_{S,\ldots,1} \\ & \ddots & \\ \iota_{S,1,\ldots} & & \iota_{S,\ldots,\ldots} \end{bmatrix},$$

wobei

$$\iota_{S,\boldsymbol{x}} = \sum_{i=1}^{n_A} \iota_{D_{i,j},\boldsymbol{x}}.$$

Da die Trefferbilder I_D Binärbilder sind, nehmen die Matrizenelemente der Summenbilder I_S maximal den Wert n_A an, d.h. die Anzahl der Bilder der Ansicht. Um die gemeinsamen Trefferpositionen eines Teils zu finden, werden die Positionen in I_S bestimmt, an denen der Maximalwert auftritt. Dies entspricht einer Schwellwertoperation oder Binarisierung. Das Ergebnis ist ein Schnittbild

$$I_{\cap,j} = \begin{bmatrix} \iota_{\cap,1,1} & & \iota_{\cap,\ldots,1} \\ & \ddots & \\ \iota_{\cap,1,\ldots} & & \iota_{\cap,\ldots,\ldots} \end{bmatrix} \tag{6.19}$$

mit den Werten

$$\iota_{\cap,\boldsymbol{x}} = \begin{cases} 1, & \text{falls } \iota_{S,\boldsymbol{x}} = n_A \\ 0, & \text{sonst} \end{cases}.$$

Aufgrund der Dilatation in Gleichung 6.18 enthalten die Schnittbilder $I_\cap$ an den Rändern der Trefferbereiche Positionen, die in den ursprünglichen Trefferbildern nicht auftreten. Da diese bei der Objekterkennung instabil sind, muß die Dilatation durch eine analoge Erosion auf dem Schnittbild kompensiert werden. Es werden daher erodierte Schnittbilder

$$I_{E,j} = \begin{bmatrix} \iota_{E,1,1} & & \iota_{E,\ldots,1} \\ & \ddots & \\ \iota_{E,1,\ldots} & & \iota_{E,\ldots,\ldots} \end{bmatrix}$$

berechnet. Die Matrizenelemente nehmen so die Werte

$$\iota_{E,\boldsymbol{x}} = \prod_{|\boldsymbol{x}-\boldsymbol{y}|<\vartheta} \iota_{\cap,\boldsymbol{y}}$$

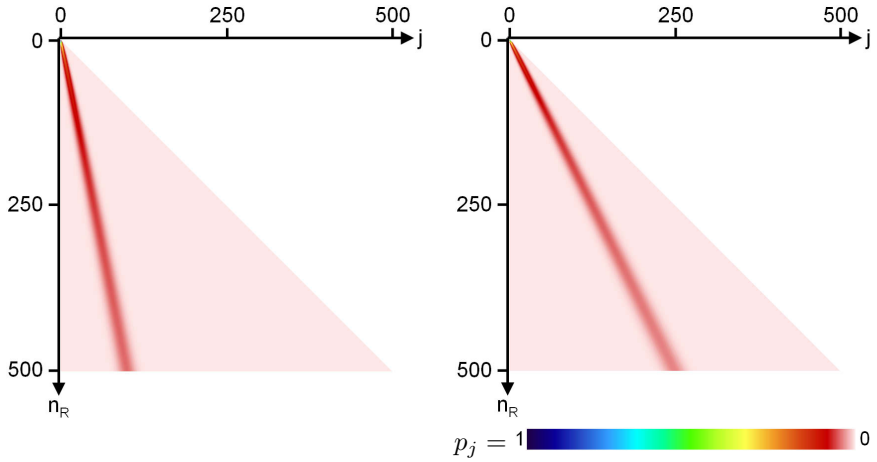


Abbildung 6.64: **Binomialverteilungen.** Die Diagramme stellen Gleichung 6.20 für verschiedene Parametrisierungen dar. Links: $p = 0{,}2$, rechts: $p = 0{,}5$. Da p_j mit steigendem n_R schnell fällt, wird über den größten Teil des Diagramms nur ein kleiner Ausschnitt der Farblegende genutzt.

an. Nun wird ein neuer Ansichtsknoten erzeugt. Für alle mit dem Wert 1 besetzten Positionen in den erodierten Trefferbildern können nun die entsprechenden Teileknoten als Unterknoten in den Ansichtsknoten eingefügt werden. Auf diese Art entsteht jedoch ein unnötig komplexes Modell. Da ein Unterknoten immer dann erkannt wird, wenn ein gleiches Teil innerhalb der Ortstoleranz erkannt wird, kann ein kompakteres Modell berechnet werden, indem die erodierten Schnittbilder unterabgetastet werden. Das Abtastverhältnis richtet sich nach der Ortstoleranz. Da die Schnittbilder Strukturen enthalten, die feiner als die Ortstoleranz sind, wird nicht mit einem regelmäßigen Raster abgetastet. Bei einem regelmäßigen Raster bestünde die Gefahr, daß die Trefferpositionen so häufig auf Zwischenrasterplätze fiele, daß das Modell die Trefferanordnung im erodierten Schnittbild nicht mehr repräsentieren würde. Stattdessen beginnt die Abtastung immer mit einer positiven Trefferposition im erodierten Schnittbild. Die benachbarten Positionen unterhalb der Ortstoleranz werden dann übersprungen. Abbildung 6.71 zeigt ein Beispiel.

6.4.3 Hintergrundunterdrückung durch Ansichtsmodelle

Aus der Forderung, daß ein Teil in allen Bildern einer Objektansicht vertreten sein muß, um es in das Ansichtsmodell aufzunehmen, ergibt sich, daß der Schwellwert theoretisch auf 100% gesetzt werden darf, ohne daß sich während der Objekterkennung schlechtere Ergebnisse für die trainierte Stichprobenelemente ergeben. Aufgrund von Rauschen ist jedoch ein niedrigerer Schwellwert wünschenswert, da dies die Verallgemeinerbarkeit des Modells erhöht, was die Toleranz des Modells gegen Abweichungen von der Trainingsstichpro-

be verbessert. Die Untergrenze des Schwellwerts wird durch die Anzahl der fälschlicherweise erkannten Hintergrundbilder festgelegt.

Da die Hintergrundbilder beliebige Muster annehmen können, kann über die entsprechenden Trefferbilder bestenfalls angenommen werden, daß einzelne Treffer statistisch unabhängig voneinander auftreten. An einer bestimmten Koordinate im Treffersummenbild I_S können bis zu n_R Treffer durch verschiedene Teile auftreten, wobei 2^{n_R} Kombinationen von Teilen möglich sind. Die Anzahl der Kombinationen mit genau j erkannten Teilen wird durch den Binomialkoeffizienten $\binom{n_R}{j}$ gegeben. Die Anzahl der Trefferkombinationen mit mindestens j Treffern in n_R Teilen ergibt der Term $\sum_{i=j}^{n_R}\binom{n_R}{i}$ durch Summation über alle großen Trefferanhäufungen. Dies berücksichtigt jedoch noch nicht die Wahrscheinlichkeit, mit der überhaupt Teile detektiert werden. Wenn angenommen wird, daß alle Teile mit der gleichen Wahrscheinlichkeit p erkannt bzw. mit der Wahrscheinlichkeit $1-p$ nicht erkannt werden, ergibt sich die Wahrscheinlichkeit

$$p_j = \binom{n_R}{j} p^j (1-p)^{n_R-j} \tag{6.20}$$

für genau j Treffer (siehe Abbildung 6.64) und die Wahrscheinlichkeit

$$p_{j\leq} = \sum_{i=j}^{n_R} \binom{n_R}{i} p^i (1-p)^{n_R-i} \tag{6.21}$$

für mindestens j Treffer an der gleichen Position. Wie man aus Abbildung 6.64 ableiten kann, verhält sich $p_{j\leq}$ wie eine verschliffene Sprungfunktion, die für kleine Werte von j nahe Eins und für große Werte von j nahe Null liegt. Die Position des Übergangs ergibt sich aus der Position des Maximum in Gleichung 6.20. Wie Abbildung 6.64 zeigt, hängt dieses von der Wahrscheinlichkeit p der Teiledetektion ab, nicht jedoch von der Anzahl an Teilen n_R. Die Teileanzahl hat dagegen einen schwachen Einfluß auf die Steilheit des Übergangs von Eins nach Null.

Ignoriert man für einen Augenblick die numerischen Probleme bei der Berechnung der Binomialverteilung in Gleichung 6.20, dann kann über die Gleichung 6.21 eine Anzahl von Teilen $\vartheta = j$ bestimmt werden, für die sich eine ausreichend niedrige Trefferwahrscheinlichkeit ergibt. Aufgrund des steilen Abfalls der Binomialverteilung abseits des Maximums ergeben sich schnell sehr niedrige Werte.

Diese Methode ist in der Praxis aufgrund von Eigenschaften des Erkennungsverfahren und Toleranzen beim Mustervergleich eingeschränkt. Zum einen ändert die Dilatation aus Gleichung 6.18 die Trefferwahrscheinlichkeit p. Dies kann teilweise dadurch kompensiert werden, daß p erst nach der Dilatation berechnet wird. Allerdings führt die Dilatation auch zu geometrischen Abhängigkeiten zwischen den Treffern, was die Annahme der statistischen Unabhängigkeit unterläuft. Die statistische Unabhängigkeit wird außerdem durch die Existenz von Ober- und Untermengen innerhalb der modellierten Muster beeinträchtigt. Diese Informationen wurden zwar aufgrund der logischen UND-Operation aus der Clusterung der Teilekandidaten, nicht jedoch aus den Teilen

Versuch	Trainierte Elemente	ς_{Min} [Pixel]	ς_{Max} [Pixel]	Schrittweite [Pixel]	Objektgröße μ, σ	
A	4	21	101	8	181	50
B	4	21	101	8	104	50
C	2	21	101	8	200	143
D	4	21	101	8	123	56
E	4	21	101	8	218	119
F	64	75	231	14–18	151	57
G	8	21	249	8–30	193	99
H	16	39	165	12–24	172	68
I	32	29	219	10–28	158	65
J	102	41	201	20–44	153	67
K	8	21	81	10	128	34

Tabelle 6.10: Versuchsparameter für das Training von Ansichten. In jedem Versuch wurde ein optimales Ansichtsmodell für einen bestimmten Knoten im Stichprobendendrogramm erzeugt. Solche Knoten repräsentieren jeweils eine bestimmte Menge an Stichprobenelementen (Spalte 2). Zur Optimierung wurde die Ortstoleranz ς systematisch variiert (Spalten 3–5). Um den Versuchsaufwand zu reduzieren, wurden für höhere Ortstoleranzen größere Schrittweite gewählt. Der Schwellwert wurde jeweils auf 90% gesetzt. Die rechten zwei Spalten geben die mittlere Größe und die Standardabweichung der Größe der jeweils trainierten Stichprobenelemente an.

selbst ausgeschlossen. Dies führt dazu, daß Ober- und Untermuster teilweise an den gleichen Positionen erkannt werden. Außerdem können weitere, bisher unidentifizierte Effekte auftreten.

In der Praxis wirkt sich dies in einer vergrößerten Standardabweichung der Binomialverteilung in Gleichung 6.20 aus. Aufgrund des bisher ungeklärten Zusammenhangs der Geometrie auf die Trefferwahrscheinlichkeiten ist allerdings auch keine geschlossene Berechnung des Schwellwerts ϑ möglich. Aus diesem Grund wird der Schwellwert experimentell eingestellt, indem die Anzahl der erkannten positiven und negativen Beispiele der Stichprobe ausgewertet wird.

6.4.4 Kritische Variablen für die Ansichtsparametrisierung

Als nächstes wird experimentell untersucht, zu welchen Knoten des Dendrogramms über die Stichprobe (vgl. Abb. 6.61) Ansichtsmodelle erzeugt werden sollen, und wie diese optimal parametrisiert werden. Dazu werden probeweise verschieden parametrisierte Ansichtsmodelle erzeugt und anhand einer Stichprobe getestet.

Zunächst müssen Knoten des Stichprobendendrogramms ausgewählt werden. Hier steht jeder Knoten für die Gruppe von Stichprobenelementen an den

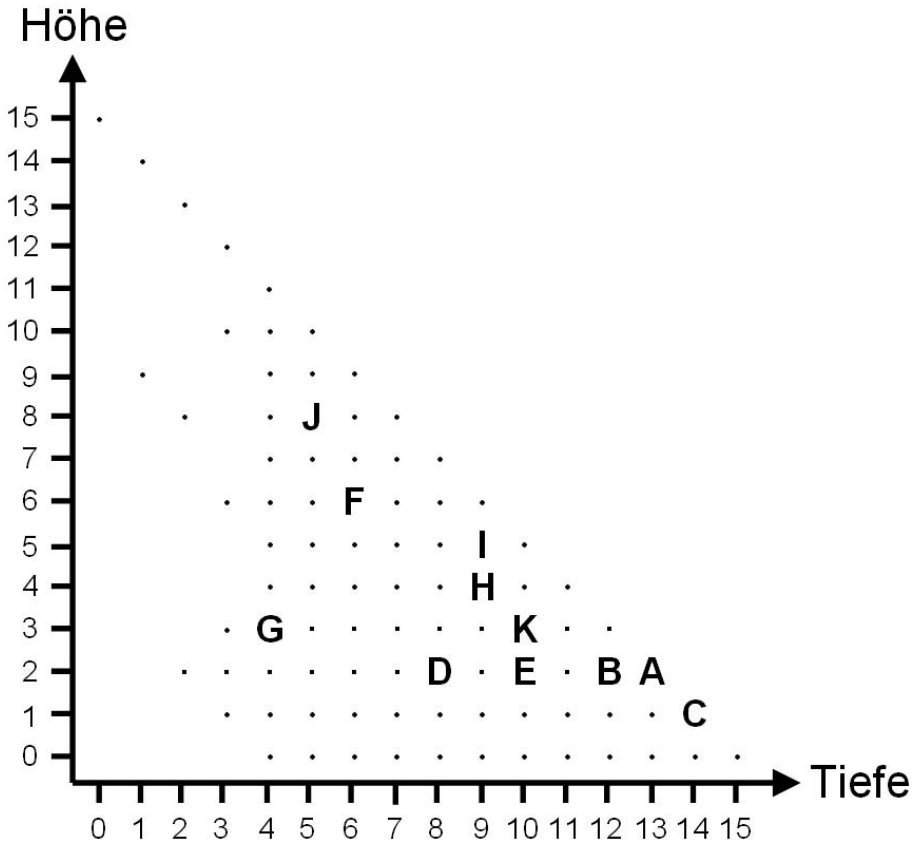


Abbildung 6.65: **Höhe und Tiefe von Knoten im Stichprobendendrogram.** Schwarze Punkte markieren die Höhe und Tiefe aller im Stichprobendendrogramm auftretender Knoten. Die Wurzel des Dendrogramms liegt in diesem Diagramm mit einer Tiefe von Null und einer Höhe von 15 oben links. Die Blattknoten des Dendrogramms besitzen eine Höhe von Null und werden daher durch die unterste Reihe von Punkten markiert. Die Höhe und Tiefe der in den Messungen A–K untersuchten Knoten ist durch die entsprechenden Buchstaben gekennzeichnet.

Versuch	Beste Parametrisierung		Max.	Anzahl
	ς [Pixel]	ϑ	Precision	Teile
A	53	95%	1	24 354
B	29	85%	1	3038
C	21	81%	1	4274
D	29	91%	1	1807
E	21–29	85%	1	1142–6244
F	195	100%	1	64 248
G	119	100%	0,087	2744
H	99	98%	1	19 992
I	191	100%	1	149 930
J	201	100%	1	222 279
K	51	98%	1	7592

Tabelle 6.11: In jedem Versuch (A-K) wurde das Ansichtsmodell ermittelt, das die zu dem betreffenden Knoten des Stichprobendendrogramms zusammengefaßten Donald-Bilder mit einem optimalen positiven Vorhersagewert erkennt. Die Tabelle zeigt zu jedem Versuch die optimale Ortstoleranz und den optimalen Schwellwert, sowie den erzielten positiven Vorhersagewert. Der Wertebereich des positiven Vorhersagewerts reicht von 0 bis 1, wobei der Wert 1 das Optimum darstellt. Die letzte Spalte gibt an, wieviele Teileknoten zur Modellierung des besten Ansichtsmodells erzeugt wurden. Die Anzahl der Teileknoten kann die Größe des visuellen Teilealphabets übersteigen, da gleiche Teile für unterschiedliche Bildpositionen erzeugt werden können.

Versuch	Beste Parametrisierung		Max.	Anzahl
	ς [Pixel]	ϑ	Accuracy	Teile
A	101	64%	0,801	147 020
B	85	66%	0,805	127 816
C	85	54%	0,783	204 668
D	45	59%	0,765	13 076
E	101	55%	0,782	253 513
F	117	99%	0,776	4764
G	51	99%	0,824	98
H	119	76%	0,704	48 132
I	99	97%	0,805	16 720
J	87	94%	0,826	12 153
K	81	75%	0,819	47 782

Tabelle 6.12: Parametrisierungen von Ansichtsmodelle mit optimaler Korrektklassifikationsrate.

Blattknoten des jeweiligen Teilbaums. Die Größe der Gruppe hängt von der Höhe und der Tiefe der ausgewählten Knoten im Dendrogramm ab. Die Höhe gibt dabei die Anzahl an Hierarchiestufen zwischen dem ausgewählten und den Blattknoten an. Die Tiefe gibt den Abstand zum Wurzelknoten des Dendrogramms an. Da höhere Knoten mehr Stichprobenelemente umfassen, sind die zugehörigen Ansichtsmodelle aufgrund des Schnitts über alle Teiletreffer allgemeiner. Knoten mit einer geringeren Höhe entsprechen dagegen gleichartigeren Gruppen von Stichprobenelementen, wodurch selektivere Ansichtsmodelle zu erwarten sind. Bei der Wahl der Knoten im Dendrogramm kann über die Höhe ein Kompromiß zwischen Selektivität und Verallgemeinerbarkeit getroffen werden.

Neben der Höhe eines Knotens spielt auch die Knotentiefe eine Rolle. Da diese entgegengesetzt zur Höhe definiert ist, sind hier weitgehend gegensätzliche Ergebnisse zu erwarten. Das Stichprobendendrogramm ist jedoch im allgemeinen nicht ausbalanciert, d.h. die Blattknoten des Dendrogramms haben unterschiedliche Tiefen. Daraus ergeben sich für Knoten einer bestimmten Höhe unterschiedliche Tiefen und umgekehrt. Aus diesem Grund müssen Höhe und Tiefe separat betrachtet werden. Abbildung 6.65 zeigt alle Kombinationen einer bestimmten Höhe und einer bestimmten Tiefe, die im Stichprobendendrogramm auftreten.

Um festzustellen, ob die Güte der Ansichtsmodelle eher von der Tiefe oder der Höhe abhängt, werden für elf verschiedene Knoten Ansichtsmodelle erzeugt. In Abbildung 6.65 sind die untersuchten Kombinationen aus Knotentiefe und -höhe mit den Buchstaben A bis K markiert. Für die Wurzelknoten der Ansichtsmodelle muß eine Schwelle ϑ und eine Ortstoleranz ς gewählt werden. Dabei beeinflußt die Ortstoleranz auch die Schnittmenge der in allen Beispielobjekten einer gewählten Ansicht auftretenden Teile. Da unklar ist, wie die Parameter ϑ und ς von der Höhe und der Tiefe im Stichprobendendrogramm abhängen, werden für jedes Ansichtsmodell mehrere Parametrisierungen überprüft. Diese sind in Tabelle 6.10 angegeben.

Trainingsziele auf Ansichtsebene

Zur Auswertung der Ansichtsmodelle wird die trainierte Stichprobe reklassifiziert. Aufgrund der Arbeitsweise des vorliegenden Modellierungsverfahrens ist zunächst zu erwarten, daß alle trainierten Elemente erkannt werden. Interessanter ist jedoch, wie die Anzahl der erkannten Vordergrund- und Hintergrundbilder von der Schwelle ϑ abhängt. Diese Abhängigkeit wird durch Auszählen der Treffer ermittelt. Auf Basis dieser Zahlen wird die Parametrisierung der Ansichtsmodelle verfeinert. Dies ist sinnvoll, da bei den Ansichtsmodellen eine relativ grobe Schrittweite für die Ortstoleranz gewählt wurde, um den rechentechnischen Aufwand auf ein praktikables Maß zu beschränken. Aufgrund der groben Schrittweite können Optima auf Zwischenschritten nicht erkannt werden. Um dieses Problem zu lösen, wird die Abhängigkeit zwischen Schwelle und Ortstoleranz ausgenutzt und die Feinoptimierung über die Schwelle vorgenommen. Dies geschieht recheneffizient über die Optimierung bezüglich der bereits durch das Auszählen bekannten Trefferzahlen für den Vordergrund und den Hintergrund.

Die Schwelle wird zudem genutzt, um eine Optimierung des Modells auf verschiedene Ziele zu erreichen. Diese hängen vom letztendlichen Einsatzzweck der Objekterkennung ab. Da in dieser Arbeit das Problem im Allgemeinen betrachtet wird, werden stellvertretend drei relevante Optimierungsziele angestrebt. Diese sind

1. ein hoher positiver Vorhersagewert (engl. *Precision*),
2. eine hohe Korrektklassifikationsrate (engl. *Accurracy*) und
3. eine hohe Sensitivität (engl. *Recall*).

Ein hoher positiver Vorhersagewert bedeutet, daß das Modell sehr selektiv ist. Das Objekterkennungssystem erkennt dann zwar nicht alle Objekte, allerdings sind die wenigen Erkennungen, die angezeigt werden, dann auch sehr zuverlässig. Um den hohen positiven Vorhersagewert zu erreichen, wird für jedes Ansichtsmodell die Schwelle ϑ so eingestellt, daß das Verhältnis in Gleichung 1.1 den maximalen Wert ergibt. Die Anzahl der richtig Positiven der Gleichung ist hier die Anzahl der erkannten Stichprobenelemente aus der trainierten Ansicht. Die Anzahl der falschen Positiven bezieht sich auf Treffer in den Hintergrundbildern.

Um eine hohe Korrektklassifikationsrate zu erzielen, wird das Verhältnis der korrekten positiven und negativen Erkennungen zur Gesamtzahl aller Treffer optimiert (Gleichung 1.2). Erste Tests zeigen Vorteile für die Berechnung der richtig positiven Treffer eines Ansichtsmodells jeweils über die gesamte Stichprobe, statt nur über die Stichprobenelemente der gerade betrachteten Ansicht. Möglicherweise wirkt sich hier die größere Anzahl an Beispielen positiv aus. Die Optimierung auf prinzipiell nicht trainierte Ansichten erfordert eine höhere Generalisierbarkeit des Modells, was die Gefahr von falschen Positiv-Detektionen mit sich bringt. Eine weitere Modelloptimierung auf der nächsthöheren Kategorieebene kann daher hauptsächlich über die Reduktion der Falsch-Positiv-Rate geschehen.

Als drittes soll die Sensitivität optimiert werden. Dies ist allerdings auf Ansichtsebene kein sinnvolles Kriterium, da die Ansichtsmodelle sich nur aus den Teilen zusammensetzen, die in allen zu einer Ansicht zusammengefaßten Stichprobenelemente vorkommen. Aus diesem Grund ist die Sensitivität für plausible Parameterbereiche bereits bei der Modellerzeugung maximal. Die Optimierung kann daher nur verbessert werden, indem auf der nächsthöheren Ebene der Kategorien falsche Treffer minimiert werden. Damit ist das Verfahren identisch mit dem vorangehenden.

Die Tabellen 6.11 und 6.12 zeigen die durch die Optimierung des positiven Vorhersagewertes bzw. der Korrektklassifikationsrate ermittelten Parametrisierungen. Wenn mehrere Parametrisierungen das gleiche Ergebnis ergaben, wurde die Parametrisierung mit dem niedrigsten Schwellwert gewählt, um die Generalisierbarkeit des Modells zu erhöhen und damit eine zu starke Anpassung an das Rauschen in der Stichprobe zu vermeiden. Waren ferner als Ergebnis mehrere Ortstoleranzen möglich, wurde die kleinste gewählt.

Erster Parameter	**Zweiter Parameter**	**Korrelations-koeffizient**
Objektgröße	Höhe	−0,1749
Objektgröße	Tiefe	0.0597
Standardabweichung der Größe	Höhe	−0,3206
Standardabweichung der Größe	Größe	0,7893
Schwelle	Höhe	0,7264
Schwelle	Tiefe	−0,7020
Ortstoleranz	Höhe	0,9240
Ortstoleranz	Tiefe	−0,6910
Anzahl Teileknoten	Ortstoleranz	0,8117

Tabelle 6.13: Korrelationskoeffizient von Ansichtsparametern bei der Optimierung auf den positiven Vorhersagewert. Der Korrelationskoeffizient ist normiert, d.h. die Werte liegen zwischen 0 (unkorreliert) und ±1 (ideale Korrelation).

Ansichtsmodelierung bei Optimierung auf den positiven Vorhersagewert

Anhand dieser Ergebnisse wird nun nach den Einflußfaktoren gesucht, die für die Erzeugung guter Ansichtsmodelle entscheidend sind.

Als erstes wird der Zusammenhang zwischen der Objekterscheinung und der Position eines Knotens im Stichprobendendrogramm untersucht. Da sich die Objektgröße im Gegensatz zu anderen relevanten visuellen Eigenschaften leicht bestimmen läßt, wird diese beispielhaft für die Erscheinung der Objekte betrachtet. Dazu wird für die Messungen A–K jeweils die mittlere Größe der modellierten Stichprobenelemente und die zugehörige Standardabweichung berechnet (siehe Tabelle 6.10). Die Abbildungen 6.66 a) und b) zeigen Streudiagramme der Größe über der Höhe und Tiefe. Wie an den beiden Diagrammen erkennbar ist, läßt sich nicht von der Objektgröße auf die Höhe oder Tiefe im Stichprobendendrogramm schließen. Tabelle 6.13 zeigt, daß die Größe und die Lage im Stichprobendendrogramm unkorreliert sind. Abbildung 6.66 a) deutet jedoch an, daß die Größe bei einer niedrigen Knotenhöhe, d.h. bei kleinen Clustern, stärker schwankt als bei großen Clustern nahe der Wurzel des Stichprobendendrogramms. Diese Beobachtung wird durch das Diagramm in Abbildung 6.66 c) bestätigt. Dieses zeigt einen gewissen Abfall der Standardabweichung der Größe mit zunehmender Höhe im Dendrogramm. Die Größe und die Höhe sind mit einem Koeffizienten von −0,32 schwach korreliert (Tabelle 6.13). Wie die Abbildungen 6.66 b) und d) zeigen, besteht dieser Zusammenhang zwischen der Größe und der Tiefe jedoch nicht, oder er ist zu schwach ausgeprägt, um mit den Messungen A–K erfaßt zu werden.

