Zur Erkennung verformbarer Objekte anhand ihrer Teile

Martin Stommel

Arbeitsgruppe Künstliche Intelligenz, Universität Bremen
Am Fallturm 1, 28359 Bremen
mstommel@tzi.de

Abstract: Aufgrund der Vielzahl möglicher visueller Erscheinungen lassen sich verformbare Objekte mit den Mitteln der digitalen Bildverarbeitung nur schwer zuverlässig erkennen. Zur Lösung dieses Problems wird in dieser Arbeit ein kompositioneller Ansatz untersucht, bei dem ein Objekt als Hierarchie von Teilen und Unterteilen in geometrischen Beziehungen beschrieben wird. Für jedes Teil läßt sich die Behandlung der Ausprägung und der Position lokaler Merkmale gezielt parametrisieren, was eine hohe Flexibilität ergibt. Die Parametrisierung des Modells beruht auf Beobachtungen der Statistik von Merkmalsverbünden, ihren geometrischen Eigenschaften und Abhängigkeiten von der Hierarchieebene. Die Methode ist ferner durch die Modellierung mehrerer Objektansichten und die gleichzeitige Lokalisation und Klassifikation gekennzeichnet. Die Leistungsfähigkeit des 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.

1 Erkennung verformbarer Objekte

Viele technische Anwendungen [Naw01] basieren auf der automatischen Auswertung von Bildern durch einen Rechner, da Kameras schnell sind, die meisten Oberflächen erkennen und berührungslos sowohl nahe als auch ferne Objekte aufnehmen. Die Qualität eines Bildverarbeitungssystems hängt stark davon ab, in welchem Umfang man Einfluß auf verschiedene Umwelteinflüsse und den technischen Aufbau am Einsatzort nehmen kann. Die Erscheinung eines Objekts hängt beispielsweise von der Beleuchtung, der Objektoberfläche soweit vorhanden und der Lage des Objekts ab. Eine besondere Schwierigkeit stellen verformbare Objekte dar, da diese besonders stark in ihrer Erscheinung variieren. Oft müssen für diese Einflüsse enge Einsatzgrenzen definiert werden, um eine bestimmte Zuverlässigkeit des Gesamtsystems gewährleisten zu können. Einen weiteren Einfluß stellt die Kamera selbst dar, die geometrische Verzerrungen, Rauschen, Zeilenverschiebungen, Helligkeitsänderungen oder Kompressionsartefakte bewirken kann. Solche Einflüsse sind allerdings in der Regel sehr viel schwächer als die Variabilität der Objekte an sich.

Die Erkennung verformbarer Objekte wird in dieser Arbeit genauer untersucht. Den Schwerpunkt bildet die Analyse der für das Training wichtigen Einflußgrößen, insbesondere die Geometrie. Der vorgestellte Ansatz leistet dabei sowohl eine Lokalisation der Ob-

jekte im Bild als auch deren Klassifikation. Die Wirksamkeit der entwickelten Methoden wird am Beispiel einer Cartoon-Datenbank gezeigt.

2 Grundsätzlicher Aufbau eines Bildverarbeitungssystems

Die Objekterkennung im Rechner geschieht üblicherweise durch eine Vorverarbeitung, eine Merkmalsextraktion und eine anschließende Klassifikation [Ros69]. Die Vorverarbeitung dient dazu, die Bildqualität soweit zu erhöhen (z.B. durch Rauschfilterung), daß im nächsten Schritt robuste Merkmale berechnet werden können. Merkmale repräsentieren typische Eigenschaften von Objekten. Die Klassifikation ordnet das Bildmaterial anhand der Merkmale bestimmten Klassen von Objekten zu. Die typischen Merkmale einer Klasse sind das *Modell*.

Durch einen Experten speziell auf ein Anwendungsgebiet zugeschnittene Merkmale erreichen oft hohe Erkennungsraten. Eine stärkere Übertragbarkeit auch auf andere Anwendungsgebiete erhofft man sich jedoch von der Kombination eher allgemeingültiger Merkmale und dem Einsatz heuristischer Lernverfahren, die sehr gute Modelle aus einer repräsentativen Trainingsstichprobe ableiten können. Die Güte des Modells wird durch die Klassifikation einer separaten Teststichprobe geschätzt. Die Ergebnisse lassen sich als Vertauschungs- oder Wahrheitstabelle, oder kompakter durch die *Korrektklassifikationsrate* (engl. accuracy), den *positiven Vorhersagewert* (engl. precision) und die *Sensitivität* (engl. recall) darstellen. Letztere geben an, welcher Anteil der Stichprobe korrekt klassifiziert wurde, welcher Anteil der Zuordnungen zu einer bestimmten Klasse korrekt ist, und welcher Anteil einer bestimmten Klasse korrekt erkannt wurde.