Diese Beobachtungen lassen sich dadurch erklären, daß Knoten einer geringen Höhe, d.h. Knoten nahe den Blattknoten des Dendrogramms, einheitlich große Stichprobenelemente umfassen. Auf höheren Ebenen des Dendrogramms werden zunehmend uneinheitliche Gruppen von Stichprobenelementen vereint,

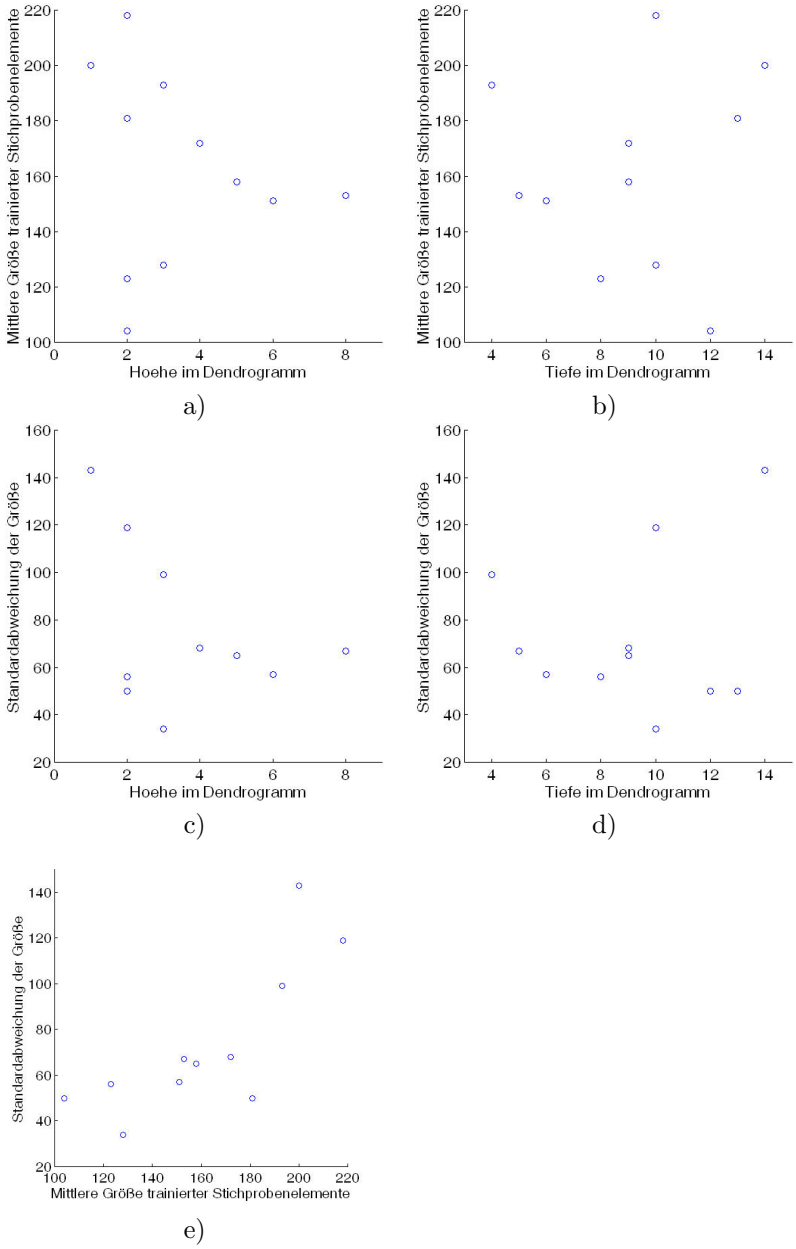


Abbildung 6.66: **Zusammenhang zwischen Objektgröße und Stichprobendendrogramm.** a,b) Die Objektgröße hängt nicht von der Knotenposition im Stichprobendendrogramm ab. c) Die Streuung der Objektgröße innerhalb der zu einer Ansicht zusammengefaßten Stichprobenelemente nimmt dagegen mit der Höhe im Dendrogramm ab. e) Ansonsten hängt die Streuung nur von der mittleren Größe der Stichprobenelemente ab.

sodaß sich die mittlere Größe der Elemente einzelner Knoten zunehmend dem Mittelwert aller Stichprobenelemente nähert. Die steigende Standardabweichung der Größe spiegelt die wachsende Heterogenität der Stichprobenelemente wider. Da die Gesamtheit der Knoten einer bestimmten Höhe im Prinzip immer alle Stichprobenelemente umfaßt, bleibt der Mittelwert über der Höhe konstant. Der schwächer ausgeprägte Zusammenhang zur Tiefe im Dendrogramm kann damit erklärt werden, daß Knoten einer bestimmten Tiefe oft unterschiedlich große Teilbäume im Dendrogramm darstellen. Die Anzahl der Stichprobenelemente und damit auch deren Heterogenität schwankt daher stärker für die Tiefe stärker als für die Höhe. Insgesamt bestätigen diese Beobachtungen die korrekte Funktionsweise der Clusterung von Stichprobenelementen zu mehr oder weniger homogenen Ansichten je nach Höhe im Dendrogramm. Die beobachteten Effekte sind dabei jedoch deutlich schwächer als die Abhängigkeit zwischen der Größe und der Standardabweichung der Größe (siehe Abbildung 6.66 e) und Tabelle 6.13). Die Objektgröße als leicht vermeßbarer Aspekt der visuellen Erscheinung eignet sich daher auch bei diesem Ansatz schlecht zur Erzeugung guter Ansichtsmodelle.

Als nächstes wird untersucht, wie die Parametrisierung und Qualität der auf einen hohen positiven Vorhersagewert optimierten Ansichtsmodelle von der Höhe und Tiefe im Stichprobendendrogramm abhängt. Wie Tabelle 6.11 zeigt, konnte für fast alle getestete Knoten des Stichprobendendrogramms der optimale positive Vorhersagewert von Eins erreicht werden. Dies ist grundsätzlich ein sehr vielversprechendes Ergebnis. Die Qualität der ermittelten Ansichtsmodelle wird darüberhinaus auch durch die Höhe des optimalen Schwellwerts mitbestimmt. Die gefundenen optimalen Schwellwerte reichen von 81% bis 100%. Da ein Schwellwert von 100% bedeutet, daß alle Teile des Ansichtsmodells für eine positive Erkennung benötigt werden, sind solche Modelle anfällig gegenüber Störungen und Verdeckungen. Dies betrifft die Knoten aus den Versuchen F, G, I und J. Aus Abbildung 6.65 ist ersichtlich, daß diese Knoten eine vergleichsweise hohe Höhe und niedrige Tiefe aufweisen und damit nahe an der Wurzel des Stichprobendendrogramms liegen. Dafür sprechen auch die hohen Korrelationskoeffizienten von $\pm 0{,}7$ (vgl. Tabelle 6.13) für die Abhängigkeit zwischen dem Schwellwert und der Höhe bzw. Tiefe. Die besten Modelle ergeben sich daher offenbar für die Abstraktion von kleineren Bildermengen in der Nähe der Blattknoten des Stichprobendendrogramms. Die Streudiagramme des Schwellwerts über der Höhe bzw. der Tiefe der modellierten Knoten im Stichprobendendrogramm in den Abbildungen 6.67 a) und b) zeigen diesen Zusammenhang deutlich. Hier ergibt sich ein deutlicher Anstieg des Schwellwerts ab einer Höhe von etwa 3 bis 4, also für die gemeinsame Modellierung von 8 bis 16 Stichprobenelementen in einem Ansichtsmodell. Die Abhängigkeit von der Tiefe im Stichprobendendrogramm zeigt einen Abfall des Schwellwerts unter 100% für Knoten mit einer Mindesttiefe von 8.

Da der Schwellwert aufgrund der beschriebenen Feinoptimierung als von der Ortstoleranz abhängiger Parameter betrachtet wird, kann dieser nicht direkt zur Auswahl vielversprechender Knoten des Stichprobendendrogramms herangezogen werden. Da jedoch die Ortstoleranz frei wählbar ist, wird als nächstes

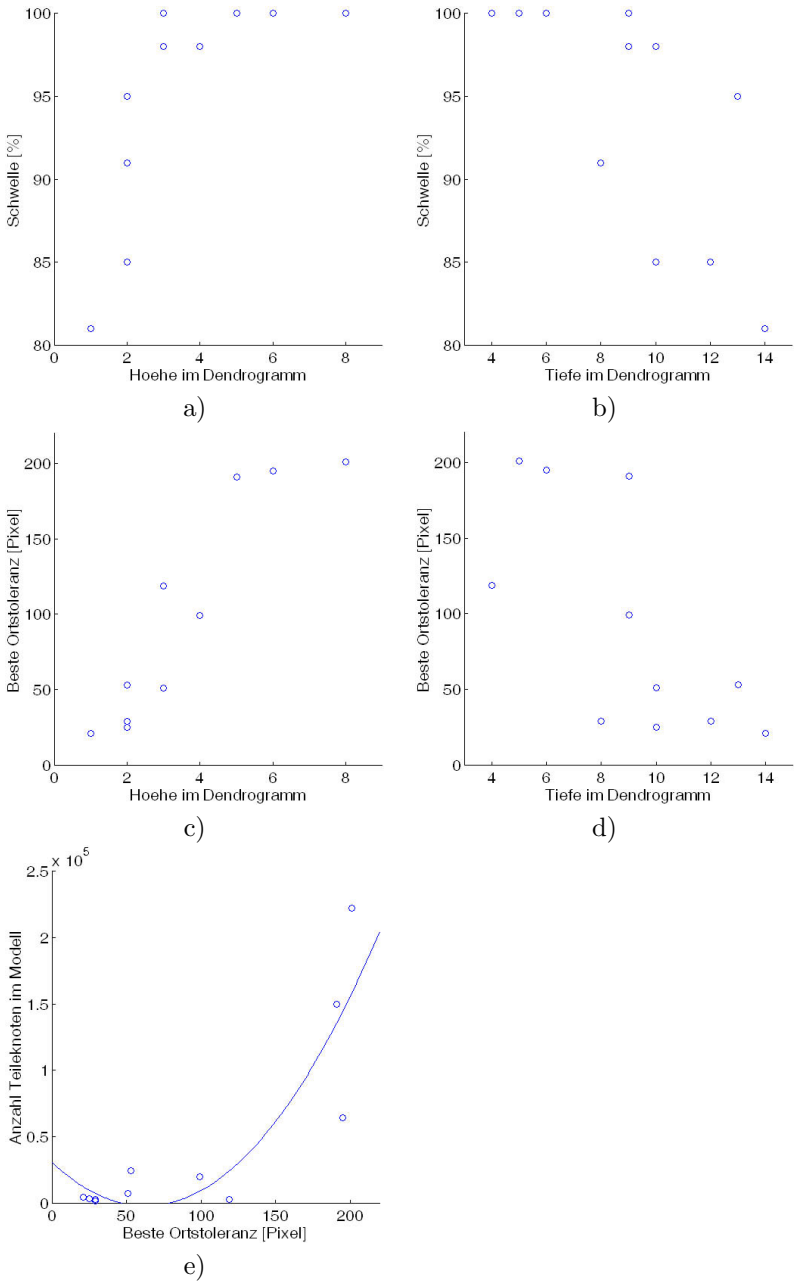
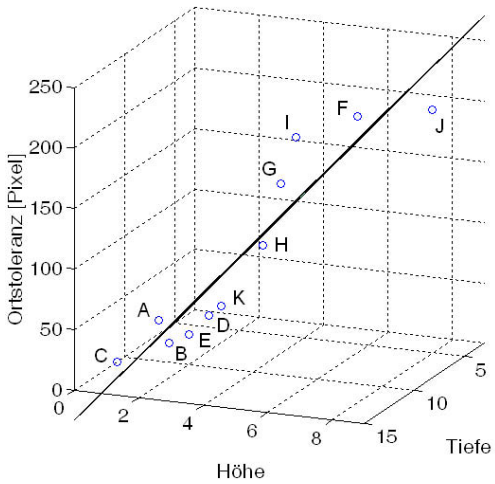
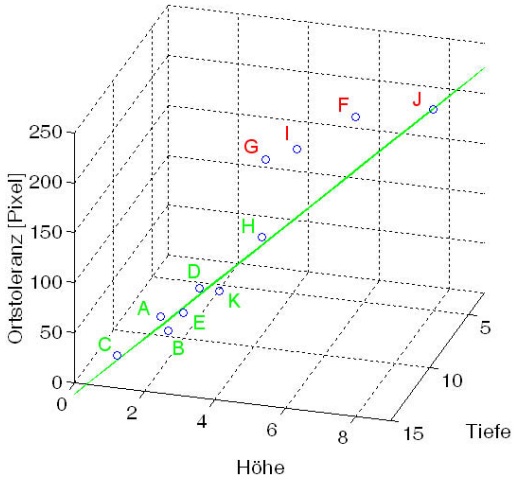


Abbildung 6.67: **Parameter der durch die Optimierung des positiven Vorhersagewerts erzeugten Ansichtsmodelle.** a, b) Der Schwellwert der optimalen Ansichtsmodelle hängt stark von der Höhe und Tiefe der modellierten Knoten im Stichprobendendrogramm ab. c, d) Ähnlich verhält sich die Ortstoleranz. e) Die Anzahl der Teile im Modell steigt stark mit der Ortstoleranz.



a) Anpassung der Ebenengleichung an die Ergebnisse der Versuche A–K



b) Anpassung an die Ergebnisse der Versuche A–E, H und K

Abbildung 6.68: **Lineare Näherung der Ortstoleranz für Ansichtsmodelle.** a) Die in den Versuchen A–K ermittelte optimale Ortstoleranz bezüglich des positiven Vorhersagewertes wurde durch eine Ebenengleichung über der Höhe und Tiefe der Knoten im Stichprobendendrogramm angenähert. Das Diagramm zeigt einen Blick entlang der Ebene, sodaß die Ebene als Gerade erscheint. b) Im unteren Diagramm wurden nur die Werte aus den Messungen A–E, H und K durch die Ebenengleichung angenähert (grüne Markierungen). Aufgrund eines ermittelten optimalen Schwellwerts unter 100% ergeben sich für diese Messungen die besten Modelle.

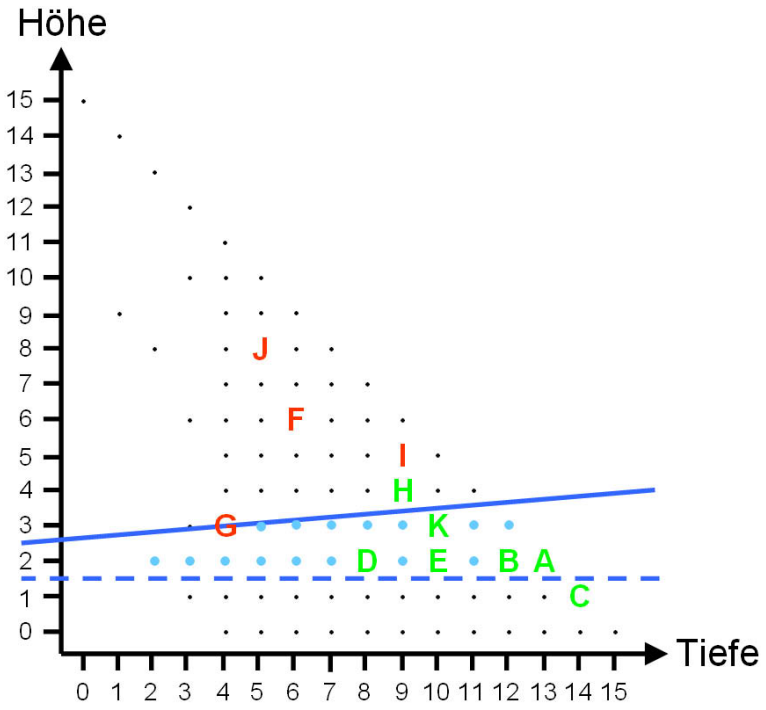


Abbildung 6.69: **Auswahl von Knoten im Stichprobendendrogramm.** Die durchgezogene blaue Linie gibt die maximale Höhe und minimale Tiefe von Knoten zur Modellierung von Ansichten an. Diese ergibt sich aus der linearen Näherung der Ortstoleranz über der Höhe und Tiefe in Verbindung mit dem Schwellwert von 90 Pixeln. Die gestrichelte Linie markiert die Mindesthöhe von Knoten für die Abstraktion über mehrere Stichprobenbilder. Die zur Modellierung in Frage kommenden Kombinationen aus Höhe und Tiefe sind blau markiert.

geprüft, wie diese von der Höhe und Tiefe abhängt. Die hohen Beträge der entsprechenden Korrelationskoeffizienten von 0,9 und 0,7 (Tabelle 6.13) sowie die Übereinstimmung des Vorzeichens verglichen mit den Korrelationskoeffizienten des Schwellwerts belegen, daß die Ortstoleranz alternativ zum Schwellwert herangezogen werden kann, um gut modellierbare Knoten des Stichprobendendrogramms zu finden. Ähnlichkeiten zeigen auch die Streudiagramme der Ortstoleranz in Abbildung 6.67 c) und d) zu denen des Schwellwerts in den Abbildungen a) und b). Um ein mathematisch geschlossenes Kriterium zu ermitteln, mit dem aussichtsreiche Ansichtsknoten bestimmt werden können, wird die Ortstoleranz über der Höhe und Tiefe durch eine Ebenengleichung angenähert. Aufgrund des hohen Korrelationskoeffizient von 0,9 zwischen der Ortstoleranz und der Höhe sind hier gute Ergebnisse zu erwarten. Die starke Streuung der Diagramme in Abbildung 6.67 regt auf der anderen Seite keine besonderen Funktionen höherer Ordnung an. Eine quadratische Anpassung der Ebenenparameter an die optimalen Ortstoleranzen der Versuche A–K liefert die Gleichung

$$\varsigma_{\text{A–K}} = -2{,}51 \cdot \text{Tiefe} + 29{,}42 \cdot \text{Höhe} + 12{,}3. \tag{6.22}$$

Abbildung 6.68 a) zeigt diese Ebenengleichung. Wie in der für die Darstellung gewählten Perspektive zu erkennen ist, sind die Meßwerte leicht S-förmig um die ermittelte Ebenengleichung herum angeordnet. Diese Abweichung führt dazu, daß die Ebene den globalen Verlauf der Meßwerte zwar gut wiedergibt, allerdings ungenaue Näherungswerte für die optimalen Ortstoleranzen der Versuche A–E, H und K liefert, die aufgrund des niedrigen ermittelten Schwellwerts für die Ansichtsmodellierung interessant sind. Um für die interessanten Kombinationen von Höhe und Tiefe eine genaue Näherung der Ortstoleranz zu erhalten, wurde an diese eine zweite Ebenengleichung angepaßt. Für diese ergibt sich die Gleichung

$$\varsigma_{\text{beste}} = 4{,}48 \cdot \text{Tiefe} + 32{,}21 \cdot \text{Höhe} - 79{,}4, \tag{6.23}$$

welche in Abbildung 6.68 b) dargestellt ist.

Eine weitere wichtige Eigenschaft des Modells ist die Größe. Wie Abbildung 6.67 e) in Form der Kreismarkierungen zeigt, steigt die Anzahl der zu einem Ansichtsmodell kombinierten Teile stark an, wenn Modelle für viele Stichprobenelemente erzeugt werden, die eine hohe Ortstoleranz erfordern. Da die Ortstoleranz einen quadratischen Bereich in der Bildebene definiert, sollte sich hier eine quadratische Abhängigkeit ergeben. Die Anpassung einer quadratischen Funktion an die Anzahl der Teile liefert die Gleichung

$$\text{Teileanzahl} = 8{,}3393 \cdot \varsigma^2 - 1044{,}2 \cdot \varsigma + 30514,$$

welche in Abbildung 6.67 e) als Kurve eingezeichnet ist. Die Meßwerte lassen sich jedoch schlecht durch diese Funktion darstellen: Es ergibt sich ein Tiefpunkt mit einer negativen Teileanzahl, der zudem bei einer Ortstoleranz von über 50 Pixeln liegt, statt im Nullpunkt. Hier werden offenbar wieder die komplexen Geometrieeinflüsse deutlich. Für die Modellerzeugung bedeutet der Anstieg der

Teileanzahl, daß nur für die Knoten des Stichprobendendrogramms Ansichtsmodelle erzeugt werden, deren Teileanzahl noch ein praktikables Laufzeitverhalten erlaubt.

Als Konsequenz aus den untersuchten Abhängigkeiten zwischen den verschiedenen Modellparametern wird für die Ansichtsmodellierung folgende Strategie angewandt:

- Knoten des Stichprobendendrogramms werden durch ihre Höhe und Tiefe im Graphen beschrieben.
- Die Auswahl geeigneter Knoten zur Modellierung als Ansicht geschieht über die Betrachtung der Ortstoleranz.
- Die Ortstoleranz wird durch eine Ebenengleichung über der Höhe und Tiefe angenähert (vgl. Gleichung 6.22).
- Es werden die Knoten ausgewählt, für die sich eine Ortstoleranz unter einer bestimmten Schwelle ergibt.
- Die Schwelle wird so gewählt, daß sich eine behandelbare Modellgröße ergibt.
- Die Parametrisierung der Ansichtsmodelle bezüglich der Ortstoleranz geschieht über eine zweite Ebenengleichung, die nur die Knoten des Stichprobendendrogramms umfaßt, für die gute Modelle erzielbar sind (vgl. Gleichung 6.23). Eine minimale Ortstoleranz von 5 Pixeln wird festgelegt.
- Der Schwellwert wird anschließend auf einen hohen positiven Vorhersagewert optimiert. Dies geschieht anhand der Trainingsstichprobe.

Experimentell hat für die Auswahl von Knoten des Stichprobendendrogramms eine durch Gleichung 6.22 angenäherte maximale Ortstoleranz von 90 Pixeln als gerade noch tolerierbar herausgestellt. Auf diese Weise ergibt sich für die Klassifikation einzelner Stichprobenbilder eine Laufzeit im Minutenbereich. Die Laufzeit ist stark abhängig von der Speicherauslastung des Versuchsrechners (max. 8GB RAM) durch die erzeugten Hypothesen und kann sich im Fall der Speicherauslagerung durch das Betriebssystem (hier Windows Vista64) schlagartig auf den Stundenbereich ausdehnen. Um die Modellgröße weiter einzugrenzen, wird bei der Auswahl von Knoten des Stichprobendendrogramms eine Mindesthöhe von zwei festgelegt, sodaß ein Ansichtsmodell über mindestens drei Bilder generalisiert. Abbildung 6.69 zeigt die ausgewählten Knoten. Auf diese Weise ergibt sich ein Objektmodell mit insgesamt 800 000 Knoten gemessen über alle Merkmale, Teile und Ansichten.

Ansichtsmodelierung bei Optimierung der Korrektklassifikationsrate

Wie Tabelle 6.12 zeigt, resultieren alle getesteten Knoten des Stichprobendendrogramms in einer Korrektklassifikationsrate von etwa 79 Prozent. Eine Abhängigkeit von der Höhe und Tiefe ist nicht erkennbar (vgl. Tabelle 6.14).

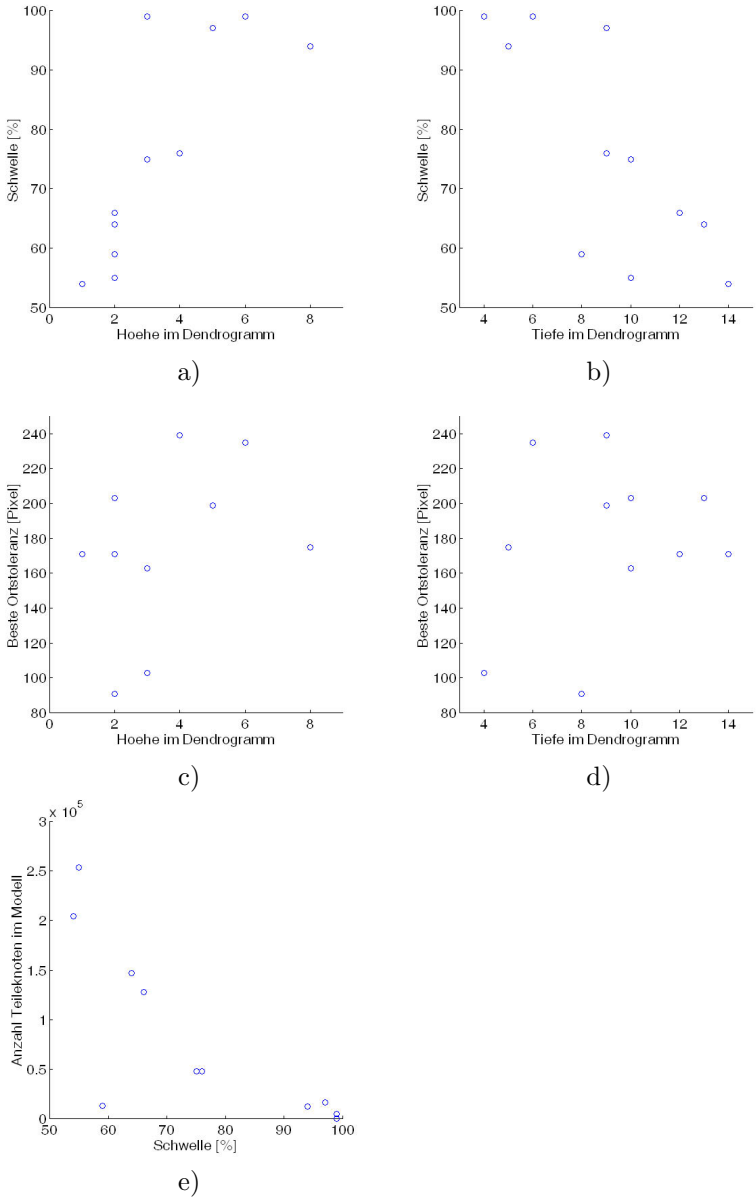


Abbildung 6.70: **Parameter der durch die Optimierung der Korrektklassifikationsrate erzeugten Ansichtsmodelle.** a, b) Der Schwellwert hängt von der Höhe und Tiefe der modellierten Knoten im Stichprobendendrogramm ab. c, d) Für die Ortstoleranz ist der Zusammenhang jedoch höchstens schwach zu erkennen. e) Für kleine Schwellwerte ergeben sich schnell unattraktive Modellgrößen.

Erster Parameter	**Zweiter Parameter**	**Korrelations-koeffizient**
Korrektklassifikationsrate	Höhe	0,1395
Korrektklassifikationsrate	Tiefe	−0,1535
Schwelle	Höhe	0,7937
Schwelle	Tiefe	−0,7729
Ortstoleranz	Höhe	0,3205
Ortstoleranz	Tiefe	0,2366
Anzahl Teileknoten	Schwelle	−0,7806

Tabelle 6.14: Korrelationskoeffizient von Ansichtsparametern bei der Optimierung der Korrektklassifikationsrate. Die Werte sind wieder auf den Bereich von −1 bis +1 normiert.

Die Optimierung über alle Positivbeispiele der Stichprobe bewirkt hier, daß viel mehr Elemente erkannt werden als für die Modellerzeugung herangezogen werden. Die Stichprobenelemente, für die das jeweilige Ansichtsmodell erzeugt wurde, haben daher nur einen geringen Einfluß auf das Ergebnis. Daß die Schwellwerte teilweise bis auf 55 Prozent (Versuch E) sinken, deutet zudem an, daß die Ähnlichkeit zwischen dem erlernten und dem klassifizierten Muster eine geringere Rolle spielt als bei dem vorherigen Optimierungsverfahren. Dafür spricht auch die im Mittel bei 178 Pixeln liegende Ortstoleranz im Vergleich zu nur 92 Pixeln für die vorherige Methode. Die Ergebnisse beruhen daher möglicherweise auf einer generellen Bevorzugung der positiven Stichprobenelemente durch das visuelle Teile-Alphabet.

Die für die Optimierung des positiven Vorhersagewertes identifizierte Abhängigkeit zwischen dem Schwellwert und der Höhe und Tiefe im Stichprobendendrogramm zeigt sich auch bei der Optimierung auf eine hohe Korrektklassifikationsrate. Wie Tabelle 6.14 zeigt, ergeben sich ähnliche hohe Korrelationskoeffizienten zwischen dem Schwellwert und der Höhe bzw. Tiefe wie bei der Optimierung auf den positiven Vorhersagewert (vgl. Tabelle 6.13 oder Abbildung 6.70 oben). Die besten Ansichtsmodelle ergeben sich im Stichprobendendrogramm also wieder für Knoten in der Nähe der Blattknoten, d.h. für die Verallgemeinerung über wenige Beispielbilder.

Wie die Abbildungen 6.70 c) und d) zeigen, ist die Abhängigkeit zwischen der optimalen Ortstoleranz und der Lage im Stichprobendendrogramm kaum ausgeprägt. Es kann daher keine Näherungslösung in Form einer Ebenengleichung angepaßt werden. Dies erschwert die Parametrisierung von Ansichtsmodellen.

Im Vergleich zur vorherigen Optimierung ergeben sich zudem größere Modelle. Wie Abbildung 6.70 e) zeigt, ergeben sich insbesondere für die Modelle mit einem niedrigen Schwellwert besonders viele Teile und hohe Testlaufzeiten, wodurch sie für die weiteren praktischen Versuche derzeit unattraktiv sind.

Aus diesen Gründen wird zur Optimierung des Modells auf die Korrektklassifikationsrate die gleiche Strategie zur Auswahl von Knoten des Stichprobendendrogramms eingesetzt wie für die Optimierung auf den positiven Vorher-

sagewert. Auch die Ortstoleranz wird auf die durch Gleichung 6.23 gegebenen Werte gesetzt. Der Schwellwert der Ansichtsmodelle wird dagegen so gewählt, daß sich eine maximale Korrektklassifikationsrate ergibt. Das resultierende Modell ist auf der Ansichtsebene suboptimal. Die Trainingsmethode liefert daher eine Untergrenze für die theoretisch erzielbare Korrektklassifikationsrate des Gesamtmodells.

Geometriebedeutung auf Ansichtsebene

Auf der Teileebene des Modells ergibt sich nach Gleichung 6.17 eine Ortstoleranz von 7–21 Pixeln für die im Teilealphabet auftretenden Teilegrößen von 10–60 Pixeln. Die geometrischen Abweichungen, die auf der Merkmalsebene toleriert werden, liegen damit bei 4–12 Prozent der mittleren Objektgröße. Die Größe der Teile selbst liegt bei 6–36 Prozent der mittleren Objektgröße. Auf der Ansichtsebene liegt die Ortstoleranz noch einmal höher als auf der Teileebene. Für die ausgewählten Knoten des Stichprobendendrogramms liefert Gleichung 6.23 Ortstoleranzen zwischen 5 und 72 Pixel. Im Vergleich zur mittleren Objektgröße sind dies 3 bis 43 Prozent.

Dieser Anstieg der Ortstoleranz auf höheren Abstraktionsebenen des Modells zeigt sich deutlich in den Positionen der modellierten Teile. Abbildung 6.71 zeigt ein Beispiel für die Teilepositionen ausgewählter Prototypen des Teilealphabets für ein Ansichtsmodell aus Versuch K mit einer Ortstoleranz von $\varsigma = 50$ Pixel. Die Abbildung stellt sowohl die gemeinsamen Trefferbereiche von Prototypen über alle zu der Ansicht zusammengestellten Stichprobenbilder dar (blaue Markierung) als auch die für die tatsächliche Modellierung unterabgetasteten Teilepositionen (schwarze Markierung). Aufgrund der hohen Ortstoleranz ergeben sich für die meisten Einträge des Teilealphabets Trefferpositionen in einem überwiegend zusammenhängenden Bereich um das Objektzentrum. Die Größe dieser Bereiche variiert zwar, die geringe Formenvielfalt der Trefferbereiche zeigt jedoch, daß die Geometrie auf dieser Abstraktionsebene von untergeordneter Bedeutung ist.

Die mit der Hierarchieebene steigende Ortstoleranz bildet so eine Parallele zum Aufbau des visuellen Kortex. Dort steigt die Größe der rezeptiven Felder von Gehirnzellen mit der Hierarchiestufe [Hub88, SD90, TRS01]. Die Ortstoleranz läßt sich mit den rezeptiven Feldern der Gehirnzellen vergleichen, da nur innerhalb dieser Felder ein Muster erkannt wird. Die geringere Bedeutung der Geometrie auf Ansichtsebene könnte ausgenutzt werden, um kompakterer Modelle zu erzeugen. Beispielsweise könnten die Schnittbilder vereinfacht durch den Schwerpunkt und einen Radius dargestellt werden, sodaß sie sich durch ein einzelnes Teil modellieren ließen.

6.5 Training und Test auf Kategorieebene

Die Stichprobenelemente, auf die einzelne Ansichtsmodelle trainiert wurden, stehen beispielhaft für einen bestimmten Ausschnitt aller möglichen Ansich-

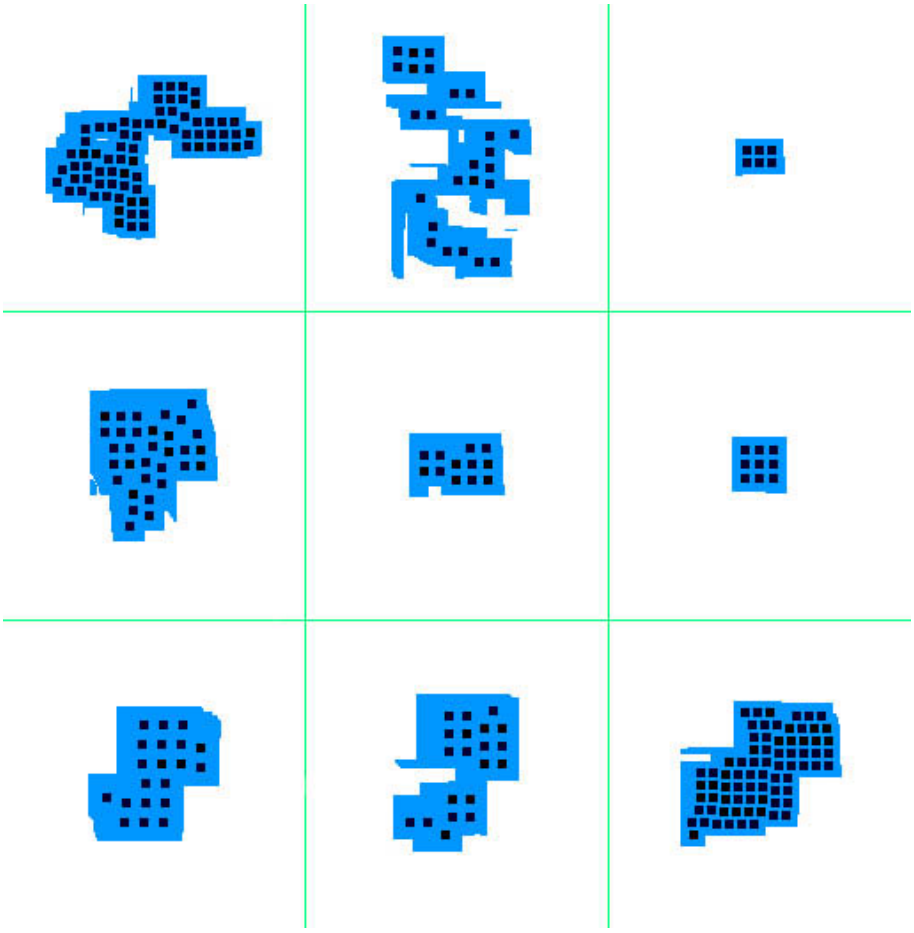


Abbildung 6.71: **Teilegeometrie auf Ansichtsebene.** Das Diagramm zeigt die Positionen von Teilen in einem trainierten Ansichtsmodell für neun verschiedene Prototypen des visuellen Teilealphabets. Die Positionen der Teile sind schwarz markiert. Die blauen Flächen geben die Teilepositionen des Schnittbildes aus Gleichung 6.19 an. Dies sind die Teilepositionen, die in allen zu einer Ansicht zusammengefaßten Stichprobenbildern auftreten. Die schwarz markierten Teilepositionen wurden durch Unterabtastung ermittelt. Am Rand der Schnittfiguren wurden keine Teile erzeugt, um die Dilatation in Gleichung 6.18 zu kompensieren.

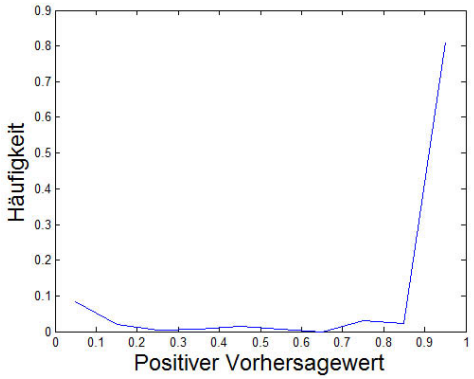


Abbildung 6.72: **Histogramm des positiven Vorhersagewerts von Ansichtsmodellen.** Der positive Vorhersagewert von Ansichtsmodellen ist bimodal verteilt. Ca. 8% der Modelle haben einen extrem niedrigen positiven Vorhersagewert von unter 10%. Sie erzeugen daher viele falsche Treffer.

ten des trainierten Objekts. Die Menge von Objektdarstellungen, auf die ein Ansichtsmodell anspricht, ist daher nicht nur auf die trainierten Stichprobenbilder beschränkt. Aus diesem Grund liefern Ansichtsmodelle nicht nur für die trainierten Stichprobenelemente Treffer, sondern zum Teil auch auf anderen, nicht trainierten Positivbeispielen. Die Ansichtsmodelle stellen auf diese Weise wieder ein visuelles Alphabet dar, dessen Elemente auf der Kategorieebene kombiniert werden. Da einzelne Stichprobenelemente mitunter auch durch fremde Ansichtsmodelle erkannt werden, enthält das visuelle Alphabet auf Ansichtsebene ein gewisses Maß an Redundanz. Diese wird auf der Kategorieebene genutzt, um die Zuverlässigkeit der Erkennung zu steigern und das Gesamtmodell auf verschiedene Ziele zu optimieren.

6.5.1 Optimierung auf den positiven Vorhersagewert

Um das visuelle Alphabet auf Ansichtsebene sinnvoll nutzen zu können, muß zunächst für jedes Stichprobenelement ermittelt werden, durch welche Ansichtsmodelle es erkannt wird. Diese Information wird in Form von binären *Ansichtsmatrizen*

$$I_P \begin{bmatrix} \iota_{P,1,1} & & \iota_{P,\ldots,1} \\ & \ddots & \\ \iota_{P,1,\ldots} & & \iota_{P,\ldots,\ldots} \end{bmatrix} \text{ und } I_N \begin{bmatrix} \iota_{N,1,1} & & \iota_{N,\ldots,1} \\ & \ddots & \\ \iota_{N,1,\ldots} & & \iota_{N,\ldots,\ldots} \end{bmatrix}$$

für die positiven bzw. negativen Stichprobenelemente zusammengestellt. Ein Wert von $\iota_{P,i,j} = 1$ bzw. $\iota_{N,i,j} = 1$ gibt an, daß ein durch j indiziertes Ansichtsmodell das positive bzw. negative Stichprobenelement i erkennt.

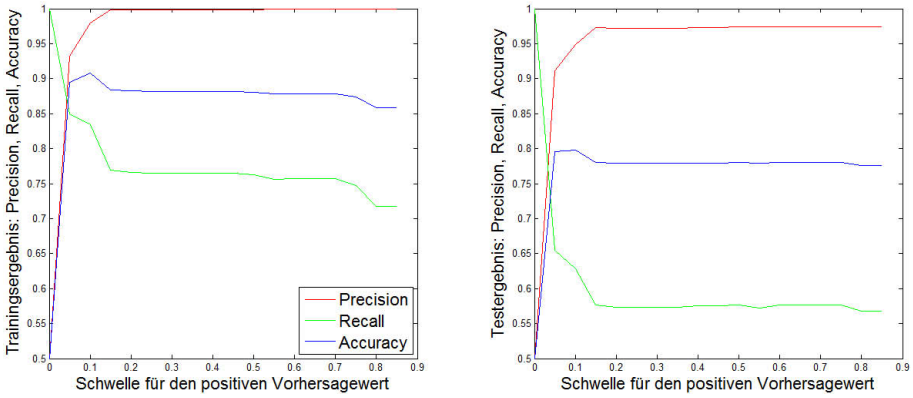


Abbildung 6.73: **Versuchsergebnisse bei Variierung des Schwellwerts für den positiven Vorhersagewert.** Die Ergebnisse sind sehr gleichförmig für einen Schwellwert über 20 Prozent.

Um den positiven Vorhersagewert zu optimieren, werden nun für jedes einzelne Stichprobenelement i die Ansichtsmodelle $j_1, j_2, \ldots$ gesucht, die zu einer positiven Erkennung führen, d.h. für die $\iota_{P,i,j} = 1$ ist. Die Anzahl der erkennenden Ansichtsmodelle wird durch die Formel

$$n_{\mathfrak{P}}(i) = \sum_j \iota_{P,i,j}$$

gegeben. Um die Redundanz der gleichzeitigen Treffer der Ansichtsmodelle voll auszuschöpfen, werden alle $n_{\mathfrak{P}}$ erkennenden Ansichtsmodelle durch einen gemeinsamen Kategorieknoten zusammengefaßt. Da die Ansichtsmodelle bereits vollständige Objekte repräsentieren, wird für alle Ansichtsknoten die gleiche Position angenommen. Die Ortstoleranz wird so eingestellt, daß sich für die vorhandene Stichprobe keine Einschränkungen ergeben. Der Schwellwert wird so eingestellt, daß sich ein maximaler positiver Vorhersagewert ergibt. Da die Anzahl der richtigen Treffer durch die Auswahl der Ansichtsmodelle bereits feststeht, kann der positive Vorhersagewert nur durch die Minimierung der falschen Treffer optimiert werden. Auf jedes Negativbeispiel der Stichprobe können dabei unterschiedlich viele Ansichtsmodelle des Kategorieknotens ansprechen. Maximal sprechen

$$n_{\mathfrak{N}} = \max_k \sum_{\iota_{P,i,j}=1} \iota_{N,k,j}$$

Ansichtsmodelle sowohl auf das zu erlerndende positive Stichprobenelement i als auch auf ein beliebiges Negativbeispiel k an. Anhand der Bedingung $\iota_{P,i,j} = 1$ unter dem Summenzeichen wird jedes einzelne Ansichtsmodell j ermittelt, welches das Positivbeispiel i erkennt. Durch die Summation über $\iota_{N,k,j}$ wird dann berechnet, wieviele dieser Ansichtsmodelle auch Treffer auf einem Negativbeispiel k liefern. Die Maximierung über alle Negativbeispiele führt dann zu dem

Trainingsstichprobe

		Real	
		Objekt	Hintergrund
Erkannt	Objekt	605	0
	Hintergrund	195	800

Positiver Vorhersagewert	100%
Korrektklassifikationsrate	88%
Sensititivität	76%

Teststichprobe

		Real	
		Objekt	Hintergrund
Erkannt	Objekt	458	12
	Hintergrund	342	788

Positiver Vorhersagewert	97%
Korrektklassifikationsrate	78%
Sensititivität	57%

Tabelle 6.15: Trainings- und Testergebnisse bei Auslegung des Modells auf einen hohen positiven Vorhersagewert.

Stichprobenelement k, für das maximal viele Ansichtsmodelle j, die das positive Stichprobenelement i erkennen, Treffer liefern.

Der Schwellwert ϑ des Kategorieknotens wird daher gerade so eingestellt, daß das Maximum nach Gleichung 5.7 keine Erkennung auslöst, d.h.

$$\vartheta = n_{\mathfrak{N}}/n_{\mathfrak{P}}.$$

Das resultierende Modell erzeugt allerdings trotz der Optimierung des positiven Vorhersagewertes noch viele falsche Treffer. Ein Histogramm der optimierten positiven Vorhersagewerte (siehe Abb. 6.72) zeigt, daß sich für etwa 8 Prozent der Ansichtsmodelle extrem niedrige positive Vorhersagewerte ergeben. Diese Ansichtsmodelle verhalten sich offenbar als Ausreißer, indem sie anders als die meisten Ansichtsmodelle unselektiv auf die Eingaben reagieren. Damit verschlechtern sie das Gesamtergebnis der Objekterkennung. Um ein zuverlässiges Gesamtmodell auf Kategorieebene zu erzielen, werden diese Ansichtsmodelle mit Hilfe eines Schwellwerts von 50% für den positiven Vorhersagewert von der Modellbildung ausgeschlossen. Da die Verteilung des positiven Vorhersagewerts anscheinend bimodal ist, ist die Wahl des Schwellwerts unkritisch (siehe Abb. 6.73).

Die Ergebnisse auf der Trainingsstichprobe können anhand der Ansichtsmatrizen I_P und I_N direkt berechnet werden. Ein Objekt gilt dann als detektiert, wenn mindestens ein Kategorieknoten der behandelten Cartoon-Klasse auf ein Bild anspricht. Da das Training auf die Modellierung jedes einzelnen Stichprobenelements ausgerichtet ist, sind die Ergebnisse für die Trainingsstichprobe nicht auf die Wirklichkeit anwendbar. Insbesondere kann die Repräsentativität der Trainingsstichprobe für die Erscheinungsvielfalt der Objekte nicht aus den

Ergebnissen der Trainingsstichprobe abgelesen werden. Aus diesem Grund wird das gewonnene Modell auch auf eine Teststichprobe angewandt. Diese besteht wie die Trainingsstichprobe aus 800 randomisierten Positiv- und Negativbeispielen. Die Trainings- und die Teststichprobe haben keine gemeinsamen Elemente.

Die Ergebnisse des Trainings und des Tests zeigt Abbildung 6.15. Für beide Stichproben ergibt sich ein hoher positiver Vorhersagewert von mindestens 97 Prozent. Das Erkennungsverfahren ist daher sehr zuverlässig. Die Korrektklassifikationsrate liegt für die Teststichprobe bei 78 Prozent, was in anbetracht der Variation in der Erscheinung des trainierten Objekts ein sehr guter Wert ist. Die geringere Sensitivität von 76 Prozent in der Trainingsstichprobe und insbesondere von 57 Prozent in der Teststichprobe zeigt eine hohe Spezialisierung des Modells an. Das Modell kann die Beispiele der Trainingsstichprobe aufgrund der konservativen Auslegung auf den positiven Vorhersagewert nicht in vollem Maße verallgemeinern. Dies bedeutet für ein Anwendungsgebiet, das einen hohen positiven Vorhersagewert erfordert, daß auf eine nicht zu kleine Stichprobe geachtet werden muß.

6.5.2 Auslegung des Modells auf Sensitivität und Korrektklassifikationsrate

Die konservative Auslegung des Modells auf die Vermeidung von falsch positiven Treffern ist für Anwendungen ungünstig, bei denen fälschlicherweise ignorierte Objekte höhere Kosten verursachen als fälschlicherweise angezeigte Objekte. Um die Leistungsfähigkeit der Modellierung für einen solchen Anwendungsbereich zu untersuchen, wird nun auch eine sensitivere Modellkonfiguration überprüft.

Dazu wird zunächst wieder nach dem oben genannten Verfahren für jedes Positivbeispiel der Stichprobe ein Kategorieknoten erzeugt. Dabei umfaßt jeder Kategorieknoten wieder alle Ansichtsmodelle, die das betreffende Stichprobenelement erkennen. Im Gegensatz zu dem obigen Verfahren wird der Schwellwert jedoch nicht auf eine starke Hintergrundunterdrückung, sondern auf eine hohe Korrektklassifikationsrate optimiert. Diese wird zudem nicht bezüglich der wenigen zu dem jeweiligen Ansichtsmodell gehörenden Stichprobenelemente durchgeführt, sondern bezüglich aller Stichprobenelemente. Da die Schwellwerte auf eine größere und variablere Menge von Stichprobenelementen angepaßt werden, erkennt jeder einzelne Kategorieknoten mehr verschiedene Objekte als in der konservativen Konfiguration. Die Berücksichtigung der Korrektklassifikationsrate verhindert dabei die Erzeugung einer völlig unselektiven Knotenmenge, die beliebige Muster als Objekt erkennt.

Die Objekterkennung geschieht dann wieder dadurch, daß ein einzelner Kategorieknoten anspricht. Da jeder Kategorieknoten auf ein bestimmtes Beispiel für eine Objektansicht trainiert wurde, ergibt die Gesamtheit der Kategorieknoten eine hohe Sensitivität. Die Optimierung der Schwellwerte auf die Korrektklassifikationsrate sichert dabei eine gewisse Mindestselektivität des gesamten Modells.

Praktische Versuche zeigen, daß wieder ein bestimmter Anteil der Ansichtsmodelle gleichermaßen auf positive wie auf negative Beispiele der Stichprobe

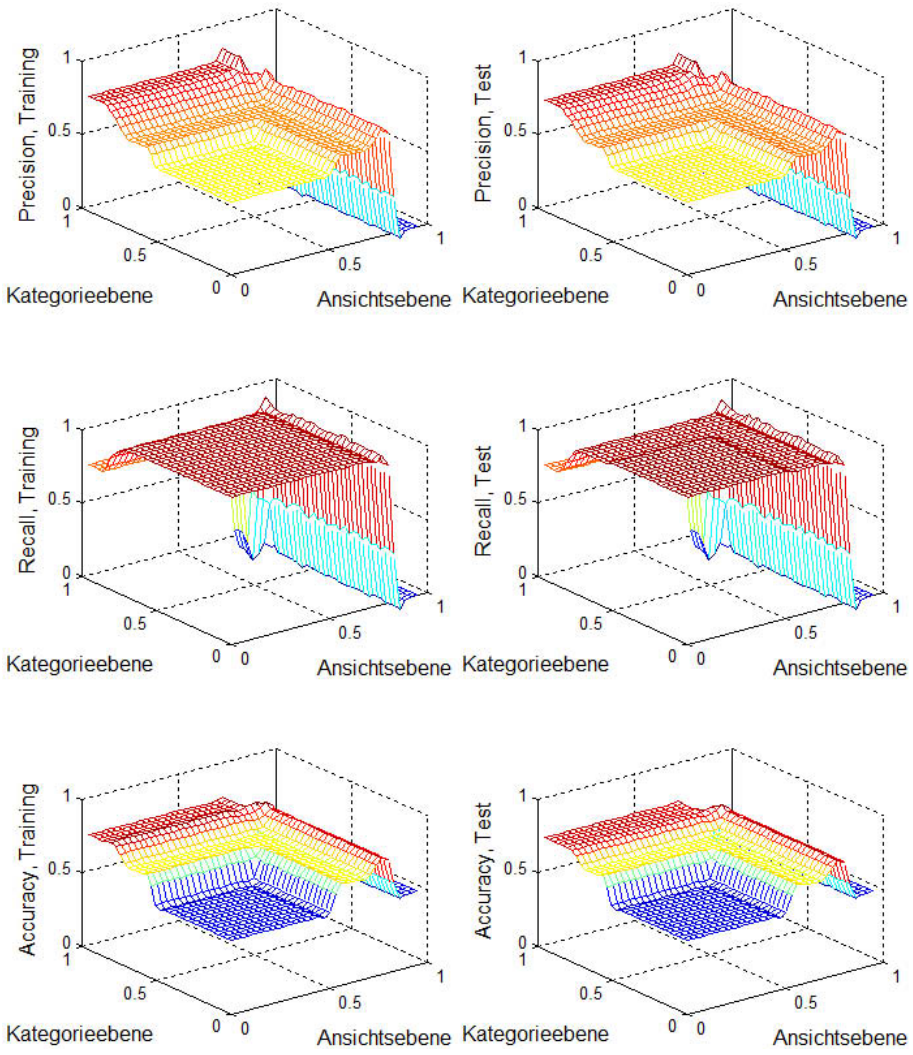


Abbildung 6.74: **Modellgüte bei Variation der Schwellwerte für die Korrektklassifikationsrate.** Varriiert wurden die Schwellwerte für die Korrektklassifikationsrate auf der Ansichtsebene und der Kategorieebene. Die Diagramme links geben den positiven Vorhersagewert (Precision), die Sensitivität (Recall) und die Korrektklassifikationsrate (Accuracy) für die Trainingsstichprobe an. In der rechten Spalte sind die entsprechenden Diagramme für die Teststichprobe angegeben. Aus Darstellungsgründen wurden im Fall einer Division durch Null aufgrund fehlender Treffer Diagrammwerte auf Null gesetzt.

reagiert. Aus diesem Grund wird ein Schwellwert für die Korrektklassifikationsrate einzelner Ansichtsmodelle eingeführt, mit dem Ansichtsmodelle aus dem Gesamtmodell ausgeschlossen werden können, wenn sie das Gesamtergebnis verschlechtern. Ähnliches gilt für die resultierenden Knoten auf Kategorieebene. Aufgrund der relativ niedrigen Schwellen für die zur Erkennung nötigen Teile liefern die Kategorieknoten oft überlappende Treffer. Aus diesem Grund sind nicht alle Knoten nötig, um eine hohe Stichprobenüberdeckung zu erreichen. Es wird daher ein weiterer Schwellwert eingeführt, mit dem die geeignetsten Kategorieknoten ausgewählt werden. Dazu wird für jeden Kategorieknoten die Korrektklassifikationsrate auf der Trainingsstichprobe bestimmt. Liegt diese unter dem Schwellwert, wird der Knoten aus dem Modell entfernt.