Aufgrund gewisser Probleme, die tatsächliche dreidimensionale Struktur von Objekten aus einer Trainingsstichprobe zuverlässig rekonstruieren zu können [NMN96, Low99, Sel01, BBM96, BB97, PLRS04], werden derzeit meistens erscheinungsbasierte Modelle eingesetzt. Diese speichern Objektansichten unter verschiedenen Perspektiven und Beleuchtungsarten. Der folgende Abschnitt gibt einen Überblick.

3 Stand der Forschung

Die aktuellen Verfahren unterscheiden sich bezüglich ihres hierarchischen Aufbaus, der Lernbarkeit lokaler Merkmale und der Berücksichtigung von Geometrieinformationen.

Im Falle des erscheinungsbasierten Ansatzes von Murase und Nayar [NMN96] enthält das Modell nur eine Hierarchieebene. Globale Objektansichten ohne eine weitere Zerlegung in einfachere Teile werden mittels Hauptkomponentenanalyse zwischen prototypischen Ansichten interpoliert. Pose und Beleuchtungssituation sind im Modell enthalten. Zur Behandlung verformbarer Objekte und texturierter Hintergründe wurden teilebasierte Ansätze mit einer [WWP00, FPZ06, CH06, VJ04] oder mehreren [Sel01, OB06, SWP05] Hierarchieebenen entworfen.

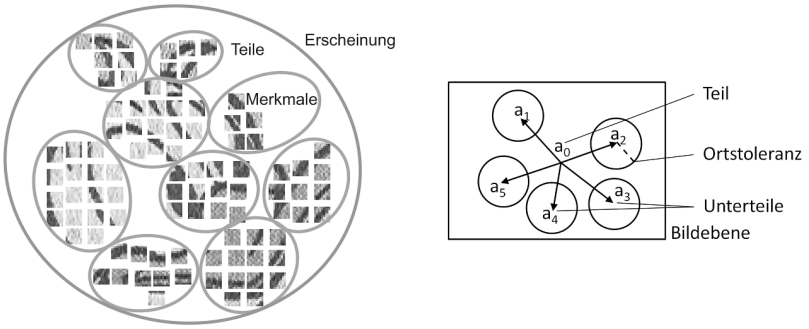

Abbildung 1: Links: Zerlegung eines Bildes in eine Hierarchie von Teilen. Rechts: Zerlegung des Teils a_0 in eine geometrische Anordnung der Teile $a_1 \ldots a_5$ mit einer gewissen Ortstoleranz.

Bei Ansätzen mit einer Ebene steht meistens die automatische Erlernbarkeit der Objektteile [WWP00, VJ04] oder deren geometrischer Abhängigkeiten [FPZ06, CH06] im Vordergrund. Um Teile zu erlernen, werden diese meistens durch hochdimensionale Merkmalsvektoren modelliert (etwa [Low99, BETvG06, KS04]), was den Einsatz von Clusterungsverfahren und weiteren Techniken des Machine Learnings erlaubt [MLS05, SS07, LS03, FPZ07]. Die Modellierung geometrischer Abhängigkeiten zwischen den Teilen ist eng mit den Cliquenproblem verbunden [CH06], was bei vollvernetzten Teilen die Modellgröße beschränkt [FPZ06]. Crandall und Huttenlocher stellen in diesem Zusammenhang Vorteile einer sternförmigen Vernetzung gegenüber einem einfachen Bag-of-Features-Modell fest, jedoch nicht für stärkere Vernetzungen.

Die Behandlung mehrerer Hierarchieebenen wird auf unterschiedliche Weise begründet. Serre et al. [SWP05] nehmen das menschliche Gehirn mit seiner hohen Leistungsfähigkeit zum Vorbild. Ommer und Buhmann [OB06] stellen Kompositionalität als allgemeingültiges und stabiles Prinzip dar mit Begründungen u.a. aus der Gestaltpsychologie. Selinger [Sel01] optimiert die Stufen der Hierarchie jeweils auf qualitativ leicht unterschiedliche Aspekte der Objekterkennung. Die häufigsten Methoden, um mit teilebasierten Modellen Objekte zu erkennen, sind Mehrheitsentscheide [Sel01, MLS05] und Entscheidungsbäume [BBM96, BB97, VJ04, SK06].