Die Optimierung der beiden Schwellwerte verursacht kaum Rechenaufwand und geschieht daher durch erschöpfende Suche. Abbildung 6.74 zeigt die Güte des Gesamtmodells über den Schwellwerten. Für die Trainingsstichprobe wird auf diese Art eine maximale Korrektklassifikationsrate von 81 Prozent erreicht für die Schwellen von 75 Prozent auf der Ansichtsebene und 80 Prozent auf der Kategorieebene. Die Sensitivität ist mit 95 Prozent sehr hoch, etwas geringer ist der positive Vorhersagewert. Für die Teststichprobe liefert diese Konfiguration einer Korrektklassifikationsrate von 78 Prozent, eine Sensitität von 93 Prozent und einen positiven Vorhersagewert von 72 Prozent. Die hohe Ähnlichkeit der Ergebnisse für die beiden Stichproben belegt die Repräsentativität der Ergebnisse.

Im Vergleich zu der konservativen Konfiguration aus dem vorigen Abschnitt ist die Sensitivität jetzt deutlich höher. Dies geht allerdings zulasten des positiven Vorhersagewertes. Die Korrektklassifikationsrate ist für beide Modellauslegungen ungefähr gleich. Wie Tabelle 6.16 für die zweitbeste Schwellwertkonfiguration zeigt, entfällt der größte Teil der verbleibenden Fehlerrate jetzt auf falsch positive Treffer. Die sensitivere Konfiguration liefert daher bezüglich der Korrektklassifikationsrate ein genauso gutes Modell wie die konservative Konfiguration. Es wird nur ein anderes Trainingsziel verfolgt.

Aufgrund der in Abschnitt 6.4.4 angegebenen Maßnahmen zur Beschleunigung des Trainings stellen die Ergebnisse eine praktische Untergrenze für die erreichbare Korrektklassifikationsrate dar. Da das Modell auf Ansichtsebene aus Geschwindigkeitserwägungen heraus nicht konsequent auf eine hohe Korrektklassifikationsrate optimiert ist, sind theoretisch auch höhere Werte möglich.

Trainingsstichprobe

		Real	
		Objekt	Hintergrund
Erkannt	Objekt	733	250
	Hintergrund	67	550

Positiver Vorhersagewert	75%
Korrektklassifikationsrate	80%
Sensititivität	92%

Teststichprobe

		Real	
		Objekt	Hintergrund
Erkannt	Objekt	708	278
	Hintergrund	92	522

Positiver Vorhersagewert	72%
Korrektklassifikationsrate	77%
Sensititivität	89%

Tabelle 6.16: Vertauschungsmatrix für die zweitbeste Kombination von Schwellwerten bei der Auslegung des Modells auf eine hohe Korrektklassifikationsrate und Sensitivität. Der Schwellwert auf Ansichtsebene beträgt 70%, der auf Kategorieebene 80%.

Kapitel 7

Fazit und Ansätze für zukünftige Arbeiten

Die Auswertung der visuellen Umgebungsinformation ist derzeit eines der drängendsten Probleme, wenn es um die zielgerichtete Interaktion eines technischen Systems mit einer dynamischen Umgebung geht. Um dieses Problem zu lösen, wird in dieser Arbeit die visuelle Erkennung von Objekten, insbesondere die der schwierigen Klasse der verformbaren Objekte untersucht. Die Ergebnisse führen zu einem Erkennungssystem, das neuartig ist bezüglich des Trainings von visuellen Alphabeten auf verschiedenen Abstraktionsebenen. Berücksichtigt werden auch neuere Erkenntnisse darüber, wie verschiedene Objektansichten im Gehirn dargestellt werden. Ein besonderes Kennzeichen ist zudem, daß für alle Teile des Modells eine individuelle Gewichtung von räumlicher Teile-Anordnung und Merkmalscharakteristik einstellbar ist. Dabei sind mehrere Ansichten eines Objekts trainierbar.

7.1 Zusammenfassung des entwickelten Verfahrens

Um eine Toleranz gegen geometrische Verformungen zu erreichen, wird ein auf Teilen basierendes Modell eingeführt. Der für eine robuste Erkennung erforderliche Grad an geometrischer Freiheit kann für jedes Teil eingestellt werden. Das Modell ist in Anlehnung an den visuellen Kortex hierarchisch aufgebaut, was die Beschreibung der verschiedenen Erscheinungen der modellierten Objektklassen auf unterschiedlichen Abstraktionsebenen erlaubt. Die Objekterkennung basiert auf Mehrheitsentscheiden.

Wichtig für die Allgemeingültigkeit der hier gewonnenen Ergebnisse ist, daß das hier vorgestellte Objekterkennungssystem auf eine große Vielfalt von Objekten anwendbar ist. Für die niedrigsten Hierarchiestufe des Modells werden daher verschiedene, im Falle von Kanten und Flächen sogar komplementäre Merkmale

kombiniert. Die Eignung dieser Merkmale zur Erkennung realer Objekte wurde in früheren Arbeiten gezeigt [SK05, SK06].

Um die höheren Hierarchiestufen zu trainieren, werden die statistischen Abhängigkeiten zwischen den Teilen untersucht. Für die Merkmalsebene wird gezeigt, daß die Geometrieinformationen im Vergleich zur Merkmalsausprägung den Haupteinfluß auf die Modellqualität darstellt. Eine Analyse der Verbundwahrscheinlichkeiten von Zweierkombinationen von Merkmalen identifiziert die räumliche Nähe von Merkmalen in der Bildebene als entscheidenden Einfluß. Für eine mögliche Relevanz von unzusammenhängenden, weit entfernten Merkmalen werden keine Hinweise gefunden. Dies ist in Übereinstimmung mit dem Ansatz von Ommer und Buhmann [OB06], auch wenn diese eine psychologisch motivierte Begründung nennen. Da in der vorliegenden Arbeit zudem ein einfacher Zusammenhang zwischen dem räumlichen Abstand und den Modellparametern gefunden wird, kann durch Clusterung von Merkmalen eine Menge von Teilekandidaten für die Bildung der nächsthöheren Hierarchieebene ermittelt werden.

Inspiriert durch das biologische Vorbild wird die Menge von Teilekandidaten zu einem allgemeingültigen visuellen Alphabet verdichtet, daß aus prototypischen Teilen besteht. Für den hierzu notwendigen Vergleich der in den Teilekandidaten gespeicherten Mustern wird ein Kriterium entwickelt, daß trotz der für den Menschen visuellen Plausibilität mathematisch leicht beschreibbar ist. Für die Modellparametrisierung auf dieser Ebene sind die Größe der Teile, die Größe der Objekte, und mögliche Ober- und Untermengen der enthaltenen Merkmale von Bedeutung.

Die Teile des visuellen Alphabets werden auf der nächsten Hierarchiestufe zu typischen Objektansichten zusammengestellt. Zur Identifikation typischer Objektansichten werden die Bilder einer Stichprobe nach ihrer Ähnlichkeit sortiert. Dadurch können aus dem visuellen Alphabet Teile ausgewählt und kombiniert werden, die für die jeweilige Ansicht typisch sind. Die Sortierung wird hierarchisch in Form eines Binärbaums dargestellt, wobei die Knoten des Baums je nach ihrem Abstand von den Blattknoten bzw. von der Wurzel mehr oder weniger einheitlichen Mengen von Bildern entsprechen. Die Abhängigkeit der Modellparameter von der Lage der Knoten im Baum kann linear genähert werden. Dies erlaubt die Bildung eines weiteren, abstrakteren visuellen Alphabets von Ansichtsmodellen. Anstelle einer expliziten Berechnung der realen geometrischen Verhältnisse zwischen den Ansichtsmodellen, werden diese jedoch nur aufgezählt. Als biologische Motivation dienen die Erkenntnisse, daß Affen anscheinend die Zusammengehörigkeit beliebiger visueller Mustern erlernen können [Miy93], und daß Affen verschiedene Rotationslagen eines Objekts separat erlernen können und dazu jeweils unterschiedliche Teilmuster kombinieren [NLR08] (vgl. Abschnitt 3.1.8).

Auf der höchsten hier modellierten Hierarchiestufe werden schließlich verschiedene Ansichtsmodelle zu Kategorieknoten kombiniert. Verschiedene Strategien der Verknüpfung von Ansichtsmodellen erlauben die Auslegung des Modells auf verschiedene Erkennungsziele, beispielhaft einen hohen Anteil erkannter Objekte oder eine geringe Anzahl von falschen Treffern. Bei der Erstellung der

Kategorieknoten kommt die in dem visuellen Alphabet von Ansichtsmodellen vorhandene Redundanz der Zuverlässigkeit der Objekterkennung zugute.

Die durchgeführten Versuche zeigen, daß die Ortstoleranz, die für die geometrischen Teilebeziehungen zugelassen wird, von der Abstraktionsebene im Modell abhängt. Auf der Merkmalsebene kommen Merkmalspunkte zum Einsatz, deren Position z.T. mit Subpixelgenauigkeit bestimmt wird. Im Teile-Alphabet dürfen die Relativpositionen innerhalb von 7–21 Pixel variieren. Auf der Ansichtsebene erhöht sich die Toleranz auf 5–72 Pixel für die Positionen der Teile. Auf der Kategorie-Ebene des Modells spielt die Ortstoleranz keine Rolle mehr, da praktisch alle Ansichten an der gleichen Stelle erkannt werden. Für die in der Literaturdurchsicht herausgestellte Abwägung zwischen Constellation-Modell oder Bag-of-features muß daher die Abstraktionsebene berücksichtigt werden, auf der die Objektmodellierung geschieht.

Die Zuverlässigkeit der Objekterkennung wird anhand der Klassifikation einer Cartoon-Stichprobe nachgewiesen. Die Elemente der Stichprobe umfassen dabei nicht nur perspektivische Variationen, sondern auch Bewegungen und Mimik. Durch die Erzeugung spezieller Modelle für die verschiedenen Ansichten kann die Stichprobe trotz der starken Variation mit einer hohen Korrektklassifikationsrate von mindestens 78 Prozent erkannt werden. Der Ansatz ist dabei flexibel genug, um das Modell je nach Anwendungsbereich auf die Vermeidung falscher Treffer oder die Detektion vieler Objekte auszurichten. Bei gleicher Korrektklassifikationsrate wird so entweder ein positiver Vorhersagewert von 97 Prozent oder eine Sensitivität von 89 Prozent erreicht, jeweils zulasten des anderen Wertes. Der allgemeine Ansatz erreicht nicht die hohe Korrektklassifikationsrate von 92 Prozent des speziell auf die Cartoon-Datenbank zugeschnittenen Klassifikators. Der speziell zugeschnittene Klassifikator ist jedoch im Anwendungsbereich stark eingeschränkt und eignet sich nicht einmal zur Erkennung vergleichbarer Cartoons früherer Ausgaben. Der hier vorgestellte allgemeinere Ansatz ist dagegen auf vielfältiges Bildmaterial anwendbar. Die hier eingesetzten Merkmale wurden bereits erfolgreich zur Erkennung von Autos in realen Verkehrsszenen eingesetzt [SK05, SK06]. Die Integration weiterer Merkmale ist im übrigen ohne Änderungen am Modell oder Verfahren problemlos möglich. Darüberhinaus führt das Verfahren gleichzeitig zur Klassifikation eine Objektlokalisation durch, sodaß keine zusätzliche Nachverarbeitung nötig ist.

7.2 Ansätze für zukünftige Arbeiten

Im folgenden werden einige Aspekte der Objekterkennung und Modellerzeugung aufgegriffen, die als Ansatzpunkte für weiterführende Arbeiten in Frage kommen.

7.2.1 Auswertung der Statistik von Mehrheitsentscheiden

Derzeit wird ein Teil eines Objekts erkannt, wenn genügend viele der jeweiligen Unterteile erkannt wurden. Für die Objekterkennung ist auf der einen Seite

die Anzahl der erkannten Unterteile kritisch, da sie in direktem Zusammenhang zur Geschwindigkeit der Objekterkennung steht. Neben der Gesamtmenge an erkannten Teilen ist auch deren Verteilung relevant, da sie mitbestimmt, wie gut richtige von falschen Treffern unterschieden werden können. Histogramme über die Häufigkeit verschiedener Teile zeigen, daß eine sehr kleine Menge an Teilen einen sehr großen Anteil an der Gesamtmenge hat. Häufige Teile werden aufgrund ihrer geringeren Unterscheidungsleistung bei der Objekterkennung oft niedriger gewichtet [Sel01, NS98b, MLS06]. Die Berechnung der Teilegewichtung erfordert jedoch zahlreiche Multiplikationen, was die Objekterkennung stark verlangsamt. Statt mit einer Gewichtung kann ein Teil auch mit einem Wahrscheinlichkeitswert versehen werden, der es erlaubt, Teile abhängig von ihrer Häufigkeit von weiteren Mehrheitsentscheidungen auszuschließen. Aufgrund von komplexen Geometrieeinflüssen kann hier jedoch schwer abgeschätzt werden, wieviele Unterteile eines Teils zu dessen Erkennung erforderlich sind. Alternativ können besonders häufige Teile vollständig aus dem Modell ausgeschlossen werden. In Anbetracht der starken Ungleichverteilung der Teile sollte dies ohne Nachteile für die Güte der Objekterkennung möglich sein, unter Umständen ergibt sich sogar ein besseres Signal/Rausch-Verhältnis. Insbesondere wird jedoch mit einer starken Geschwindigkeitssteigerung aufgrund der geringeren Anzahl an zu verarbeitenden Teilen gerechnet. Die genauen Auswirkungen müssen experimentell ermittelt werden.

7.2.2 Bewertung von visuellen Alphabeten

Die aktuellen auf Teilen basierenden Objektmodelle verwenden zunehmend allgemeingültige visuelle Alphabete. Die Darstellungskapazitäten schwanken dabei jedoch erheblich zwischen einigen zehn [FPZ07] und mehreren Millionen [SOP07] gespeicherten Mustern. Da sich die in den Modellen gespeicherten visuellen Alphabete zunehmend als kritischer Faktor für die Objekterkennung herausstellen, stellt sich die Frage nach entsprechenden Bewertungsmaßen. Dazu bieten sich die informationstheoretische Einheit Bit oder die in der Neurophysiologie verwendete Spärlichkeit [RTT97, FRAJ07] an. Letztere ist speziell in Hinblick auf die Reaktion von Zellverbänden auf visuelle Stimuli ausgelegt. Diese enthalten in der Regel ein gewisses Maß an Redundanz, so daß hier insbesondere der Fall auftritt, daß mehrere Zellen auf ein Muster ansprechen. Aufgrund der Parallelen zwischen Zellverbänden im Gehirn und Klassifikatoren für visuelle Alphabete im Rechner bietet es sich an, die neurophysiologischen Kriterien auf ihre Eignung zur Beschreibung technischer Objekterkennungssysteme zu prüfen.

Um mit dem Spärlichkeitsmaß zu arbeiten, muß dieses sowie geeignete Alternativen in das bestehende System integriert werden. Anschließend muß die Beschreibbarkeit der während eines Modelltrainings erzeugten visuellen Alphabete durch die neuen Ausdrucksmittel geprüft werden. Die neuen Bewertungsmaße stellen eine Bereicherung für die Interpretation von Experimenten dar.

7.2.3 Dekorrelation von Teil-Ganzes-Beziehungen im visuellen Alphabet

Die Qualität der Objekterkennung hängt entscheidend davon ab, daß in Hintergrundbildern, die nicht das modellierte Objekt zeigen, keine falschen Treffer erzeugt werden. Bei Systemen auf Basis von Mehrheitsentscheiden wird angenommen, daß die Teile eines Objekts mit einer gewissen (in guten Fällen niedrigen) Wahrscheinlichkeit gleichverteilt über der Bildebene auftreten. Grimson und Huttenlocher [GH88] zeigen, daß bei einer unkorrelierten Teiledetektion die Summe der an einer bestimmten Bildposition erkannten Teile binomialverteilt ist und von der Anzahl der Teile des Modells und der Wahrscheinlichkeit falscher Treffer für einzelne Teile abhängt. Dies ist für geringe Standardabweichungen eine gute Nachricht, da die Binomialverteilung abseits des Maximums steil fällt und bald für eine zuverlässige Objekterkennung ausreichend niedrige Trefferwahrscheinlichkeiten liefert. Laut Grimson und Huttenlocher weicht die tatsächliche Verteilung erkannter Teile in dem Maße von der idealen Verteilung ab, in dem statistische Abhängigkeiten zwischen den Detektionen verschiedener Teile bestehen. Diese zeigen sich darin, daß bestimmte Kombinationen von Teilen gehäuft an einer einzelnen Bildkoordinate detektiert werden.

Tatsächlich ähneln die Trefferverteilungen des Versuchssystems grob einer Binomialverteilung, allerdings ist die Standardabweichung größer als theoretisch zu erwarten. Um die Zahl falscher Treffer zu reduzieren, sollen die Elemente des visuellen Alphabets dekorreliert werden, d.h. das systematische, gleichzeitige Auftreten unterschiedlicher Teile soll verringert werden. Eine Berechnung der gegenseitigen Teilekompatibilität ergab, daß das erzeugte visuelle Alphabet derzeit Korrelationen in Form von Elementen enthält, die zueinander in einer Teil-Ganzes-Beziehung stehen (beispielsweise ist das Minuszeichen zweimal im Gleichheitszeichen enthalten). Um zu verhindern, daß ein allgemeineres Muster immer dann erkannt wird, wenn auch das jeweils speziellere Muster erkannt wird, könnten hemmende Pfade in das Modell integriert werden. Die Erkennung des spezielleren Musters kann dann das Ansprechen des allgemeineren Klassifikators verhindern.

Dazu muß sowohl das Modell als auch der Klassifikator angepaßt werden. Teil-Ganzes-Beziehungen im visuellen Alphabet können während des Trainings leicht erkannt werden. Aufwendiger wird die Festlegung des Umfangs der gegenseitigen Unterdrückung. Die Auswirkung auf die Objekterkennung müssen anschließend experimentell bestimmt werden.

7.2.4 Untersuchung von Teile-Abhängigkeiten durch Dekomposition

Wie Kapitel 6.2 gezeigt hat, ist der konstruktive Ansatz, Modelle durch schrittweise Hinzunahme und Optimierung einzelner weiterer Teile zu erzeugen, aufgrund suboptimaler lokaler Maxima im Konfigurationsraum der Modelle problematisch. Das Problem der schlechten Konvergenz konstruktiver Methoden ist auch in der Neurophysiologie bekannt: In dem Experiment von Albright und

Gross [AG90] war es nicht möglich, visuelle Stimuli für Zellen des inferioren Temporallappens aus einfacheren Merkmalen zu kombinieren. Der umgekehrte Weg, komplexe Stimuli in einfachere zu zerlegen, führte dagegen zur Entdeckung eines visuellen Alphabets moderat komplexer Teile [Tan96].

Als Erweiterung dieser Arbeit könnte daher ebenfalls ein dekompositioneller Ansatz überprüft werden. Dazu wird ein Modell durch den in Kapitel 6.2 erwähnten genetischen Algorithmus trainiert. Aus diesem Modell werden anschließend Teile entfernt, wobei jeweils die Klassifikationsgüte für Verschiebungen von Teilen bestimmt wird. Auf diese Weise kann untersucht werden, wie sich das durch den genetischen Algorithmus gefundene gute Maximum im Konfigurationsraum von den suboptimalen Maxima des konstruktiven Ansatzes unterscheidet. Wie Kapitel 6.2 zeigt, sind solche Einblicke in den Konfigurationsraum sehr aufschlußreich.

Anhang A

Symbolverzeichnis

A.1 Symbole zu Kapitel 3

$\boldsymbol{A} = (a_1, a_2, ...)$	Die stochastisch unabhängigen Attribute der Teile im Constellation-Modell. Die Attribute beschreiben in erster Linie die lokale Bildstruktur, die durch ein Teil repräsentiert wird. Es wird meistens vereinfacht angenommen, daß zwischen den Bildstrukturen verschiedener Teile eines Objekts kein statistischer Zusammenhang besteht. Dies vereinfacht die Auswahl von Bildregionen zur Modellierung als Teil.
$\boldsymbol{A}_I$	Die Attribute von Teilekandidaten in einem bestimmten Bild I.
$\boldsymbol{A}_M$	Die Attribute der Teile eines Modells M.
$\boldsymbol{A}^*, \boldsymbol{L}^*$	Optimale Modellattribute. Durch das Training des Modells werden sowohl die als Teil repräsentierten Bildregionen als auch die Relativpositionen der Teile so gewählt, daß das Modell die anhand einer Stichprobe vorgegebenen Klassen sensibel und trennscharf erkennt.
$\mathfrak{A}$	Vektorförmiges Attribut eines Terminal- oder Nichtterminalzeichens einer in der syntaktischen Mustererkennung nach Fu [Fu82] eingesetzten Grammatik G. Die Attribute parametrisieren geometrische Objekte.
α	Viola und Jones [VJ04] berechnen per Boosting einen Klassifikator, der sich aus einer Reihe von Einzelklassifikatoren zusammensetzt. Diese werden jeweils mit einem individuellen Vorfaktor α gewichtet. Die einzelnen Klassifikatoren detektieren die für eine positive Erkennung erforderlichen lokalen Bildmerkmale.

B	Die Teileabhängigkeiten im Constellation-Modell. Insbesondere werden hier die Relativpositionen der verschiedenen Teile in der Bildebene modelliert.
$b_{ij} \in B$	Die zumeist geometrische Abhängigkeit zwischen zwei Teilen i und j kann beispielsweise als Histogramm des gemeinsamen Auftretens über der Bildebene gespeichert werden.
$\mathfrak{c}$	Länge einer Clique in einem Graphen. Crandall et al. [CFH05] schlagen diesen Begriff vor, um die Komplexität einer Teile-Anordnung im Constellation-Modell zu messen.
Γ	Ommer und Buhmann [OB06] bezeichnen als Komposition Γ eine Gruppe von räumlich benachbart auftretenden Deskriptoren oder auch eine Gruppe, die selbst wieder aus mehreren Kompositionen besteht.
$\boldsymbol{e}$	Ein Eigenvektor. Murase und Nayar [NMN96] benutzen Eigenvektoren, um die Bilder einer Stichprobe in einem im Vergleich zu den Originalbildern niedrigdimensionalen Raum darzustellen.
$\varepsilon(\dots)$	Viola und Jones [VJ04] berechnen den Klassifikationsfehler, indem sie die Ausgabe des Klassifikators mit der jeweiligen Sollklasse der Elemente einer Stichprobe vergleichen. Die Stichprobenelemente sind dabei unterschiedlich gewichtet, damit der Klassifikationsfehler im Verlauf des Trainings die noch zu lernenden Elemente anzeigt.
f	Ein lokales Merkmal. Viola und Jones [VJ04] erreichen durch den Einsatz von Integral-Bildern eine besonders schnelle Merkmalsextraktion. Die Merkmale reagieren auf Helligkeitsunterschiede zwischen jeweils zwei bis vier aneinandergrenzenden, rechteckigen Bildregionen.
$\boldsymbol{f}$	Symbol für ein Bild in der von Murase und Nayar [NMN96] gewählten alternativen Bilddarstellung.
$\mathfrak{f}$	Eine Klassifikationsfunktion nach Viola und Jones [VJ04], die für jede Bildposition anzeigt, ob dort ein bestimmtes Merkmale gefunden wurde oder nicht.
G	Eine Grammatik $(V_n, V_t, P, \mathfrak{K})$ zur Szenenmodellierung bestehend aus den Nichtterminalzeichen V_n, den Terminalzeichen V_t, den Produktionsregeln P und den Konfigurationen $\mathfrak{K}$. Terminalzeichen sind Primitive einer Szene, beispielsweise Kanten oder Rechtecken. Die Produktionsregeln beschreiben, wie die Primitive zu Szenen zusammengesetzt werden. Szenen werden als Konfigurationen bezeichnet.

g, q	Funktionen auf Attributen von geometrischen Objekten bei der syntaktischen Mustererkennung nach Fu [Fu82]. Die Funktionen sichern die geometrische Konsistenz eines Modells, indem sie bestimmte Bedingungen für die Parametrisierungen der Objekte einer Szene formulieren. Zudem unterstützen sie den Erkennungsprozeß, indem sie Zustandsinformationen über bereits behandelte Objekte bereitstellen.
H	Die Entropie (Informationstheorie)
$\boldsymbol{h}$	Hypothese über ein mögliches Objekt. Da in einem Bild nicht die gleiche Anzahl an Teilen wie im Modell vorliegt oder Teile aufgrund von Rauschen nicht zuverlässig erkannt werden, kommen mehrere Zuordnungen von detektierten Teilen zu Teilen des Modells in Frage.
$\mathfrak{H}$	Die Gesamtmenge einzelner Hypothesen $\boldsymbol{h}$.
$\theta_1, \theta_2, \dots$	Parameter des visuellen Arbeitsbereichs nach Murase und Nayar [NMN96]. Der visuelle Arbeitsbereich umfaßt alle in einem bestimmten Anwendungsbereich möglichen Objekterscheinungen. Die für den Anwendungsbereich relevanten Größen, welche die Objekterscheinung beeinflussen, sind die Parameter des visuellen Arbeitsbereichs. Diese beschreiben beispielsweise die Lage eines Objekts relativ zur Lichtquelle.
ϑ	Ein Schwellwert zur Merkmalsextraktion, eingeführt in der Beschreibung des Objekterkennungssystems von Viola und Jones [VJ04].
$\boldsymbol{I}, \boldsymbol{I}_1, \boldsymbol{I}_2, \dots$	Einzelne Bilder einer Stichprobe. Ein Bild wird als Vektor geschrieben, der sich aus den Intensitätswerten der einzelnen Bildpunkte zusammensetzt.
$\boldsymbol{I}_\Sigma$	Ein über mehrere Bilder gemitteltes Bild.
$\iota_1, \iota_2, \dots$	Einzelne Punkte eines Bildes, insbesondere deren Intensität. Ein fortlaufender Index zählt die Bildkoordinaten ab.
$\mathfrak{K}$	Konfigurationen der Grammatik G zur syntaktischen Mustererkennung [Fu82]. Die Konfigurationen stellen die in der Grammatik möglichen Regelableitungen dar und entsprechen damit den modellierbaren Szenen.
$\boldsymbol{L} = (l_1, l_2, \dots)$	Die stochastisch abhängigen Attribute im Constellation-Modell. Dies ist eine von Crandall et al. [CFH05] getroffene Verallgemeinerung, die in erster Linie auf die Bildkoordinaten der modellierten Teile abzielt. Die Modellierung der relativen Teilepositionen ist der Kerngedanke des Constellation-Modells.
$\boldsymbol{L}_M$	Die stochastisch abhängigen Attribute eines Modells M.
$\boldsymbol{L}_I$	Die Attribute der in einem Bild I gefundenen Teilekandidaten.

Λ	Eine Anzahl an Eigenvektoren.
λ	Ein Eigenwert.
$\boldsymbol{M}$	Ein Modell für Objekte. Im Constellation-Modell besteht ein Modell beispielsweise aus einem Vektor aus Teilen, ihren insbesondere geometrischen Beziehungen B, den Teilepositionen $\boldsymbol{L}$ und den als Teil modellierten Bildregionen $\boldsymbol{A}$.
$\boldsymbol{M}_H$	Ein Modell für den Hintergrund. Dieses faßt die statistischen Eigenschaften aller Bildregionen, die kein zu erkennendes Objekt darstellen, zusammen.
m	Im Constellation-Modell wird ein Objekt als eine Menge von m Teilen beschrieben, die in der Bildebene in einer bestimmten geometrischen Anordnung vorliegen.
ν	Ein Terminal- oder Nichtterminalzeichen der Grammatik G.
n_B	Anzahl der Beschränkungen auf einer Produktionsregel der Grammatik G.
n_C	Anzahl Akkumulatorzellen bei der Hough-Transformation. Die Anzahl hängt davon ab, wie fein die Parameter der zu erkennenden Objekte quantisiert werden können.
n_I	Die Anzahl der Bilder in einer Stichprobe.
n_x	Die Anzahl an Punkten in einem Bild (= Breite x Höhe).
$n_{\mathfrak{H}}$	Die Anzahl an Hypothesen bei der Objekterkennung durch die verallgemeinerte Hough-Transformation [Bal81]. Eine Hypothese basiert auf der Erkennung eines Merkmals eines Objekts. Das gefundene Merkmal wird in der R-Tabelle nachgeschlagen. Der Wert der dort angegebenen Akkumulatorzelle wird daraufhin erhöht. Wenn sich eine ausreichende Menge an Hypothesen in bestimmten einem Objekt zugeordneten Akkumulatorzellen angehäuft hat, gilt ein Objekt als erkannt. Da fälschlicherweise erkannte Merkmale nicht mit der in der R-Tabelle kodierten Objektstruktur übereinstimmen, geht man davon aus, daß sich Rauschen gleichmäßig auf den Akkumulator verteilt, statt sich in einzelnen Zellen anzuhäufen.
$\mathfrak{n}$	Beim Vergleich eines teilebasierten Modells mit einem Bild wird oft zunächst eine Menge von $\mathfrak{n}$ Teilepositionen bestimmt, die aufgrund bestimmter Eigenschaften besonders aussichtsreich für einen näheren Vergleich mit den Teilen des Modells sind. Dies reduziert die Anzahl der Vergleiche zwischen dem Bild und dem Modell, da ansonsten alle Bildpositionen mit dem Modell verglichen werden müßten.
$\mathcal{O}(n)$	Die Komplexität der Ordnung n.

$P = \mathfrak{p}_1, \mathfrak{p}_2, \ldots$	Produktionsregeln der Grammatik G zur Szenenbeschreibung. Sie beschreiben, wie geometrische Primitive zu komplexeren Objekten zusammengesetzt werden. Die Produktionsregeln werden durch Beschränkungen auf den Attributwerten der zusammengesetzten Teile ergänzt. Diese sichern sinnvolle und konsistente Szenenparametrisierungen.
$p(a\|b)$	Die Wahrscheinlichkeit für das Auftreten von a unter der Bedingung des Auftretens von b.
ϖ	Die Polarität $+1, -1$. Ein Vorfaktor zur kompakteren Darstellung der Merkmalsextraktion durch Viola und Jones [VJ04].
Q	Eine Kovarianzmatrix. Murase und Nayar [NMN96] berechnen eine Kovarianzmatrix über eine Bildermenge. Sie stellen so fest, welche Punkte in den Bildern einer Stichprobe voneinander abhängen, um eine kompaktere Darstellung zu bestimmen.
$R, \bar{R}$	Menge von Referenzteilen (R) und Nicht-Referenzteilen ($\bar{R}$) zur Beschreibung der Teile-Abhängigkeiten im Constellation-Modell. Die Referenzteile bilden eine Clique, d.h. sie sind alle gegenseitig voneinander abhängig. Die Nicht-Referenzteile sind untereinander unabhängig, hängen allerdings jeweils von allen Referenzteilen ab.
$\mathbf{r}$	Vektor von einem Kantenpunkt zum Referenzpunkt des zugehörigen Objekts bei der verallgemeinerten Hough-Transformation. Die Menge solcher Vektoren wird in Form der R-Tabelle zusammengefaßt und kodiert die Struktur eines Objekts.
ρ	Im Constellation-Modell dient der Wahrscheinlichkeitsquotient ρ der Objektdetektion. Dieser berechnet sich aus der Wahrscheinlichkeit, daß die zu einem Bild extrahierten Merkmale zu einer bestimmten Objektklasse gehören, dividiert durch die Wahrscheinlichkeit, daß die Merkmale zum Hintergrund gehören. Wenn der Quotient größer als Eins ist, gilt ein Objekt als erkannt.
S	Die potenzielle Klassenzugehörigkeit eines Teils aus einem Modell. Vor dem Berechnen der Klassenzugehörigkeit im Verlauf des Trainings betrachten Ommer und Buhmann [OB06] diese als Zufallsvariable.
s	Die während des Trainings festgelegte Klassenzugehörigkeit eines Teils aus einem Modell.
σ	Eine Standardabweichung - je nach Zusammenhang ein Meßwert aus einer Stichprobe oder ein Verfahrensparameter.

$\boldsymbol{T} = (t_1, t_2, \dots)$	Die Teile $t_1, t_2, \dots$ des Constellation-Modells als Vektor zusammengefaßt. In den gängigen Varianten des Constellation-Modells werden lokale Bildregionen als Teil bezeichnet. Gute Teile werden häufig mit Hilfe eines Lernverfahrens aus einer Stichprobe abstrahiert. Die gefundenen Teile werden dann beispielsweise durch lokale Farb- und Gradientenstatistiken [CH06] oder prototypische Deskriptoren eines Codebooks [LMS06a] beschrieben.
t_r	Ein Referenzteil einer Teilekonstellation. Das Constellation-Modell [WWP00] zielt insbesondere auf die Modellierung der geometrischen Anordnung von Objektteilen in der Bildebene ab. Die Teilepositionen werden dabei relativ zueinander angegeben. Je nach Komplexität des betrachteten Modells werden die Teilepositionen relativ zu einem oder zu mehreren als Referenz ausgezeichneten Teilen angegeben.
V_n	Eine Menge von Nichtterminalzeichen aus der Grammatik G.
V_t	Eine Menge von Terminalzeichen aus der Grammatik G.
φ	Bezeichner für eine Kantenrichtung. Eingeführt in der Beschreibung der Hough-Transformation.
w	Ein Gewichtungswert. Viola und Jones [VJ04] gewichten die Elemente ihrer Trainingstichprobe unterschiedlich, um das Training auf die noch nicht gelernten Stichprobenelemente zu konzentrieren.
X	Eine Bildermatrix. Die Spalten der Matrix repräsentieren jeweils einzelne Bilder. Eine Zeile der Matrix entspricht einer Bildkoordinate. Die Werte der Matrix sind die Intensitätswerte der Bilder. Murase und Nayar [NMN96] erzeugen eine Bildermatrix als Zwischenergebnis bei der Berechnung des universellen Eigenraums.
$\boldsymbol{x} = (x, y)$	Eine Bildkoordinate.

A.2 Symbole zu den Kapiteln 2 und 4 bis 6

A, A'	Kleine Flächen innerhalb eines Pixels.
$\mathfrak{b}$	Eine binäre Funktion, zur Abfrage der detektierten Unterknoten U eines Teils t. Die Funktion liefert für eine Bildkoordinate x, y und den Index c des Unterknotens $u_c \in U$ den Wert Eins zurück, falls der Unterknoten innerhalb der Ortstoleranz um die im 'Idealposition' für den Unterknoten liegt. Ansonsten liefert die Funktion Null zurück. Die 'Idealposition' ist dabei die Position, an der eine Detektion des Unterknotens zu erwarten ist, wenn für das komplette Teil t die Position x, y angenommen wird. Die 'Idealposition' ergibt sich daher aus der angegebenen Bildkoordinate zuzüglich der Relativposition des Unterknotens gegenüber dem Teilezentrum. Eine Detektion des Unterknotens u_c an der 'Idealposition' wird daher als Beleg gewertet, daß an der Position x, y das Teil t vorliegt.
β	Eine Menge von Vektoren. In der Beschreibung des Clusterungsverfahrens (Abschnitt 6.2.5) bezeichnet β eine Menge von benachbarten Vektoren, die zu einem Cluster zusammengefaßt wurden. Die Cluster werden vereinfacht durch den Mittelwert $\boldsymbol{v}$ der Vektoren in β beschrieben.
c	Ein Index in einer Liste von Unterknoten. Beispiel: $u_c, c = 3$ bezeichnet den dritten Unterknoten in U.
γ	Eine Abstandsnorm, z.B. 1=Häuserblockabstand, 2=euklidische Norm. Abstände werden gemäß der bekannten Gleichung $L_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \sqrt[\gamma]{\sum_i \lvert x_i - y_i \rvert^\gamma}, \boldsymbol{x} = (x_1, x_2, \dots), \boldsymbol{y} = (y_1, y_2, \dots)$ berechnet.
$\boldsymbol{D}$	Der Deskriptor eines lokalen Merkmals. In dieser Arbeit haben die Deskriptoren maximal drei Dimensionen und speichern eine Gradientenrichtung oder die Intensität von Flächenmittelpunkten, die Orientierung von Skelettlinien und den Abstand zur nächsten extrahierten Kante.
$\tilde{\boldsymbol{D}}$	Die direkt zu einem Deskriptor $\boldsymbol{D}$ benachbarten Deskriptoren. Diese ergeben sich durch Addition bzw. Subtraktion jeweils einer Quantisierungsstufe auf die einzelnen Deskriptorkomponenten. Sie werden anstelle des Ursprunglichen Deskriptors eingesetzt, um die Toleranz gegen Rauschen zu verbessern.
$d_1, d_2, d_3, \dots$	Die einzelnen Komponenten eines Deskriptors.

$\mathfrak{d}$	Die Tiefe eines Knotens im Modell. Der Wurzelknoten hat die Tiefe Null. Wenn es mehrere Wege von dem Knoten zu einem Wurzelknoten gibt, hat $\mathfrak{d}$ den maximalen Abstand von einem Wurzelknoten über alle vorhandenen Pfade.
$\partial\iota/\partial x$	Die partielle Ableitung von ι nach x.
∂r	Eine kurze Strecke in einem Bild, eingeführt für die Abstandstransformation.
E	Eine Menge von Punkten.
$\zeta_1, \zeta_2, \zeta_3, \ldots$	Die Anzahl an Quantisierungsstufen zur Diskretisierung eines Deskriptors. Die Diskretisierung erlaubt das Abzählen der auftretenden Merkmalsausprägungen und die Einordnung von Merkmalen als Unterknoten in das Teilemodells.
$\boldsymbol{h}$	Eine Hypothese der Form $(\kappa_j, x, y, c, \mathfrak{d})$, welche das Auftreten des durch den Schlüssel κ_j bezeichneten Teils t_j an der Position (x, y) vorhersagt. Die Werte c und $\mathfrak{d}$ dienen dem vereinfachten Zugriff auf das Modell und bezeichnen den Index des Unterknotens, der zur Erzeugung der Hypothese geführt hat, und die Hierarchieebene des vorhergesagten Knotens.
$\mathfrak{H}_j$	Die Menge aller Hypothesen, die einen Knoten t_j vorhersagen.
ϑ	Ein Schwellwert zur Detektion eines Knotens. Der Schwellwert gibt an, wieviele Unterknoten eines Teils erkannt werden müssen, damit das Teil selbst erkannt wird.
$\vartheta_\nabla, \vartheta_r$	Schwellwerte werden auch zur Abstandstransformation benötigt. Der Schwellwert ϑ_∇ stellt sicher, daß der Kantenabstand auf einer Skelettlinie ein lokales Minimum bildet, d.h. daß die Änderung des Abstands zumindest sehr klein ist. Der Schwellwert ϑ_r verhindert die Berechnung von Skelettlinien in der direkten Nachbarschaft einer Kante, da die Berechnung dort durch Rauschen zu stark beeinträchtigt wird.

ϑ_Farbe	Ein Schwellwert für den Farbklassifikator in Kapitel 2. Der Schwellwert gibt an, welcher Anteil der als Vordergrund klassifizierten Bildpunkte auf eine bestimmte Farbe entfallen muß, damit ein Stichprobenbild als ein bestimmtes Objekt erkannt wird. Diese Regel basiert auf der Beobachtung, daß ein Objekt typischerweise Punkte in mehreren Farben enthält, und diese in einem typischen Verhältnis zueinander stehen. Die Schwellwerte verhindern somit, daß eine in einem Bild fehlende Farbe durch ein häufigeres Auftreten einer anderen Farbe kompensiert werden kann.
I_A	Eine quadratische Matrix mit den gegenseitigen Abständen aller Elemente einer Menge von Vektoren. Die Matrix hat so viele Spalten und Zeilen wie die Menge Vektoren enthält. Jeder Zeile und jeder Spalte ist ein Vektor der Menge zugeordnet. Ein Matrizenelement $\iota_{i,j}$ der Spalte i und Zeile j gibt den Abstand der entsprechenden Vektoren zurück.
$I_\cap, I_E$	Eine Matrix mit den gemeinsamen Trefferpositionen eines bestimmten Teils über mehrere Bilder. Ein Wert $\iota_{i,j} = 1$ an der Koordinate i, j der Matrix zeigt an, daß das Teil in allen Bildern an der Position i, j gefunden wurde. Andernfalls hat ι den Wert Null.
I_D	Ein dilatiertes Trefferbild. Ein Trefferbild zeigt als Binärbild alle Bildpositionen an, an denen ein bestimmtes Teil detektiert wurde. Das dilatierte Trefferbild zeigt alle Positionen an, in deren Nähe ein bestimmtes Teil detektiert wurde. Dazu wird auf dem Trefferbild eine Dilatation mit einer quadratischen Maske mit Radius ς durchgeführt. Alle Werte der Maske sind auf Eins gesetzt. Die Dilatation mit dem quadratischen Muster wird aus Geschwindigkeitsgründen durchgeführt, da sie sich effizienter umsetzen läßt als die Dilatation mit einer Kreisfläche.
$I_\mathfrak{D}, I_h, I_v$	Drei Matrizen in Bildgröße. Die Koordinaten der Matrizen entsprechen den Bildkoordinaten. Die Matrix I_D gibt zu jeder Koordinate an, wieviele Pixel sie von der nächsten Koordinate eines Kantenpunktes entfernt ist. Die Koordinate der nächsten Kantenpunkte wird in I_h und I_v gespeichert. Dabei speichert I_h den horizontalen Anteil und I_v den vertikalen Anteil der Koordinate.

I_K	Eine Kandidatenmatrix, welche die Abdeckung einer Menge von Mustern durch ein visuelles Alphabet auf Teile-Ebene angibt. Das Element $\iota_{i,j}$ der Spalte i und Zeile j der Kandidatenmatrix zeigt durch den Wert 1 an, daß ein Muster mit den Index j durch einen Teilekandidaten mit dem Index i erkannt wird. Andernfalls hat ι den Wert Null. Die Kandidatenmatrix wird in Abschnitt 6.3.1 eingeführt und in Abschnitt 6.3.2 berechnet.
I_M	Eine Matrix in Bildgröße, in der alle Positionen markiert sind, an denen ein bestimmter Modellknoten erkannt wird. Für eine solche Matrix wird in dieser Arbeit der Begriff *Trefferbild* gewählt. Trefferbilder werden als Binärbilder implementiert, die morphologische Operationen ermöglichen.
I_N, I_P	Binäre Matrix, die für jedes negative (I_N) bzw. positive (I_P) Stichprobenelement angibt, durch welche Ansichtsmodelle es erkannt wird. Die Spalten entsprechen den Stichprobenelementen, die Zeilen den Ansichtsmodellen.
I_S	Die elementweise Summe mehrerer Trefferbilder. Die Trefferbilder enthalten nur die Werte Null und Eins. Im konkreten Fall wird über die Bilder summiert, welche die Detektion eines bestimmten gemeinsamen Teils anzeigen. Wenn sich für eine bestimmte Koordinate eine Summe ergibt, die der Anzahl der summierten Trefferbildern entspricht, bedeutet dies, daß das betrachtete Teil in allen Bildern an der gleichen Position erkannt wird.
I_T	Eine Teilematrix. Diese gibt an, welche Teile in welchen Stichprobenelementen erkannt werden. Wenn ein Element $\iota_{i,j}$ der Teilematrix den Wert 1 oder -1 hat, bedeutet das, daß ein Teil i in einem Stichprobenbild j gefunden wurde. Die Teilematrix enthält positive Werte für Stichprobenelemente, die das gesuchte Objekt enthalten, und negative Werte für Stichprobenelemente, die nur den Hintergrund, aber nicht das Objekt zeigen. Wenn mehrere Stichprobenelemente die gleichen Teile enthalten, wird angenommen, daß es sich um verwandte Ansichten eines Objekt handelt, die gemeinsam modelliert werden können.

I_U	Eine unterabgetastete Version der Abstandsmatrix I_A. Die Matrix I_U besitzt um den Faktor n_U weniger Spalten als die ursprüngliche Matrix. Die unterabgetastete Matrix wird zur schnellen Bestimmung des Minimums der Matrix I_A eingesetzt (siehe Clusterungsverfahren, Abschnitt 6.2.5). Dazu wird I_U mit Werten belegt, die höchstens so groß sind wie der Minimalwert der entsprechenden Zellen in I_A. Mit Hilfe der unterabgetasteten Matrix kann daher schnell entschieden werden, ob ein Teilbereich der Matrix I_A für die Bestimmung des Minimalwerts interessant ist. Uninteressante Bereiche können übersprungen werden.
ι, z.B. ι_{i,j,I_A}	Je nach Zusammenhang entweder ein Matrixelement, z.B. das Element an der Position i, j der Matrix I_A, oder auch konkret die Intensität eines Bildpunkts. Die Variable ι wird immer dort verwendet, wo in mehr oder weniger abstrakter Form mit Bilddaten gearbeitet wird, und der genaue Typ der Bilddaten zwar klar ist, aber nicht die Kernaussage des Textes darstellt.
κ	Ein Schlüssel zur Identifikation eines Knotens des Teile-Modells, eingeführt in Kapitel 4. Für Merkmals-Knoten dient der Merkmalsdeskriptor als Schlüssel, da identische Merkmale bei der Objekterkennung nicht voneinander unterschieden werden müssen. Bei den komplexen Knoten des Modells entspricht der Schlüssel einer laufenden Nummer und ist für das gesamte Modell eindeutig.
L, L_i	Die Positionen der Unterknoten eines Teils im allgemeinen bzw. eines bestimmten Teils t_i. Die Position eines Teils wird per Vereinbarung in den Schwerpunkt der Positionen der Unterknoten gelegt. Die Positionen der Unterknoten werden in karthesischen Koordinaten relativ zum Schwerpunkt angegeben.
LUT	Eine Tabelle, die zu jedem Knoten des Modells eine Liste der Elternknoten angibt. Sie entspricht dem Gedanken nach der R-Tabelle in der erweiterten Hough-Transformation und dient die Erzeugung von Hypothesen über das Auftreten von abstrakteren Teilen.
m, m_i	Die Anzahl der Unterknoten eines Knotens im allgemeinen bzw. des Teils t_i.

$\mathfrak{m} = (\mathfrak{m}_1, \mathfrak{m}_2, \mathfrak{m}_3, \dots)$	Ein Vektor mit Zahlen, der eine Indexumbenennung der Form $neu = \mathfrak{m}_{alt}$ erlaubt. Die Indexumbenennung wird in Gleichung 6.13 eingeführt, um eine Abstandsmatrix über beliebige Vektoren einer Menge zusammenzustellen. Einer Spalte oder Zeile mit dem Index i der Abstandsmatrix kann mit Hilfe der Übersetzungstabelle $\mathfrak{m}$ ein Vektor mit dem Index $\mathfrak{m}_i$ zugeordnet werden. Die Indices der Vektoren in der Menge müssen daher nicht mit den Spalten oder Zeilen der Abstandsmatrix übereinstimmen. Das hat keine besondere konzeptionelle Auswirkung, sondern beschreibt nur eine rechnentechnisch vorteilhafte Implementierung: Es muß kein Speicher freigegeben oder belegt werden, um Neuzuordnungen von Vektoren zur Abstandsmatrix vorzunehmen.
n_A	Die Anzahl an Bildern, die eine bestimmte Objektansicht, d.h. eine bestimmte Pose des Objekts aus einer bestimmten Perspektive, darstellen. Die Bilder, die zu einer bestimmten Objektansicht gehören, werden dadurch erkannt, daß in ihnen die gleichen Teile detektiert werden.
n_D	Die Dimensionalität von Merkmalsdeskriptoren oder im Zusammenhang mit Clusterungsverfahren die Dimensionalität der Eingabevektoren der Clusterung.
n_E	Die Anzahl von Vektoren in einer Menge. Diese können Bildkoordinaten sein oder die Eingabeelemente einer Clusterung darstellen.
n_P	Die Anzahl der positiven Elemente in der Stichprobe, d.h. der Stichprobenbilder, die das zu erkennende Objekt zeigen.
$n_{\mathfrak{P}}$	Die Anzahl der Ansichtsmodelle, die ein bestimmtes positives Stichprobenelement erkennen. Der Wert wird in Abschnitt 6.5.1 aus der Ansichtsmatrix I_P berechnet.
n_N	Die Anzahl der negativen Elemente der Stichprobe, d.h. die Anzahl der Hintergrundbilder, die das gesuchte Objekt nicht enthalten.

$n_{\mathfrak{N}}$	Die maximale Anzahl von Ansichtsmodellen eines Kategorieknotens, die auf ein beliebiges negatives Stichprobenelement anspricht. Zur Berechnung von $n_{\mathfrak{N}}$ werden zuerst alle Ansichtsmodelle gesucht und zu einem Kategorieknoten zusammengefaßt, welche ein bestimmtes Positivbeispiel der Stichprobe erkennen. Da die einzelnen Ansichtsmodelle keine perfekte Trennschärfe besitzen, sprechen sie in einem gewissen Maße auch auf Negativbeispiele der Stichprobe an. Daher wird die Anzahl der Ansichtsmodelle, die fälschlicherweise Treffer liefert, für jedes Negativbeispiel ermittelt. Der dabei auftretende Maximalwert wird in der Variablen $n_{\mathfrak{N}}$ gespeichert. Für die zu unterschiedlichen Positivbeispielen gebildeten Kategorieknoten ergeben sich verschiedene Werte von $n_{\mathfrak{N}}$.
n_R	Charakteristische Objektansichten werden dadurch erkannt, daß in einer bestimmten Untermenge der Stichprobe die gleichen Teile detektiert werden. Der Wert n_R gibt die Anzahl der gemeinsamen Teile einer Objektansicht an.
n_T	Die Anzahl der Teile im visuellen Alphabet auf Teile-Ebene. Mit Hilfe des Alphabets auf Teile-Ebene werden verschiedene Objektansichten modelliert.
n_U	Ein Abtastverhältnis. Eingeführt wird n_U als Skalierungsfaktor zwischen einer Abstandsmatrix I_A und einer unterabgetasteten Variante I_U. Letztere erlaubt eine schnelle Bestimmung des Minimums in I_A. Die Methode beschleunigt die in Abschnitt 6.2.5 beschriebene Clusterung stark. Dabei ist der genaue Wert von n_U jedoch unkritisch.
r_b, r_d, r_e, r_p	Verschiedene Längen von Strecken in einem Bild zur Bestimmung der subpixelgenauen Lage einer Kante.
r_x, r_y, r_z	Weitere Längen von Strecken in einem Bild. Bei der Erläuterung der Abstandstransformation bezeichnen r_x und r_z die Abstände von Punkten $\boldsymbol{x}$ und $\boldsymbol{z}$ von der nächsten Kante. Für einen Punkt $\boldsymbol{x}$ ist der Abstand dabei erst noch zu berechnen, wohingegen der Kantenabstand für den Nachbarpunkt $\boldsymbol{z}$ bereits bekannt ist. Der Abstand r_x wird dann aus den Abständen der Nachbarpunkte hergeleitet. Dabei werden unter Umständen unterschiedliche Kantenpunkte $\boldsymbol{y}$ untersucht, die jeweils einen Abstand r_y von $\boldsymbol{x}$ haben.
r_t	Der Durchmesser eines Teils eines Objekts in einem Bild gemessen in Pixeln. Abbildung 6.47 zeigt die Verteilung der Größen von Teilekandidaten vor der Aufnahme in das Modell.

r_a	Der Durchmesser einer Objektansicht bzw. eines Beispielobjekts in einem Bild gemessen in Pixeln.
ς	Die Ortstoleranz eines Teils aus dem Modell gemessen in Pixeln. Sie definiert den Durchmesser einer Umgebung um jeden einzelnen Unterknoten des Teils. Unterknoten, die innerhalb der Umgebung auftreten, werden für die Erkennung des gesamten Teils gewertet. Die Positionen, an denen die Unterknoten eines Teils auftreten, dürfen also um ς Pixel variieren, ohne daß dies einen Einfluß auf die Erkennung des Teils hat.
ς_{Farbe}	Eine Farbtoleranz für den Farbklassifikator in Kapitel 2. Die Farbtoleranz wird als Vorfaktor für die Wahrscheinlichkeit, daß ein Punkt einer bestimmten Farbe zum Vordergrund gehört, eingeführt. Der Vorfaktor kompensiert Unterschiede in den Auftrittswahrscheinlichkeiten verschiedener Farben, sodaß auch Farben von kleineren Objektteilen zum Klassifikationsergebnis beitragen können, wenn sie für ein Objekt charakteristisch sind. Die Farbtoleranzen werden experimentell ermittelt.
$t_1, t_2, t_3, \ldots$	Die Knoten oder Teile des Modells. Sie repräsentieren lokale Regionen der Objektabbildungen durch Merkmale in bestimmten Anordnungen in der Bildebene. Sie werden in Kapitel 4 eingeführt. Ein Knoten faßt eine Anzahl von m Unterknoten U oder Merkmalen auf einer niedrigeren Abstraktionsebene, deren Positionen L und die für die Erkennung benötigten Parameter ς und ϑ zusammen. Außerdem läßt sich jeder Knoten des Modells eindeutig über einen Schlüssel κ identifizieren.
U, U_i	Eine Menge von Unterknoten eines Teils im allgemeinen bzw. eines bestimmten Teils t_i. Die Unterknoten werden durch die Indices ihrer Teile (hier z.B. das i beschrieben, welche gleichbedeutend mit den Knotenschlüsseln κ sind.
$u_1, u_2, u_3, \ldots$	Einzelne Indices einer Unterknotenmenge U.
$\boldsymbol{v}$	Ein komponentenweise Mittelwert mehrerer Vektoren, insbesondere der Mittelwert eines Clusters von Vektoren.
φ	Die Richtung des Gradienten in einem Punkt. Der Gradient zeigt in die Richtung des stärksten Anstiegs von Schwarz nach Weiß.
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	Bildkoordinaten, je nach Zusammenhang absolut oder relativ zu anderen Koordinaten.

Ξ_j	Eine Akkumulationsfunktion für den Knoten j. Ähnlich dem Akkumulator der Hough-Transformation zeigt die Funktion Ξ_j an, wieviele Unterknoten eines Teils an den passenden Positionen relativ zu einer Bezugskoordinate x, y gefunden wurden. Wenn genügend Unterknoten gefunden wurden, gilt ein Teil als erkannt an der Position x, y. Die Entscheidung, ob genügend Unterteile vorliegen, wird mit Hilfe des Schwellwerts ϑ entschieden.
$\psi(\dots)$	Eine kodierte, insbesondere kompaktierte Darstellung eines Vektors. Bei der Beschreibung der Clusterung von Merkmalspunkten nach ihrer Position in der Bildebene (Abschnitt 6.2.5) ist ψ die Identitätsfunktion, d.h. die Clusterung arbeitet direkt auf den Bildkoordinaten der Merkmale. Bei der Clusterung von Teilekandidaten (Abschnitt 6.3.1) sind die Eingabevektoren jedoch sehr hochdimensional, sodaß die Berechnung der Abstände zwischen den Vektoren sehr aufwendig ist. Da jedoch die meisten Vektorkomponenten den Wert Null haben, können die Vektoren verkürzt dargestellt werden, indem zusammenhängende Folgen von Nullen in Form einer Lauflängenkodierung zusammengefaßt werden. Die Funktion ψ wird dann so gewählt, daß sie diese Kodierung leistet. Abstandberechnungen können auf der kodierten Darstellung sehr viel schneller durchgeführt werden.

Anhang B

Merkmalsrauschen in der Cartoon-Datenbank

Es folgen Messungen der Streuung von Intensitätswerten in homogenen Farbflächen. Diese beziehen sich zum einen auf das Rauschen innerhalb einzelner Bildbereiche und zum anderen auf die Streuung über mehrere eingescannte Seiten hinweg.

B.1 Streuung der Intensität innerhalb eines Scans

Wie die Tabelle B.1 zeigt, rauschen die Intensitätswerte innerhalb einzelner Bilder der Cartoon-Datenbank ohne vorherige Filterung enorm. Dies ist vor allem auf das grobe Druckraster zurückzuführen. Eine Standardabweichung von 0,2204 für den Rot-Kanal in der türkisen Fläche bedeutet beispielsweise, daß etwa zwei Drittel aller Punkte über 44 Prozent des darstellbaren Farbraums streuen. Die übrigen Punkte streuen noch weiter.

Farbe	**Nr.**	**Mittelwert**			**Standardabweichung**		
		Rot	**Grün**	**Blau**	**Rot**	**Grün**	**Blau**
Schwarz	27.	0,322	0,315	0,301	0,0595	0,0601	0,0596
	28.	0,321	0,315	0,291	0,0892	0,0881	0,0933
	29.	0,272	0,273	0,272	0,0734	0,0822	0,0837
	30.	0,304	0,281	0,257	0,0775	0,0822	0,0817
Weiß	31.	0,928	0,896	0,84	0,0548	0,0627	0,069
	32.	0,885	0,847	0,785	0,0567	0,0603	0,0634
	33.	0,978	0,926	0,877	0,0292	0,0497	0,055
	34.	0,934	0,894	0,847	0,0561	0,0619	0,0627

Tabelle B.1 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	Nr.	Mittelwert			Standardabweichung		
		Rot	Grün	Blau	Rot	Grün	Blau
Orange	35.	0,94	0,569	0,265	0,0588	0,1702	0,1553
	36.	0,939	0,556	0,169	0,0496	0,1119	0,14
	37.	0,936	0,562	0,167	0,0583	0,1574	0,1335
	38.	0,934	0,579	0,227	0,0598	0,1711	0,1896
	39.	0,964	0,589	0,264	0,042	0,1273	0,1461
Grün	40.	0,439	0,627	0,327	0,0915	0,0805	0,1048
Hellgrün	41.	0,805	0,821	0,545	0,0998	0,0705	0,1351
Hellblau	42.	0,686	0,775	0,833	0,1219	0,0816	0,0613
Blau	43.	0,408	0,466	0,571	0,0912	0,0919	0,0969
Braun	44.	0,778	0,585	0,515	0,0887	0,1021	0,096
Violett	45.	0,745	0,644	0,805	0,196	0,1575	0,1037
Türkis	46.	0,61	0,791	0,796	0,2204	0,1273	0,1246
Grauviolett	47.	0,843	0,792	0,848	0,1298	0,1373	0,1071
Gelb	48.	0,987	0,918	0,055	0,0297	0,0601	0,1231
Rosa	49.	0,974	0,634	0,758	0,0404	0,172	0,1073

Tabelle B.1: **Intensität und Rauschen der einzelnen Farbauszüge innerhalb von einfarbigen Flächen.** Die Intensitätswerte sind auf das Intervall von 0 bis 1 normiert, von einer möglichen Vorverarbeitung in der Scanner-Firmware abgesehen ansonsten aber ungefiltert.

B.2 Streuung der Intensität über mehrere Scans

Die Tabelle B.2 gibt die Intensität (Mittelwert aus Rot-, Grün- und Blauanteil einer Farbe) von einfarbigen Bildausschnitten der Cartoon-Vorlage an. Das Rauschen der Intensität konnte durch eine Filteroperation von einer durchschnittlichen Standardabweichung von 10,6 für Weiß, 15,2 für Orange und 15,4 für Schwarz auf 4,7 für Weiß, 4,0 für Orange und 5,5 für Schwarz gesenkt werden.

Dagegen senkt die Filterung nicht die Schwankungen, die zwischen verschiedenen eingescannten Bildseiten bestehen. Diese resultieren beispielsweise aus dem Vergilben des Papiers von älteren Bänden, möglicherweise auch aus Unterschieden im Druck. Dies zeigt sich darin, daß die für einzelne Flächen berechneten Farbmittelwerte über mehrere Flächen aus verschiedenen Bildern hinweg schwanken. Für diese Schwankungen wurde eine Standardabweichung von 9,2 für Weiß, 8,6 für Orange und 7,6 für Schwarz ermittelt. Bezüglich der in dieser Arbeit verwendeten Normierung auf das Interval 0 bis 1 entspricht dies den Werten 0,036 für Weiß, 0,034 für Orange und 0,030 für Schwarz.

Die Filterung der Eingabebilder verbessert daher zwar die Merkmalsextraktion, hat aber keine Auswirkung auf die Anzahl der Quantisierungsstufen bei der Diskretisierung von Deskriptoren.

Farbe	Nr.	Fläche	Ungefiltert		5x5-Binomialfilter	
		[Pixel]	Mittelw.	Stdabw.	Mittelw.	Stdabw.
Weiß	50.	21546	229,2	11,2	228,3	4,2
	51.	17226	225,3	15,7	224,4	6
	52.	66040	245,2	6,7	244,3	3,5
	53.	22800	227,8	9,9	226,8	4,9
	54.	58254	232	10,8	231	4,1
	55.	50218	224,4	12,7	223,4	5
	56.	44336	228,1	13,7	227,1	5,8
	57.	51246	228,3	12,1	227,3	4,9
	58.	24739	228	12,3	227	4,9
	59.	9546	225,9	15,4	224,9	5,7
	60.	25217	226	14,3	225,1	6,4
	61.	18718	241,2	8	240,2	4,2
	62.	59875	232,2	8,5	231,2	4,4
	63.	15120	222,1	11,3	221,2	3,8
	64.	17110	241,2	9,7	240,3	4,2
	65.	20060	225,8	14,2	224,9	5,7
	66.	13248	231,9	10,5	230,9	4,3
	67.	60300	231,1	11,1	230,1	4,3
	68.	7050	227,4	14,3	226,5	5
	69.	78156	229,7	12	228,8	5,3
	70.	3948	230,2	10,9	229,2	5,3
	71.	7560	246,9	5,9	246	2,3
	72.	14940	225	10,4	224,1	6,4
	73.	7590	223,9	10,5	223	5,5
	74.	20041	225,6	10,4	224,7	6,3
	75.	37800	224,8	11,1	223,8	6,2
	76.	22755	230,9	9,8	229,9	5,4
	77.	33150	232,5	10,1	231,6	5,7
	78.	8008	224,3	10,4	223,4	5,8
	79.	3783	218,3	13,1	217,4	5,5
	80.	10780	230,6	9,5	229,6	5,4
	81.	8640	226,7	9,8	225,7	4,3
	82.	71688	232	8	231,1	3
	83.	10425	222,4	12,9	221,4	5,7
	84.	75500	230	10,1	229,1	4,3
	85.	11163	241,1	9,9	240,2	5,4
	86.	5432	242,9	7,2	241,9	4
	87.	48365	253,6	1,8	253,4	1

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	Nr.	Fläche	Ungefiltert		5x5-Binomialfilter	
		[Pixel]	Mittelw.	Stdabw.	Mittelw.	Stdabw.
	88.	14554	221,6	11,2	220,6	5,5
	89.	5985	225,3	12,7	224,4	5,2
	90.	97410	246,7	7,6	245,8	3,2
	91.	14484	214,6	11,9	213,7	4,4
	92.	26786	243,9	7,3	243	3,4
	93.	15620	241,5	9,5	240,6	4
	94.	11550	226,3	12,5	225,3	4,5
	95.	12028	242,8	9,4	241,9	3,9
	96.	17094	241,7	6,7	240,8	3,5
	97.	25760	245,8	8	244,9	3
	98.	9870	224,9	11,4	224	4,6
	99.	23010	225,2	13,2	224,2	4,8
	100.	11776	244,1	7,7	243,2	3,4
	101.	11070	245,2	7,3	244,3	3,6
	102.	17760	247,8	6,7	246,9	2,7
	103	16915	244,7	6,4	243,8	3,5
	104.	8024	234,5	10,4	233,6	3,4
	105.	16577	242,8	10,3	241,8	4
	106.	102606	243,6	9,8	242,7	4,1
	107.	81972	246,1	7,6	245,2	3,2
	108.	37536	246	7,8	245	3,3
	109.	17876	248,3	6,2	247,4	2,6
	110.	13631	226,4	12,1	225,5	6,7
	111.	12558	226,7	12,1	225,8	5,2
	112.	26883	226,5	12	225,6	6,1
	113.	48633	246,5	6,1	245,7	2,7
	114.	5341	240,1	10,3	239,1	4,6
	115.	4168	220,1	14,3	219,1	6,4
	116.	10500	223,9	12,4	223	5,9
	117.	7840	240,4	9,9	239,5	5,3
	118.	5808	227,1	11,4	226,2	4,7
	119.	3016	217,6	12,8	216,7	5,4
	120.	10976	227,6	12,1	226,6	5,9
	121.	9540	219,7	14,2	218,7	6,9
	122.	5644	229,5	11,6	228,6	5,4
	123.	12510	240	11,3	239,1	5,6
	124.	4929	228,7	12,5	227,8	5,1
	125.	3721	227,8	13,6	226,9	6,3
	126.	10164	222,7	12,4	221,7	5,5
	127.	8532	222,8	13,4	221,9	5,8

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	**Nr.**	**Fläche**	**Ungefiltert**		**5x5-Binomialfilter**	
		[Pixel]	**Mittelw.**	**Stdabw.**	**Mittelw.**	**Stdabw.**
	128.	3657	224,5	13,2	223,5	5,2
	129.	27738	224,6	13,1	223,7	6,3
Orange	130.	4845	154,4	15,4	153,5	4,7
	131.	3960	150,5	18,5	149,5	5,2
	132.	2200	161,9	9,2	161,1	3,8
	133.	7956	158,1	11,3	157,2	4,5
	134.	2166	153,6	15,5	152,6	4,5
	135.	2684	155	19,1	154	4,9
	136.	2184	159,6	17,2	158,7	5,1
	137.	1537	154,9	15,9	154,1	4,5
	138.	2160	148	16,6	147,1	4,5
	139.	1155	153,9	18,5	152,9	5,1
	140.	2438	156,6	15,6	155,7	4,9
	141.	1800	162,6	11,5	161,8	4,1
	142.	2650	145,7	13,5	144,7	4,7
	143.	2829	139,5	17,8	138,5	4,9
	144.	2295	144,9	16,5	144	4,8
	145.	1120	145,5	16,8	144,5	5
	146.	1230	139	15,9	138,1	4,2
	147.	4929	147	17,2	146,1	5,1
	148.	2205	150,2	16,1	149,3	4,8
	149.	3456	156,3	15,8	155,5	4,7
	150.	1190	153,8	17	153	4,6
	151.	2340	143,5	16,7	142,6	4,6
	152.	3120	152,8	14,8	151,9	4,3
	153.	2257	148,6	14,2	147,7	5,2
	154.	5488	137,9	15,7	136,9	4,9
	155.	2268	163,5	12,3	162,6	4
	156.	5220	166,3	10,1	165,7	4,2
	157.	2107	161,6	12,9	160,7	4
	158.	1292	144,7	15,2	143,9	4,2
	159.	1312	151,5	14,5	150,6	4,4
	150.	2546	153,8	14,2	152,9	4,2
	161.	2112	158,2	15,6	157,4	4,5
	162.	2394	164,4	15,9	163,7	4,3
	163.	1155	162,3	10,2	161,4	4,1
	164.	1833	152,7	13,5	151,7	4,2
	165.	2480	155,2	14,2	154,3	4,8
	166.	1332	159,7	15,7	158,7	4,2
	167.	1305	155,8	17,6	154,9	5,9

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	Nr.	Fläche	Ungefiltert		5x5-Binomialfilter	
		[Pixel]	Mittelw.	Stdabw.	Mittelw.	Stdabw.
	168.	1665	162,7	15,1	162	4,5
	169.	2058	157,7	14,7	156,9	4,6
	170.	4081	145,1	14,9	144,2	6
	171.	2109	151,8	12,1	151,1	4,2
	172.	5459	156,1	14,7	155,3	5,1
	173.	2765	140,2	13,9	139,3	4,9
	174.	5650	139,7	14,7	138,8	5,5
	175.	2891	145,7	14	144,8	4,9
	176.	3068	145,3	15,3	144,4	5,4
	177.	2550	141,6	15,5	140,8	5,2
	178.	2106	141,2	14,8	140,4	5,1
	179.	2640	144,5	13,2	143,5	4,6
	180.	1638	149,4	14,1	148,6	5,1
	181.	3354	137,1	15,1	136,2	5,4
	182.	3712	129,2	18,2	128,3	5,5
	183.	3233	145,8	14	144,8	5
	184.	1746	135,1	14	134,2	5
	185.	4675	147,7	14,7	146,8	4,8
	186.	5396	158,6	13,2	158	4,7
	187.	5664	163,3	16	162,6	5,5
	188.	3360	140,2	13,8	139,3	4,8
	189.	1170	142,4	15,6	141,5	5,3
	190.	484	140,6	18,5	139,5	5,6
	192.	990	137,3	16,4	136,4	5
	193.	2294	148,1	15,6	147,2	5,4
	194.	380	155,6	13,5	155	3,8
	195.	3948	149,9	16,3	149,1	5,5
	196.	2990	148,9	16,5	148	5,9
	197.	2430	149,5	18,6	148,5	6,2
	198.	1225	158,9	16,1	158	5,6
	199.	850	136,5	15,1	135,6	4,5
	200.	3111	143,1	16,7	142,1	5,3
	201.	3808	134,9	14,7	134	5,4
	202.	1860	138,5	15,6	137,4	5,9
	203.	900	131,6	14,6	130,7	5,1
	204.	1740	147,2	18,4	146,4	6
	205.	3783	142,2	14,3	141,2	4,7
	206.	2835	149,5	18,4	148,5	6,2
	207.	3200	149,4	17,2	148,4	5,4
	208.	1517	140,9	17,4	140	6,2

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	**Nr.**	**Fläche**	**Ungefiltert**		**5x5-Binomialfilter**	
		[Pixel]	**Mittelw.**	**Stdabw.**	**Mittelw.**	**Stdabw.**
	209.	2646	143,8	15,5	142,9	5,7
	210.	961	136,5	14	135,7	5,1
Schwarz	211.	9720	72,2	17,3	71,3	5,8
	212.	6768	79,9	18,7	79	5,6
	213.	2580	78,2	12,1	77,2	5,1
	214.	2960	78,8	12	77,9	4,9
	215.	7210	73,2	15	72,2	4,9
	216.	3420	69	12,9	68	4,6
	217.	3520	76,5	14,6	75,6	4,9
	218.	4059	74,3	18,4	73,3	6,1
	219.	4836	70,7	15,6	69,7	5,9
	220.	4224	69,9	16,4	68,9	4,8
	221.	2664	74,6	15,7	73,6	4,9
	222.	3337	75,4	18,8	74,5	6,8
	223.	6208	73,7	10,6	72,8	4,6
	224.	2072	83,3	12,8	82,4	4,9
	225.	6750	93,2	14,3	92,3	5,3
	226.	2987	89,2	12,4	88,3	4,5
	227.	3658	89,7	17,6	88,7	5,8
	228.	29192	84,5	15,3	83,6	5,6
	229.	16605	75,6	16,6	74,7	5,6
	230.	42012	83,4	15,3	82,5	5,3
	231.	1504	83,6	21,4	82,7	7
	232.	1380	70,5	13,4	69,5	4,9
	233.	13770	77,4	14,3	76,5	4,7
	234.	1980	71,1	19	70,1	6,4
	235.	4690	71,1	14,8	70,2	4,9
	236.	2135	86,5	14,4	85,6	5,1
	237.	6251	81,9	12,6	80,9	5,6
	238.	1428	84,1	18,4	83,1	5,4
	239.	5742	75,3	17,9	74,4	5,2
	240.	1634	78,7	16	77,7	5,1
	241.	9312	83,8	16,9	82,8	5,6
	242.	3456	87,9	17,8	86,9	5,7
	243.	1350	88,6	24,3	87,4	6,4
	244.	2418	90,9	9,2	90	4,7
	245.	2838	82,8	12,5	81,8	4,7
	246.	775	85,8	17,3	84,9	5,2
	247.	1080	86,6	15,4	85,6	5,6
	248.	1155	89,1	14,2	88,1	5,3

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	Nr.	Fläche	Ungefiltert		5x5-Binomialfilter	
		[Pixel]	Mittelw.	Stdabw.	Mittelw.	Stdabw.
	249.	2067	87	19,7	86	5,7
	250.	8692	85,8	17,1	84,8	5,6
	251.	6392	68,6	15,5	67,7	6
	252.	7007	73,4	12,3	72,5	4,7
	253.	5767	67,5	12	66,6	4,4
	254.	2714	72,8	14,2	71,8	5,4
	255.	6390	66,1	13,1	65,1	4,8
	256.	36842	63,1	15,1	62,2	6
	257.	2912	65,9	16,6	65,1	5,7
	258.	10269	65,4	14,1	64,5	5,7
	259.	29106	63,1	15,5	62,1	6,4
	250.	4964	69,7	12,5	68,8	5,2
	251.	15680	67,1	16,5	66,2	6,9
	252.	6864	63,6	14	62,7	5,7
	253.	1740	74,7	13,6	73,8	4,1
	254.	1000	93,4	13,5	92,5	5,2
	255.	2401	77	11,7	76	4,6
	256.	2964	87,9	15,6	87	7
	257.	1935	90,3	19	89,4	7,9
	258.	3000	87,4	20,7	86,4	7
	259.	1702	68,1	13,8	67,3	5
	260.	5405	72	12,4	71,1	5,4
	261.	3312	74,4	14,6	73,4	5,3
	262.	3795	72,7	17,2	71,8	5,1
	263.	4830	78,1	14	77,2	5,2
	264.	3718	78,4	18,1	77,6	6
	265.	14300	77,1	14,5	76,1	5,7
	266.	7480	71,9	13,5	71	5
	267.	3478	75,1	13,3	74,1	5,2
	268.	918	85,2	17,4	84,2	6,7
	269.	2544	79	15,3	78	4,9
	270.	5029	66,5	13,7	65,6	5
	271.	1240	76,3	14,4	75,3	5,6
	272.	2016	77	16	76,2	5,3
	273.	2107	72,8	17,8	71,8	5,4
	274.	2000	78,2	16,3	77,3	6,6
	275.	4182	73,1	17,6	72,1	5,6
	276.	4428	83	13,6	82,1	5,1
	277.	6401	75,8	15,7	74,9	5,7
	278.	3968	78,4	15,8	77,4	5,4

Tabelle B.2 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Farbe	Nr.	Fläche	Ungefiltert		5x5-Binomialfilter	
		[Pixel]	Mittelw.	Stdabw.	Mittelw.	Stdabw.
	279.	1170	80,5	14,2	79,6	4,9
	280.	1764	75,2	17	74,3	6,2

Tabelle B.2: **Intensität und Rauschen innerhalb von einfarbigen Flächen.** Die Intensitätswerte sind als Mittelwert über alle Punkte und alle Farbauszüge innerhalb von einfarbigen Bereichen der Ursprungsbilder dargestellt. Der Wertebereich ist das Intervall von 0 bis 255, wie es vom Scanner geliefert wird.

B.3 Genauigkeit der Objekterkennung für verschiedene Merkmalsquantisierungen

Die Tabellen B.3 bis B.8 geben die Güte der Objekterkennung über dem Parameterraum des Modells und des Objekterkennungsverfahren an. Die Messungen werden für ein Modell einer Augenbraue und eines Schnabels einer Figur der Cartoon-Datenbank durchgeführt. Das Modell besteht jeweils nur aus einem Teileknoten und den entsprechenden Merkmalen. Der Parameterraum wird durch die Ortstoleranz und die Schwelle des Teileknotens sowie die Anzahl der Quantisierungsstufen der Merkmale parametrisiert.

Schwelle		Genauigkeit						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
5	20%	0,60	0,56	0,47	0,45	0,44	0,44	0,44
	30%	1,15	0,58	0,55	0,47	0,46	0,46	0,45
	40%	3,90	0,74	0,63	0,50	0,50	0,46	0,46
	50%	11,93	0,90	0,58	0,57	0,51	0,48	0,47
	60%	5,00	1,76	0,58	0,55	0,56	0,51	0,52
	70%	6,33	3,54	0,88	0,58	0,58	0,57	0,55
	80%	2,00	8,11	1,90	0,98	0,62	0,62	0,59
	90%	1,00	7,15	5,95	1,51	1,03	0,68	0,73
10	20%	1,46	0,67	0,57	0,54	0,50	0,48	0,46
	30%	4,34	0,97	0,68	0,64	0,60	0,53	0,49
	40%	10,29	1,86	0,75	0,66	0,68	0,56	0,52
	50%	11,46	4,84	1,26	0,88	0,75	0,61	0,57
	60%	9,44	8,16	2,87	1,29	0,98	0,76	0,63
	70%	6,47	13,62	6,62	2,23	1,47	1,00	0,82
	80%	2,00	12,35	9,02	5,89	2,25	1,59	1,18
	90%	2,00	4,95	10,58	6,87	6,21	4,11	2,82
15	20%	2,69	0,94	0,69	0,59	0,58	0,50	0,50

Tabelle B.3 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

	Schwelle	**Genauigkeit**						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
	30%	7,61	1,79	1,10	0,80	0,66	0,57	0,55
	40%	12,36	4,29	2,04	1,18	0,81	0,69	0,63
	50%	11,46	9,12	4,96	2,30	1,25	1,02	0,75
	60%	6,00	12,56	7,61	4,27	2,58	1,77	1,15
	70%	4,37	14,24	10,59	8,15	5,16	3,56	2,05
	80%	2,00	10,98	10,37	11,08	8,82	5,95	4,04
	90%	1,00	7,17	5,63	9,31	10,35	9,73	7,35
20	20%	3,32	1,22	0,80	0,65	0,55	0,53	0,48
	30%	9,16	2,94	1,54	1,12	0,77	0,71	0,61
	40%	13,27	5,94	3,73	1,88	1,30	1,09	0,81
	50%	9,59	11,37	6,35	4,40	2,65	1,51	1,21
	60%	5,96	14,18	9,50	7,36	4,37	2,81	1,79
	70%	5,37	13,65	13,00	9,66	9,16	5,91	3,65
	80%	2,00	8,29	12,03	11,46	12,12	8,22	7,40
	90%	1,00	3,71	7,26	12,54	12,31	11,14	7,21
25	20%	4,22	1,46	0,92	0,69	0,62	0,58	0,51
	30%	11,52	4,03	2,00	1,41	0,95	0,84	0,64
	40%	17,14	7,75	4,38	2,31	1,50	1,17	0,87
	50%	9,73	11,36	8,66	4,49	3,02	2,17	1,45
	60%	6,07	14,21	11,75	9,57	6,82	3,68	2,27
	70%	4,00	10,01	12,32	10,12	9,12	6,61	5,02
	80%	2,00	7,66	15,19	12,28	10,51	7,95	7,37
	90%	1,00	3,00	8,08	9,89	13,52	11,84	10,05
30	20%	5,24	1,84	1,05	0,78	0,61	0,59	0,53
	30%	11,59	4,16	2,33	1,65	1,12	0,89	0,70
	40%	13,52	8,96	4,62	3,08	2,21	1,51	0,95
	50%	7,29	13,17	9,26	5,69	4,06	2,44	1,80
	60%	6,17	14,23	11,98	10,24	6,90	5,15	2,85
	70%	2,00	10,82	12,52	11,40	9,95	8,37	5,37
	80%	2,00	6,81	13,62	15,10	13,59	12,88	8,14
	90%	1,00	4,00	6,82	10,30	11,16	9,51	11,46
35	20%	5,90	2,33	1,19	0,84	0,68	0,60	0,53
	30%	12,04	5,01	2,84	1,79	1,44	1,00	0,81
	40%	11,52	9,23	5,28	3,75	2,33	1,51	1,18
	50%	8,25	13,48	9,66	6,40	4,49	2,80	1,96
	60%	6,47	11,37	13,49	10,48	8,52	5,49	3,58
	70%	4,00	10,48	13,90	12,31	10,85	9,01	5,76
	80%	2,00	8,03	13,32	11,70	13,30	11,93	9,83
	90%	1,00	4,00	7,79	9,77	9,64	14,10	10,78
40	20%	6,74	2,70	1,24	1,00	0,72	0,62	0,57

Tabelle B.3 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Schwelle		Genauigkeit						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
	30%	10,85	6,19	3,19	2,05	1,37	1,04	0,92
	40%	12,53	11,23	6,83	4,19	2,50	1,74	1,39
	50%	7,69	15,67	11,78	8,21	5,98	3,82	2,35
	60%	7,00	11,65	12,90	11,48	8,83	6,43	4,68
	70%	3,00	10,14	13,28	12,76	12,25	9,24	6,82
	80%	1,00	7,42	11,12	13,21	12,88	11,09	10,04
	90%	0,00	3,00	6,00	9,38	10,03	11,18	9,95

Tabelle B.3: **Augenbrauenerkennung abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Werte zeigen die Genauigkeit der Objekterkennung für jede Parametrisierung des Klassifikators summiert über alle 35 Testbilder. Die Spalte ζ gibt die Anzahl der Quantisierungsstufen für die Kantenorientierung an. Die Spalte ϑ gibt den Schwellwert, ς die Ortstoleranz [Pixel] des Teileknotens des Augenbrauenmodells an.

Schwelle		Genauigkeit						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
5	20	1,88	1,30	1,16	1,08	1,05	1,03	1,01
	30	3,19	1,56	1,31	1,19	1,12	1,07	1,05
	40	5,48	1,98	1,55	1,35	1,24	1,16	1,11
	50	9,11	2,61	1,88	1,58	1,41	1,29	1,20
	60	2,62	3,86	2,33	1,88	1,63	1,47	1,35
	70	1,00	6,31	2,91	2,36	1,96	1,73	1,57
	80	1,00	6,23	4,17	2,80	2,43	2,07	1,85
	90	1,00	1,00	7,20	4,23	3,14	2,70	2,36
10	20	3,75	1,83	1,42	1,23	1,14	1,09	1,05
	30	6,05	2,69	1,81	1,47	1,29	1,19	1,13
	40	7,67	4,29	2,50	1,88	1,56	1,38	1,26
	50	6,80	6,53	3,67	2,53	1,98	1,67	1,47
	60	2,29	8,44	5,09	3,47	2,62	2,11	1,78
	70	1,00	9,98	6,94	4,81	3,66	2,89	2,35
	80	1,00	6,35	10,17	6,49	4,94	3,93	3,25
	90	1,00	1,00	9,24	10,72	7,70	6,00	4,91
15	20	4,77	2,35	1,70	1,40	1,25	1,16	1,11
	30	7,16	3,77	2,39	1,82	1,52	1,35	1,24
	40	8,74	5,88	3,55	2,53	2,00	1,68	1,47
	50	6,23	8,64	5,35	3,64	2,72	2,20	1,86
	60	2,20	10,42	7,30	5,24	3,82	2,99	2,43

Tabelle B.4 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Schwelle		**Genauigkeit**						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
	70	1,00	9,78	9,72	7,06	5,57	4,37	3,45
	80	1,00	5,53	13,46	9,40	7,34	6,01	4,95
	90	1,00	1,00	7,60	14,01	11,09	8,59	7,24
20	20	5,24	2,79	1,92	1,54	1,33	1,22	1,15
	30	7,65	4,61	2,90	2,12	1,72	1,47	1,33
	40	9,12	6,97	4,54	3,15	2,41	1,95	1,66
	50	5,71	9,42	6,83	4,80	3,50	2,75	2,23
	60	1,25	10,39	9,04	6,78	5,05	3,85	3,08
	70	1,00	5,97	10,43	8,80	7,21	5,69	4,47
	80	1,00	5,38	11,45	11,41	9,15	7,73	6,45
	90	1,00	1,00	5,64	11,61	12,33	10,69	9,08
25	20	5,54	3,14	2,12	1,66	1,41	1,27	1,18
	30	7,29	5,15	3,32	2,43	1,92	1,61	1,42
	40	9,41	7,11	5,22	3,59	2,76	2,22	1,86
	50	5,67	9,66	7,50	5,67	4,01	3,18	2,60
	60	1,25	10,37	9,44	7,69	6,01	4,51	3,60
	70	1,00	5,27	11,05	9,43	8,03	6,59	5,20
	80	1,00	5,33	10,49	11,81	10,22	8,74	7,33
	90	1,00	1,00	4,00	9,69	12,49	11,97	10,33
30	20	5,81	3,49	2,33	1,78	1,49	1,32	1,22
	30	7,31	5,50	3,72	2,70	2,11	1,73	1,51
	40	9,68	7,73	5,78	4,07	3,09	2,47	2,04
	50	4,67	10,87	8,08	6,43	4,71	3,61	2,94
	60	1,33	9,14	10,37	8,43	6,88	5,25	4,14
	70	1,00	5,33	12,13	10,60	9,07	7,73	6,10
	80	1,00	2,14	8,61	12,25	11,99	10,31	8,80
	90	1,00	1,00	3,00	6,94	12,91	12,99	11,93
35	20	5,87	3,78	2,49	1,88	1,54	1,35	1,24
	30	7,11	5,51	4,02	2,91	2,26	1,85	1,58
	40	9,47	8,15	5,98	4,39	3,37	2,69	2,20
	50	2,60	11,34	8,43	6,77	5,14	3,98	3,22
	60	1,00	8,43	10,86	8,88	7,39	5,77	4,56
	70	1,00	5,24	12,13	11,39	9,85	8,50	6,96
	80	1,00	1,00	8,38	11,80	12,99	11,44	9,94
	90	1,00	1,00	1,00	6,40	11,42	12,94	12,65
40	20	6,04	4,04	2,64	1,96	1,60	1,40	1,27
	30	7,43	5,97	4,34	3,12	2,40	1,95	1,65
	40	8,22	8,59	6,63	4,89	3,63	2,88	2,34
	50	2,78	12,04	9,09	7,41	5,78	4,39	3,51
	60	1,00	7,56	11,36	9,71	8,11	6,48	5,03

Tabelle B.4 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

Schwelle		**Genauigkeit**						
ζ	ϑ	$\varsigma = 3$	$\varsigma = 5$	$\varsigma = 7$	$\varsigma = 9$	$\varsigma = 11$	$\varsigma = 13$	$\varsigma = 15$
	70	1,00	4,31	12,02	11,64	10,88	9,36	7,71
	80	1,00	1,00	6,43	12,65	13,60	12,17	10,63
	90	1,00	1,00	1,00	6,27	9,41	11,85	12,71

Tabelle B.4: **Erkennung von 24 Augenbrauen abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Werte zeigen die Genauigkeit der Objekterkennung für jede Parametrisierung des Klassifikators summiert über jeweils 24 Donald-Bilder mit Augenbrauen. Die Spalte ζ gibt die Anzahl der Quantisierungsstufen für die Kantenorientierung an. Die Spalte ϑ gibt den Schwellwert, ς die Ortstoleranz [Pixel] des Teileknotens des Augenbrauenmodells an.

		Anzahl erkannter Augenbrauen für Schwellen ϑ							
ζ	ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
5	3	24	24	23	16	4	1	1	1
	5	24	24	24	24	24	21	11	1
	7	24	24	24	24	24	24	24	18
	9	24	24	24	24	24	24	24	24
	11	24	24	24	24	24	24	24	24
	13	24	24	24	24	24	24	24	24
	15	24	24	24	24	24	24	24	24
10	3	24	24	20	10	3	1	1	1
	5	24	24	24	24	23	18	7	1
	7	24	24	24	24	24	24	22	13
	9	24	24	24	24	24	24	24	22
	11	24	24	24	24	24	24	24	24
	13	24	24	24	24	24	24	24	24
	15	24	24	24	24	24	24	24	24
15	3	24	24	20	8	3	1	1	1
	5	24	24	24	24	23	16	6	1
	7	24	24	24	24	24	24	22	8
	9	24	24	24	24	24	24	24	22
	11	24	24	24	24	24	24	24	23
	13	24	24	24	24	24	24	24	24
	15	24	24	24	24	24	24	24	24
20	3	24	23	19	7	2	1	1	1
	5	24	24	24	23	21	10	6	1
	7	24	24	24	24	24	22	20	6
	9	24	24	24	24	24	24	23	18

Tabelle B.5 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

		Anzahl erkannter Augenbrauen für Schwellen ϑ							
ζ	ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
	11	24	24	24	24	24	24	24	21
	13	24	24	24	24	24	24	24	23
	15	24	24	24	24	24	24	24	24
25	3	24	22	18	7	2	1	1	1
	5	24	24	24	22	19	8	6	1
	7	24	24	24	24	23	21	15	4
	9	24	24	24	24	24	24	21	12
	11	24	24	24	24	24	24	24	20
	13	24	24	24	24	24	24	24	23
	15	24	24	24	24	24	24	24	24
30	3	24	21	14	6	2	1	1	1
	5	24	24	23	22	15	7	3	1
	7	24	24	24	24	23	20	11	3
	9	24	24	24	24	24	22	19	9
	11	24	24	24	24	24	24	22	19
	13	24	24	24	24	24	24	24	21
	15	24	24	24	24	24	24	24	24
35	3	24	20	12	4	1	1	1	1
	5	24	24	23	22	13	6	1	1
	7	24	24	24	23	23	19	10	1
	9	24	24	24	24	24	21	18	8
	11	24	24	24	24	24	24	21	15
	13	24	24	24	24	24	24	24	20
	15	24	24	24	24	24	24	24	24
40	3	24	20	11	4	1	1	1	1
	5	24	24	23	22	11	5	1	1
	7	24	24	24	23	23	18	8	1
	9	24	24	24	24	24	21	18	8
	11	24	24	24	24	24	24	21	12
	13	24	24	24	24	24	24	24	18
	15	24	24	24	24	24	24	24	23

Tabelle B.5: **Anzahl erkannter Augenbrauen abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Tabelle gibt für jede Parametrisierung eines Augenbrauenmodells an, wieviele Augenbrauen in den 24 Testbildern erkannt wurden.

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
	5 Quantisierungsstufen							
3	140397	21163	1988	123	18	6	2	0
5	593441	365603	167762	59070	16434	2984	416	22
7	757479	610315	429759	257919	134874	53702	16295	1997
9	830170	728600	596734	442047	296455	165141	76039	20325
11	871469	798068	697278	571817	437534	290937	167444	65127
13	895645	843717	762627	663460	545889	405640	269416	131864
15	910960	874098	810494	727380	628398	501920	364527	210209
	10 Quantisierungsstufen							
3	30462	4137	447	61	17	6	2	0
5	277690	107670	31462	7300	1525	277	62	16
7	519876	318732	157635	65381	24360	6501	1276	153
9	665371	497118	320235	176516	87586	35637	11406	1546
11	755729	622388	464292	305589	179917	88627	38294	8825
13	814423	710050	577538	426197	286124	160315	80862	26975
15	854133	770974	662581	529046	388970	244423	135109	55037
	15 Quantisierungsstufen							
3	18133	2736	320	48	16	3	2	0
5	174824	58106	15342	3255	869	179	44	14
7	380588	198736	84697	30039	10108	2779	660	109
9	537278	353218	194454	92432	40651	14639	4393	683
11	649261	484606	314722	179988	92929	41134	15724	3370
13	729164	589644	426948	276718	160786	81046	36971	10498
15	786286	672592	524030	371955	239714	133101	67933	23253
	20 Quantisierungsstufen							
3	12593	1900	222	43	14	2	0	0
5	123660	36274	8816	2005	604	139	40	7
7	302636	138517	50829	15867	5195	1572	434	89
9	460812	271904	131912	54631	21720	7573	2199	433
11	584246	399371	232471	116828	55572	22180	7897	1610
13	675648	511709	337551	193916	104885	48444	19987	5279
15	742626	603550	437031	278551	165066	86611	40160	12416
	25 Quantisierungsstufen							
3	9514	1524	167	39	13	2	0	0
5	90783	25249	6080	1349	404	97	32	7
7	247268	100792	34585	10253	3392	996	309	71
9	404100	213465	96379	36414	13957	4764	1308	311
11	534097	334023	179502	84507	37320	14563	4836	971

Tabelle B.6 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
13	633960	446079	272810	147699	75592	33076	12982	3059
15	708965	543545	367204	220316	125077	62257	27051	7882
	30 Quantisierungsstufen							
3	7434	1160	126	33	11	2	0	0
5	70355	18984	4259	923	290	83	33	3
7	206090	77934	24998	6961	2207	692	222	56
9	357947	175652	73856	25641	9348	2943	880	238
11	490955	286989	146451	63007	26268	9684	2966	646
13	597580	395617	230970	117305	56513	23147	8194	1946
15	679804	494153	321050	183006	98320	45798	18451	5105
	35 Quantisierungsstufen							
3	5833	881	90	28	7	2	0	0
5	57036	15116	3018	645	222	64	26	1
7	178241	63177	19240	5001	1516	456	176	38
9	325410	149280	59598	19535	6900	1919	608	197
11	459146	254182	123395	50086	19889	6858	1896	510
13	569451	361122	200869	96790	44643	17603	5720	1246
15	656922	460316	286896	155896	80805	35534	13461	3383
	40 Quantisierungsstufen							
3	4637	630	73	19	7	2	0	0
5	47767	11813	2151	485	175	49	20	0
7	156851	52385	14435	3761	1143	354	137	28
9	297822	129726	47745	14832	5127	1487	485	140
11	431413	227851	104951	39805	15403	5005	1416	404
13	543705	331776	177103	80911	35503	13295	4313	956
15	633802	429167	259015	134906	67527	28568	10249	2452

Tabelle B.6: **Anzahl falscher Treffer abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Werte zeigen die falschen Treffer für jede Parametrisierung des Klassifikators summiert über alle 24 Testbilder. Die Spalte ζ gibt die Anzahl der Quantisierungsstufen für die Kantenorientierung an. Die Spalte ϑ gibt den Schwellwert, ς die Ortstoleranz [Pixel] des Teileknotens des Augenbrauenmodells an.

		Anzahl erkannter Schnäbel für Schwellen ϑ							
ζ	ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
4	5	19	19	19	19	18	10	5	2
	9	19	19	19	19	19	19	19	19
	13	19	19	19	19	19	19	19	19

Tabelle B.7 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

		Anzahl erkannter Schnäbel für Schwellen ϑ							
ζ	ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
	17	19	19	19	19	19	19	19	19
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19
8	5	19	19	19	19	16	7	4	1
	9	19	19	19	19	19	19	19	14
	13	19	19	19	19	19	19	19	19
	17	19	19	19	19	19	19	19	19
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19
12	5	19	19	19	17	9	5	3	1
	9	19	19	19	19	19	19	16	8
	13	19	19	19	19	19	19	19	19
	17	19	19	19	19	19	19	19	19
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19
16	5	19	19	18	12	7	5	3	1
	9	19	19	19	19	19	16	12	6
	13	19	19	19	19	19	19	19	15
	17	19	19	19	19	19	19	19	19
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19
20	5	19	19	15	10	6	5	2	1
	9	19	19	19	19	18	13	9	6
	13	19	19	19	19	19	19	18	13
	17	19	19	19	19	19	19	19	18
	21	19	19	19	19	19	19	19	18
	25	19	19	19	19	19	19	19	19
24	5	19	18	13	8	6	4	2	1
	9	19	19	19	19	16	12	8	4
	13	19	19	19	19	19	19	17	11
	17	19	19	19	19	19	19	19	17
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19

Tabelle B.7: **Anzahl erkannter Schnäbel abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Tabelle gibt für jede Parametrisierung des Schnabelmodells an, wieviele Schnäbel in den 19 Testbildern erkannt wurden.

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
	4 Quantisierungsstufen							
5	456774	304617	144731	41476	6591	694	46	0
9	539647	511177	463927	378161	266188	153538	57203	9228
13	551289	540663	520814	484239	416728	322469	205318	89495
17	554366	549415	539460	520262	475456	397378	290574	168190
21	555590	552769	546911	536081	506169	438176	336279	215075
25	556156	554672	550793	544289	525119	465485	366473	245198
	8 Quantisierungsstufen							
5	208870	72356	19704	4162	747	104	4	0
9	465906	361985	247416	136626	60113	20609	4843	628
13	526883	482491	414481	321014	213800	118296	47294	11109
17	544406	523012	488309	427537	339795	235376	129027	46323
21	550843	539726	520030	484554	422119	327508	211338	101655
25	553751	547457	535957	513542	471119	394310	280252	159647
	12 Quantisierungsstufen							
5	103476	24267	5092	894	137	14	0	0
9	381236	246072	130797	54419	16626	4330	864	99
13	489096	411656	318555	209297	111105	48448	13564	1959
17	526749	485986	425085	341705	241147	141174	62588	15704
21	542522	519598	484228	425870	341707	240608	133055	49818
25	548918	536054	515107	477178	415005	322247	207584	98444
	16 Quantisierungsstufen							
5	56700	10271	1730	235	47	0	0	0
9	317737	170921	74566	24498	6397	1367	244	6
13	455450	358565	249150	139836	64197	22962	5094	469
17	510299	453748	378114	280994	175527	92455	34123	6444
21	534384	501444	452771	379375	285064	183957	92751	27124
25	544806	525870	496585	445333	367823	271155	160195	66445
	20 Quantisierungsstufen							
5	34230	4899	648	87	4	0	0	0
9	270799	123387	45441	11716	2591	529	71	3
13	428889	317465	198678	98235	38624	10803	2186	137
17	495087	428712	342019	235733	131480	62287	18842	2828
21	525992	485226	427216	344808	244452	142665	63972	15588
25	540541	516157	478758	418201	334351	232143	126292	44715
	24 Quantisierungsstufen							
5	21440	2548	295	24	0	0	0	0
9	223433	87958	28070	6485	1363	195	11	1

Tabelle B.8 – Fortsetzung auf der nächsten Seite

Fortsetzung von der Vorseite

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
13	400512	272773	155090	68974	23609	6291	995	41
17	478327	399168	302599	192731	99933	42100	11582	1309
21	516488	466507	398704	308629	202309	113057	45604	10098
25	535891	504611	458385	389778	297538	195797	100195	31750

Tabelle B.8: **Anzahl falscher Treffer bei der Erkennung von 19 Schnäbeln abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Werte zeigen die Anzahl der falschen Treffer für jede Parametrisierung des Klassifikators summiert über alle 19 Testbilder. Die Spalte ζ gibt die Anzahl der Quantisierungsstufen für die Orientierung von Skelettlinien an. Die Spalte ϑ gibt den Schwellwert, ς die Ortstoleranz [Pixel] des Teileknotens des Schnabelmodells an.

		Anzahl erkannter Objekte für Schwellen ϑ							
ζ	ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
		Erkennung von Augenbrauen							
1	3	24	24	24	23	17	4	1	1
	5	24	24	24	24	24	24	23	13
	7	24	24	24	24	24	24	24	24
	9	24	24	24	24	24	24	24	24
	11	24	24	24	24	24	24	24	24
	13	24	24	24	24	24	24	24	24
	15	24	24	24	24	24	24	24	24
40	3	24	20	11	4	1	1	1	1
	5	24	24	23	22	11	5	1	1
	7	24	24	24	23	23	18	8	1
	9	24	24	24	24	24	21	18	8
	11	24	24	24	24	24	24	21	12
	13	24	24	24	24	24	24	24	18
	15	24	24	24	24	24	24	24	23
		Erkennung von Schnäbeln							
1	5	19	19	19	19	19	13	5	2
	9	19	19	19	19	19	19	19	19
	13	19	19	19	19	19	19	19	19
	17	19	19	19	19	19	19	19	19
	21	19	19	19	19	19	19	19	19
	25	19	19	19	19	19	19	19	19

Tabelle B.9: **Anzahl erkannter Augenbrauen und Schnäbel abhängig von Quantisierung, Ortstoleranz und Schwelle.** Die Messung ergänzt die Tabellen B.5 bis B.8. Die Modelle für Augenbrauen und Schnäbel sind die gleichen wie in den vorigen Versuchen. Die Quantisierung bei der Augenbrauenerkennung bezieht sich daher wieder auf die Orientierung von Kanten, die der Skelettmerkmale wieder auf die Orientierung der Skelettlinien.

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
	Erkennung von Augenbrauen							
3	378789	137539	29856	4495	740	33	2	0
5	733704	576264	391420	221836	109599	40385	9519	656
7	825634	727332	595311	443176	300019	167062	78060	19189
9	869963	799492	703512	582206	451185	307295	181167	72054
11	895053	846168	769345	673501	561031	426858	289502	147078
13	911519	876166	815072	737237	642349	522178	391117	233747
15	922913	896485	849611	784775	703699	598096	476404	322329
	Erkennung von Schnäbeln							
5	477386	347064	190393	66040	12783	1271	76	0
9	543278	518510	479913	405686	303773	191634	86108	19397
13	552350	543479	526518	495522	435202	346258	234002	117683
17	554791	550522	541921	525738	485511	410479	306972	188213
21	555763	553399	548138	539005	512270	446056	346032	226682
25	556244	554933	551608	545903	529030	470598	372617	253116

Tabelle B.10: **Falsche Treffer für rein geometrische Modelle.** Da bei der Objekterkennung keine verschiedenen Merkmalsausprägungen unterschieden wurden ($\zeta = 1$), basieren die Ergebnisse nur auf der Geometrie.

	Anzahl falscher Treffer für verschiedene Schwellen ϑ							
ς	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
3	4637	630	73	19	7	2	0	0
5	47767	11813	2151	485	175	49	20	0
7	156851	52385	14435	3761	1143	354	137	28
9	297822	129726	47745	14832	5127	1487	485	140
11	431413	227851	104951	39805	15403	5005	1416	404
13	543705	331776	177103	80911	35503	13295	4313	956
15	633802	429167	259015	134906	67527	28568	10249	2452

Tabelle B.11: **Falsche Treffer bei feiner Merkmalsquantisierung.** Die Objekterkennung beruht auf dem in Abbildung 6.11 gezeigten Augenbrauenmodell. Die Kantenorientierung wurde mit 40 Stufen sehr fein quantisiert.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[AG90] T. D. Albright and C. G. Gross. Do inferior temporal cortex neurons encode shape by acting as fourier descriptor filters? *Proceedings of the International Conference on Fuzzy Logic & Neural Networks, Izuka, Japan*, pages 375–378, 1990.

[AHK01] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, pages 420–434, London, UK, 2001. Springer-Verlag.

[AR05] N. C. Aggelopoulos and E. T. Rolls. Scene perception: inferior temporal cortex neurons encode the position of different objects in the scene. *European Journal of Neuroscience*, 22:2903–2916, 2005.

[ASG07] Y. Abramson, B. Steux, and H. Ghorayeb. Yet Even Faster (YEF) real-time object detection. In K.-D. Kuhnert and M. Stommel, editors, *International Journal of Intelligent Systems Technologies and Applications (IJISTA)*, volume 2, pages 102–112. Inderscience, 2007.

[BA04] P. Blackwell and D. Austin. Appearance Based Object Recognition with a Large Dataset using Decision Trees. *Proceedings of Australasian Conference on Robotics and Automation*, 2004.

[Bal81] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[Bau98] U. Bauer. *Computational Models of Neural Circuitry in the Macaque Monkey Primary Visual Cortex*. 1998. Dissertation, Technische Fakultät der Universität Bielefeld.

[BB82] D. H. Ballard and C. M. Brown. *Computer Vision*. Prentice-Hall, 1982.

[BB97] M. Burge and W. Burger. Learning Visual Ideals. *Proc. of the 9th ICIAP, Florence, Italy*, pages 316–323, 1997.

[BBB05] M. Basu, H. Bunke, and A. Del Bimbo. Guest Editors' Introduction to the Special Section on Syntactic and Structural Pattern Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1009–1012, 2005.

[BBM96] M. Burge, W. Burger, and W. Mayr. Recognition and learning with polymorphic structural components. *Proc. of the 13th ICPR, Vienna, Austria*, 1:19–28, 1996.

[BGRS99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "Nearest Neighbor" Meaningful. In *In Int. Conf. on Database Theory*, pages 217–235, 1999.

[BNM98] S. Baker, S. K. Nayar, and H. Murase. Parametric feature detection. *IJCV*, pages 27–50, 1998.

[BWP98] M. C. Burl, M. Weber, and P. Perona. A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*, pages 628–641, London, UK, 1998. Springer-Verlag.

[BZM08] A. Bosch, A. Zisserman, and X. Munoz. Scene Classification Using a Hybrid Generative/Discriminative Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4), 2008.

[CAC04] L. Cole, D. Austin, and L. Cole2. Visual Object Recognition using Template Matching. *Australasian Conference on Robotics and Automation ACRA*, 2004.

[CD01] B. G. Cumming and G. C. DeAngelis. The Physiology of Stereopsis. *Annual Review of Neuroscience*, 24:203–238, 2001.

[CFH05] D. J. Crandall, P. F. Felzenszwalb, and D. P. Huttenlocher. Spatial Priors for Part-Based Recognition Using Statistical Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 10–17, 2005.

[CH06] D. Crandall and D. Huttenlocher. Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 16–29, 2006.

[CH07] D. J. Crandall and D. P. Huttenlocher. Composite Models of Objects and Scenes for Category Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[CK04] C. M. Cyr and B. B. Kimia. A Similarity-Based Aspect-Graph Approach to 3D Object Recognition. *Int. J. Comput. Vision*, 57(1):5–22, 2004.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[Dac00] D. M. Dacey. Parallel Pathways for Spectral Coding in Primate Retina. *Annual Review of Neuroscience*, 23:743–775, 2000.

[DH97] Z. Dodds and G. D. Hager. A Color Interest Operator for Landmark-Based Navigation. In *AAAI/IAAI*, pages 655–660, 1997.

[Dis] W. Disney. *Lustiges Taschenbuch*, volume 204, 320, 323, 327, 328, 336, 357, 367, Spezial 13, Enten Edition 7, 20, Sonderband 12. Egmont Ehapa, Berlin.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.*, 39:1–38, 1977.

[DS87] R. Desimone and S. J. Schein. Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology*, 57:835–868, 1987.

[dSG01] E. M. d. Santos and H. M. Gomes. Appearance-Based Object Recognition Using Support Vector Machines. In *SIBGRAPI '01: Proceedings of the 14th Brazilian Symposium on Computer Graphics and Image Processing*, page 399, Washington, DC, USA, 2001. IEEE Computer Society.

[Duf99] N. Duffy. Computer Science Technical Report: Overview of Appearance Based Methods in Computer Vision. Technical Report TDC-CS-1999-51, Trinity College Dublin, Department of Computer Science, Oct 1999.

[Ebn98] M. Ebner. On the Evolution of Interest Operators using Genetic Programming. In Riccardo Poli, W. B. Langdon, Marc Schoenauer, Terry Fogarty, and Wolfgang Banzhaf, editors, *Late Breaking Papers at EuroGP'98: the First European Workshop on Genetic Programming*, pages 6–10, Paris, France, 14-15 1998. CSRP-98-10, The University of Birmingham, UK.

[EIP97] S. Edelman, N. Intrator, and T. Poggio. Complex cells and object recognition (submitted), 1997.

[Ett06] E. Ettelt. *Appearence Based Object Recognition by Use of Optimized Template Trees*. PhD thesis, TU Muenchen, Germany, 2006.

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 264–271, June 2003.

[FPZ06] R. Fergus, P. Perona, and A. Zisserman. A Sparse Object Category Model for Efficient Learning and Complete Recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 443–461. Springer, 2006.

[FPZ07] R. Fergus, P. Perona, and A. Zisserman. Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition. In *International Journal of Computer Vision*, volume 71, pages 273–303, March 2007.

[För86] W. Förstner. A feature based correspondence algorithm for image matching. *ISP Comm. III, Rovaniemi, Int. Arch. of Photogrammetry*, 26(3/3), 1986.

[FRAJ07] L. Franco, E. T. Rolls, N. C. Aggelopoulos, and J. M. Jerez. Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biological Cybernetics*, 96:547–560, 2007.

[FS99] Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Society for Artificial Intelligence*, 14(5):771–780, 1999.

[FT98] Márta Fidrich and Jean-Philippe Thirion. Stability of Corner Points in Scale Space: The Effects of Small Nonrigid Deformations. *Computer Vision and Image Understanding: CVIU*, 72(1):72–83, 1998.

[Fu82] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, 1982.

[GH88] W. E. L. Grimson and D. P. Huttenlocher. On the Sensitivity of the Hough Transform for Object Recognition. Technical Report A.I. Memo No. 1044, Massachusetts Institute Of Technology, Artificial Intelligence Laboratory, May 1988.

[GMP+07] P. D. R. Gamlina, D. H. McDougala, J. Pokornyb, V. C. Smithb, K.-W. Yauc, and D. M. Daceyd. Human and macaque pupil responses driven by melanopsin-containing retinal ganglion cells. *Vision Research*, 47(7):946–954, 2007.

[GPC98] S. Geman, D. F. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, LX:707–736, 1998.

[GSR03] D. B. Grimes, A. P. Shon, and R. P. N. Rao. Probabilistic Bilinear Models for Appearance-Based Vision. *ICCV*, 02:1478–1487, 2003.

[GT97] G. M. Ghose and D. Y. Ts'O. Form Processing Modules in Primate Area V4. *Journal of Neurophysiology*, 77:2191–2196, 1997.

[HAK00] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What Is the Nearest Neighbor in High Dimensional Spaces? In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 506–515, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[HE00] J. Hegde and D. C. Van Essen. Selectivity for Complex Shapes in Primate Visual Area V2. *Journal of Neuroscience*, 20, 2000.

[HE03] J. Hegde and D. C. Van Essen. Strategies of shape representation in macaque visual area V2. *Visual Neuroscience*, 20:313–328, 2003.

[HE07] J. Hegde and D. C. Van Essen. A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4. *Cerebral Cortex*, 17:1100–1116, 2007.

[HR00] S. H. C. Hendry and R. C. Reid. The Koniocellular Pathway in Primate Vision. *Annual Review of Neuroscience*, 23:127–153, 2000.

[HT00] N. Hadjikhani and R. B. H. Tootell. Projection of Rods and Cones Within Human Visual Cortex. *Human Brain Mapping*, 9:55–63, 2000.

[Hub88] D. H. Hubel. *Eye, Brain and Vision.* Scientific American Library, 1988.

[HZ05] F. Han and S. C. Zhu. Bottom-up/Top-Down Image Parsing by Attribute Graph Grammar. *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, 2:1778–1785, 2005.

[JSWP07] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV)*, 2007.

[JVO00] P. Janssen, R. Vogels, and G. Orban. Three-Dimensional Shape Coding in Inferior Temporal Cortex. *Neuron*, 27(2):385–397, 2000.

[KBV03] G. Kayaert, I. Biederman, and R. Vogels. Shape Tuning in Macaque Inferior Temporal Cortex. *The Journal of Neuroscience*, 23(7):3016–3027, 2003.

[KC03] J. Kerr and P. Compton. Towards Generic Model-based Object Recognition by Knowledge Acquisition and Machine Learning. In G. Tecuci, editor, *Proceedings of the IJCAI-2003 Workshop on Mixed Initiative Intelligent Systems*, pages 107–113, Acapulco, Mexico, 2003. Morgan Kaufmann.

[KLSK07] K.-D. Kuhnert, M. Langer, M. Stommel, and A. Kolb. Dynamic 3d-vision, 2007.

[Kol03] H. Kolb. How the Retina Works. *American Scientist*, 91(1):28–35, 2003.

[KS06] K.-D. Kuhnert and M. Stommel. Fusion of Stereo-Camera and PMD-Camera Data for Real-Time suited precise 3D Environment Reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4780–4785, October 9-15, 2006. Also presented at the Robotic 3D Environment Cognition Workshop at the Spatial Cognition, Bremen, Germany, September 24-28, 2006.

[KWU00] S. Kastner, P. De Weerd, and L. G. Ungerleider. Texture Segregation in the Human Visual Cortex: A Functional MRI Study. *Journal of Neurophysiology*, 83:2453–2457, 2000.

[Lan83] E. H. Land. Recent advances in retinex theory and some implications for cortical computations: Color vision and the natural image. *Proc. Natl. Acad. Sci. USA*, 80(16):5163–5169, 1983.

[LBL+03] I. Levner, V. Bulitko, L. Li, G. Lee, and R. Greiner. Learning Robust Object Recognition Strategies. *The 8th Australian and New Zealand Conference on Intelligent Information Systems*, 2003.

[LCH98] B. Luo, A. Cross, and E. Hancock. Corner Detection via Topographic Analysis of Vector Potential. In *Proceedings of the 9th British Machine Vision Conference*, 1998.

[Lee96] L. Lee. Learning of Context-Free Languages: A Survey of the Literature. Technical Report TR-12-96, Harvard University, 1996.

[Lee00] T. M. C. Lee. Memory After Temporal Lobectomy. *Revista Espanola de Neuropsicologia*, 2(1-2):46–59, 2000.

[LH87] M. S. Livingstone and D. H. Hubel. Psychophysical Evidence for Separate Channels for the Perception of Form, Color, Movement, and Depth. *Journal of Neuroscience*, 7(11):3416–3468, 1987.

[LMS06a] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *Proceedings of the 17th British Machine Vision Conference (BMVC)*, Edinburgh, UK, September 2006.

[LMS06b] B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation Based Multi-Cue Integration for Object Detection. In *Proceedings of the 17th British Machine Vision Conference (BMVC)*, September 2006.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999.

[LS03] B. Leibe and B. Schiele. Analyzing Contour and Appearance Based Methods for Object Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[LXL04] C. Liu, T. Xia, and H. Li. A hierarchical Hough transform for fingerprint matching. In D. Zhang and A. K. Jain, editors, *Biometric Authentication*, pages 373–379, Hong Kong, 2004. Springer.

[Mar82] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman and Company, 1982.

[Miy93] Y. Miyashita. Inferior Temporal Cortex: Where Visual Perception Meets Memory. *Annual Reviews of Neuroscience*, 16:245–265, 1993.

[MLS05] K. Mikolajczyk, B. Leibe, , and B. Schiele. Local Features for Object Class Recognition. In *International Conference on Computer Vision (ICCV'05)*, October 2005.

[MLS06] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple Object Class Detection with a Generative Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, June 2006.

[MN93] H. Murase and S. K. Nayar. Learning Object Models from Appearance. In *Proc. of AAAI*, pages 836–843, Washington D.C., USA, 1993.

[MN95] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. Journal of Computer Vision*, 14(1):5–24, January 1995. Kluwer.

[Mot01] M. Motter. Statistische Modellierung von Bildinhalten fuer die Bilderkennung. Master's thesis, Faculty of Mathematics and Computer Science, University RWTH Aachen, Germany, 2001.

[MP69] M. L. Minsky and S. Papert. *Perceptrons: An introduction to computational geometry.* MIT Press, Cambridge, MA, USA, 1969.

[MP88] M. L. Minsky and S. Papert. *Perceptrons: An introduction to computational geometry, Third printing.* MIT Press, Cambridge, MA, USA, 1988.

[MRS08] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval (online version).* Cambridge University Press, 2008. URL: http://www-csli.stanford.edu/∼hinrich/information-retrieval-book.html.

[MS02] K. Mikolajczyk and C. Schmid. An Affine Invariant Interest Point Detector. *Proceedings of the 7th European Conference on Computer Vision (ECCV) - Part I*, pages 128–142, 2002.

[MSBJ07] S. Mikula, J. M. Stone, A. L. Berman, and E. G. Jones. *A digital stereotaxic atlas of the brain of the monkey, Macaca mulatta.* www.brainmaps.org, Center for Neuroscience, University of California, Davis and University of Wisconsin, Madison, 2007. Supported by Human Brain Project Grant Number MH/DA 52154 from the National Institutes of Health, United States Public Health Service.

[MSHM91] Y. Miyashita, K. Sakai, S.-I. Higuchi, and N. Masui. Localization of Primal Long-Term Memory in the Primate Temporal Cortex. In L. R. Squire, N. M. Weinberger, G. Lynch, and J. L. McGaugh, editors, *Memory: Organization And Locus of Change*, pages 239–249. Oxford University Press, 1991.

[MZS03] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape Recognition with Edge-Based Features. In *Proc. British Machine Vision Conf.*, 2003.

[NGDU93] H. Nakamura, R. Gattass, R. Desimone, and L. G. Ungerleider. The Modular Organization of Projections from Areas V1 and V2 to Areas V4 and TEO in Macaques. *Journal of Neuroscience*, 13(9):3681–3691, 1993.

[Nie02] A. Nieder. Die Wahrnehmung von Scheinkonturen - Wie sich das Gehirn Illusionen macht. *Neuroforum*, 3:210–217, 2002.

[NLR08] K. J. Nielsen, N. K. Logothetis, and G. Rainer. Object features used by humans and monkeys to identify rotated shapes. *Journal of Vision*, 8(2):1–15, 2008.

[NM94] S. Nayar and H. Murase. On the Dimensionality of Illumination Manifolds in Eigenspace. Technical Report CUCS021 -94, Department of Computer Science, Columbia University, New York, August 1994.

[NMN96] S. Nayar, H. Murase, and S. Nene. *Parametric appearance representation. In Early Visual Learning.* Oxford University Press, February 1996.

[NS98a] R. C. Nelson and A. Selinger. A Cubist Approach to Object Recognition. Technical Report TR689, Department of Computer Science, University of Rochester, NY, USA, 1998.

[NS98b] R. C. Nelson and A. Selinger. A Cubist Approach to Object Recognition. In *Proc. of the International Conference on Computer Vision (ICCV98), Bombay, India*, pages 614–621, 1998.

[NS98c] R. C. Nelson and A. Selinger. Large-Scale Tests of a Keyed, Appearance-Based 3-D Object Recognition System. *Vision Research, Special issue on computational vision*, 38(15-16), 1998.

[NS00] R. C. Nelson and A. Selinger. Learning 3D Recognition Models for General Object from Unlabeled Imagery: An Experiment in Intelligent Brute Force. In *International Conference on Pattern Recognition (ICPR00)*, volume 1, pages 1–8, Barcelona Spain, September 2000.

[NSM96] Y. Naya, K. Sakai, and Y. Miyashita. Activity of primate inferotemporal neurons related to a sought target in pair-association task. *Proceedings of the National Academy of Sciences of the United States of America PNAS*, 93:2664–2669, 1996.

[OB05] B. Ommer and J. M. Buhmann. Object Categorization by Compositional Graphical Models. *EMMCVPR'05, LNCS 3757*, pages 103–113, 2005.

[OB06] B. Ommer and J. M. Buhmann. Learning Compositional Categorization Models. In *ECCV'06*. LNCS 3953, Springer, 2006.

[OE97] J. F. Olavarria and D. C. Van Essen. The Global Pattern of Cytochrome Oxidase Stripes in Visual Area V2 of the Macaque Monkey. *Cerebral Cortex*, 7:395–404, 1997.

[OH97] C. F. Olson and D. P. Huttenlocher. Automatic Targed Recognition by Matching Oriented Edge Pixels. *IEEE Transactions On Image Processing*, 6(1):103–113, 1997.

[OSB06] B. Ommer, M. Sauter, and J. M. Buhmann. Learning Top-Down Grouping of Compositional Hierarchies for Recognition. *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 194–202, 2006.

[PDM+05] M. A. Pinsk, K. DeSimone, T. Moore, C. G. Gross, and S. Kastner. Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America PNAS*, 102(19):6996–7001, 2005.

[PF75] E. Persoon and K. S. Fu. Sequential classification of strings generated by SCFG's. *International Journal of Parallel Programming*, 4(3):205–217, 1975.

[PL96] A. R. Pope and D. G. Lowe. Learning Appearance Models for Object Recognition. In *Object Representation in Computer Vision*, pages 201–219, 1996.

[PL00] A. R. Pope and D. G. Lowe. Probabilistic Models of Appearance for 3-D Object Recognition. *International Journal of Computer Vision*, 40(2):149–167, 2000.

[PLRS04] J. Ponce, S. Lazebnik, F. Rothganger, and C. Schmid. Toward True 3D Object Recognition. *Congrès de Reconnaissance des Formes et Intelligence Artificielle, Toulouse, France*, 2004.

[PMS+05] H.-C. Pape, S. G. Meuth, T. Seidenbecher, T. Munsch, and T. Budde. Der Thalamus: Das Tor zum Bewusstsein und Rhythmusgenerator im Gehirn. *Neuroforum*, 2:44–54, 2005.

[PTRL03] C. L. Passaglia, J. B. Troy, L. Ruttiger, and B. B. Lee. Orientation sensitivity of ganglion cells in primate retina. *Vision Research*, 42(6):683–694, 2002-03.

[Qui93] J. R. Quinlan. *C4.5: Programms for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

[RDD+00] M. Reinhold, F. Deinzer, J. Denzler, D. Paulus, and J. Pösl. Active Appearance-Based Object Recognition Using Viewpoint Selection. *Proc. of the 5th International Fall Workshop Vision, Modeling and Visualization VMV*, pages 105–112, 2000.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*. Bradford Books (MIT Press), Cambridge, MA, 1986.

[RK04] S. Dutta Roy and N. Kulkarni. Active 3-D Object Recognition using Appearance-Based Aspect Graphs. *In Proc. IAPR-sponsored Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 40–45, 2004.

[Ros58] F. Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408, 1958.

[RP02] M. Riesenhuber and T. Poggio. Neural Mechanisms of Object Recognition. *Current Opinion in Neurobiology*, 12:162–168, 2002.

[RT95] A. W. Row and D. Y. Ts'o. Visual Topography in Primate V2: Multiple Representations across Functional Stripes. *Journal of Neuroscience*, 15(5):3689–3715, 1995.

[RT01] V. Roth and K. Tsuda. Pairwise coupling for machine recognition of handprinted japanese characters. *Computer Vision and Pattern Recognition (CVPR)*, pages 1120–1125, 2001.

[RTT97] E. T. Rolls, A. Treves, and M. J. Tovee. The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Experimental Brain Research*, 114(1):149–162, 1997.

[RW08] P. M. Roth and M. Winter. Survey of appearance-based methods for object recognition. Technical Report ICG-TR-01/08, Inst. for Computer Graphics and Vision, Graz University of Technology, Austria, 2008.

[SANM96] S. K. Nayar S. A. Nene and H. Murase. Columbia Object Image Library (coil-100). Technical Report CUCS-006-96, Columbia University, NY, February 1996.

[SB95] S. M. Smith and J. M. Brady. SUSAN – A new approach to low level image processing. Technical Report TR95SMS1c, Chertsey, Surrey, UK, 1995.

[SBO+07] D. Schleicher, L. M. Bergasa, M. Ocana, R. Barea, and E. Lopez. *Real-Time Stereo Visual SLAM in Large-Scale Environments Based on SIFT Fingerprints*, volume 4739 of *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, 2007.

[Sch05] A. Schurger. Mooney Faces, Art of Science Competition. Princeton University., 2005.

[SD90] S. J. Schein and R. Desimone. Spectral Properties of V4 Neurons in the Macaque. *Journal of Neuroscience*, 10(10):3369–3389, 1990.

[Sel01] A. Selinger. *Analysis and Applications of Feature-Based Object Recognition*. PhD thesis, University of Rochester, Computer Science Department, Rochester, NewYork, USA, July 2001.

[SK04] M. Stommel and K.-D. Kuhnert. Subpixel accurate segmentation of small images using level curves. In *Proc. Int'l Conf. on Computer Vision and Graphics 2004 (ICCVG'04)*, Warsaw, Poland, September 22-24, 2004.

[SK05] M. Stommel and K.-D. Kuhnert. Appearance based recognition of complex objects by genetic prototype-learning. In *Proc. 13th Int'l Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, Czech Republic, January 31 - February 4, 2005.

[SK06] M. Stommel and K.-D. Kuhnert. A Learning Algorithm for the Appearance-Based Recognition of Complex Objects. In *The 2006 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP 2006)*, In Proc. The 2006 International Conference on Machine Learning; Models, Technologies & Application (MLMTA'06), Las Vegas, Nevada, USA, June 26-29 2006.

[SLA02] A. Shashua, A. Levin, and S. Avidan. Manifold Pursuit: A New Approach to Appearance Based Recognition. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition*

(ICPR'02) Volume 3, page 30590, Washington, DC, USA, 2002. IEEE Computer Society.

[SM91] K. Sakai and Y. Miyashita. Neural organization for the long-term memory of paired associates. *Letters To Nature*, 354:152–155, 1991.

[SMT06] W. Suzuki, K. Matsumoto, and K. Tanaka. Neuronal Responses to Object Images in the Macaque Inferotemporal Cortex at Different Stimulus Discrimination Levels. *Journal of Neuroscience*, 26(41):10524–10535, 2006.

[SN99] A. Selinger and R. C. Nelson. A perceptual grouping hierarchy for appearance-based 3D object recognition. *Computer Vision and Image Understanding, Special issue on perceptual organization in computer vision*, 76(1):83–92, 1999.

[SO01] E. P. Simoncelli and B. A. Olshausen. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[SOP07] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104(15):6424–6429, 2007.

[SS07] M. Stark and B. Schiele. How Good are Local Features for Classes of Geometric Objects. *IEEE 11th International Conference on Computer Vision ICCV*, pages 1–8, 2007.

[SSS06] M. Schünke, E. Schulte, and U. Schumacher. *Prometheus*. Georg Thieme Verlag, 2006.

[SWP05a] T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[SWP05b] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994–1000, San Diego, CA, USA, 2005. IEEE Computer Society.

[SYK94] D. Shaked, O. Yaron, and N. Kiryati. Deriving stopping rules for the probabilistic hough transform bysequential analysis. In *Pattern Recognition. 1994. Vol. 2, Conference B: Computer Vision And Image Processing., Proceedings of the 12th IAPR International Conference on*, pages 229–234. IEEE Computer Society, 1994.

[SZ95] S. Shipp and S. Zeki. Segregation and convergence of specialised pathways in macaque monkey visual cortex. *Journal of Anatomy*, 187:547–562, 1995.

[Tan93] K. Tanaka. Column structure of inferotemporal cortex: "visual alphabet" or "differential amplifiers"? *Proceedings of 1993 International Joint Conference on Neural Networks IJCNN*, 2:1095–1099, 1993.

[Tan95] E. Tanaka. Theoretical aspects of syntactic pattern recognition. *Pattern Recognition*, 28(7):1053–1061, 1995.

[Tan96] K. Tanaka. Inferotemporal cortex and object vision. *Annual Reviews of Neuroscience*, 19:109–139, 1996.

[Tan00] K. Tanaka. Mechanisms of visual object recognition studied in monkeys. *Spatial Vision*, 13(2,3):147–163, 2000.

[TH01] R. B. H. Tootell and N. Hadjikhani. Where is 'Dorsal V4' in Human Visual Cortex? Retinotopic, Topographic and Functional Evidence. *Cerebral Cortex*, 4(11):298–311, 2001.

[Tom06] M. Tomono. 3-D Object Map Building Using Dense Object Models with SIFT-based Recognition Features. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006.

[TP89] M. Tarr and S. Pinker. Mental rotation and orientationdependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[TP91] M. Turk and A. Pentland. Face recognition using eigenfaces. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 586–591, 1991.

[TRS01] T. P. Trappenberg, E. T. Rolls, and S. M. Stringer. Effective Size of Receptive Fields of Inferior Temporal Visual Cortex Neurons in Natural Scenes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14(1)*, pages 293–300, Cambridge, MA, 2001. MIT Press.

[TT01] H. Tamura and K. Tanaka. Visual Response Properties of Cells in the Ventral and Dorsal Parts of the Macaque Inferotemporal Cortex. *Cerebral Cortex*, 11(5):384–399, 2001.

[vdHPB84] R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224:1260–1262, 1984.

[VJ04] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[Wal47] A. Wald. *Sequential Analysis*. New York: John Wiley and Sons/London: Chapman and Hall, 1947.

[Wal94] G. Wallis. *Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes*. 1994. Ph.D. thesis, Department of Experimental Psychology, Oxford Univeristy. www: ftp://ftp.mpiktueb.

[WSV01] N. Winters and J. Santos-Victor. Information Sampling for Appearance based 3D Object Recognition and Pose Estimation. In *Proceedings of the 2001 Irish Machine Vision and Image Processing Conference*, Maynooth, Ireland, September 2001.

[WT01] B. Willmore and D. J. Tolhurst. Characterising the sparseness of neural codes. *Network, Computation in Neural Systems*, 12:255–270, 2001.

[WWP00] M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *European Conference on Computer Vision (ECCV)*, pages 18–32, 2000.

[WZ06] X. Wang and H. Zhang. Good Image Features for Bearing-only SLAM. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2576–2581, 2006.

[XWF03] Y. Xiao, Y. Wang, and D. J. Felleman. A spatially organized representation of colour in macaque cortical area V2. *Letters To Nature*, 421:535–539, 2003.

[ZHP+07] F. H. Zaidi, J. T. Hull, S. N. Peirson, K. Wulff, D. Aeschbach, J. J. Gooley, G. C. Brainard, K. Gregory-Evans, J. F. Rizzo, C. A. Czeisler, R. G. Foster, M. J. Moseley, and S. W. Lockley. Short-Wavelength Light Sensitivity of Circadian, Pupillary, and Visual Awareness in Humans Lacking an Outer Retina. *Current Biology*, 17(24):2122–2128, 2007.