4 Modellierung von Objekten als Teilehierarchie

Zur Modellierung verformbarer Objekte wird in dieser Arbeit ein kompositioneller Ansatz gewählt, bei dem ein Objekt oder Teil rekursiv durch eine Menge von Unterteilen in lockeren geometrischen Beziehungen beschrieben wird (illustriert in Abb. 1, links). Zu jedem Teil wird eine ideale Position relativ zum übergeordneten Teil gespeichert. Ein Teil wird detektiert, wenn gemessen an einem Schwellwert ϑ eine ausreichende Anzahl von Unterteilen innerhalb einer maximalen Ortstoleranz ς um die jeweilige Idealposition gefunden wurde (s. Abb. 1, rechts). Durch die Wahl der Parameter ϑ und ς kann für jedes

Teil ein individueller Kompromiß zwischen geometrisch strikter Teilekonstellation und losem Bag-of-Features getroffen werden. Die Rekursion endet an Basismerkmalen, die nicht weiter unterteilt werden. Als Basismerkmale werden hier orientierte Kanten, Ecken und Skelettlinien gewählt, deren Allgemeingültigkeit sich bereits in früheren Tests auf realem Videomaterial zeigte [SK06].

Zur Objekterkennung wird in einem Bottom-up-Prozeß für jedes Teil des Modells ein Voting-Verfahren mit Verwandtschaft zur verallgemeinerten Hough-Transformation durchgeführt. Dieses leistet eine gleichzeitige Detektion und Klassifikation. Durch die Reduktion des Verfahrens auf einfache Bitoperationen wird ein sehr hoher Durchsatz erzielt. Für die Details sei auf die Originalarbeit verwiesen [Sto10].

5 Visuelle Alphabete auf mehreren Abstraktionsebenen

Um die für das Training ausschlaggebenden Größen zu ermitteln, wird das Bildmaterial auf verschiedenen Abstraktionsebenen analysiert. Gegenüber der Anwendung allgemeingültiger Optimierungsverfahren steht hier die Ableitung von Regeln im Vordergrund, die den Trainingsprozeß stärker erklären. Insbesondere soll geklärt werden, wodurch sich Teile genau definieren, welcher Zusammenhang zwischen Merkmalsausprägung und Geometrie der Teilekonstellation besteht und wie mehrere Ansichten modelliert werden. Die folgenden Abschnitte fassen die wesentlichen Ergebnisse geordnet nach der Hierarchieebene des Modells zusammen.

5.1 Extraktion lokaler Merkmale

Die niedrigste Hierarchieebene wird durch Kanten, Ecken und Flächen als Basismerkmale gebildet. Die Koordinaten der Merkmale werden operatorbasiert detektiert. Für jede Koordinate wird ein Deskriptor berechnet, welcher den Wert des Merkmals enthält. Für Kanten wird die Orientierung, für Ecken das Auftreten selbst gespeichert. Flächen werden durch die Punkte auf den Skelettlinien repräsentiert, welche durch die Orientierung der Skelettlinie, die Pixelintensität und den Abstand zur nächsten begrenzenden Kante parametrisiert sind.

Um jedem Merkmal Teile auf der nächsthöheren Ebene zuordnen zu können, werden die Deskriptoren diskretisiert und abgezählt. Dabei ist abzuwägen, daß durch zu grobe Intervalle Information verloren geht. Andererseits erhöhen feine Intervalle den Speicheraufwand, da für jede Merkmalsausprägung Hypothesen zu abstrakteren Teilen aufgestellt werden. Theoretische Betrachtungen sagen für den rauschbehafteten Fall eine Abnahme des Informationsgehalts sowohl bei einer zu groben als auch bei einer zu feinen Diskretisierung vorher. Praktische Messungen, die sowohl das Rauschen der Bildwerte als auch die Zeichnerische Variation berücksichtigen, zeigen einen grundsätzlich ähnlichen Verlauf, jedoch liegt die maximale Erkennungsgüte bei einer gröberen Diskretisierung von nur 8 bis 10 Stufen bei Orientierungswerten. Die Ursache liegt in statistischen Abhängigkeiten

räumlich benachbarter Merkmale, die Fehldetektionen des jeweils anderen Merkmals kompensieren können. Dabei werden auch starke Abhängigkeiten von der Ortstoleranz ς und der Schwelle ϑ gefunden, welche den Einfluß der Quantisierungsfeinheit überwiegen. Die hohe Abhängigkeit von der Ortstoleranz spricht gegen ein Bag-of-Features-Modell auf dieser niedrigen Ebene.

5.2 Kritische Variablen für die Teile-Modellierung

Die zweite Hierarchieebene soll Gruppen von Merkmalen zu Teilen zusammenfassen. Eine Reihe von Experimenten wird durchgeführt, um die statistischen und geometrischen Eigenschaften von Objektdarstellungen bei kompositioneller Modellierung zu klären.

Zunächst wird ein gieriger Algorithmus entworfen, der die Klassifikationsgüte durch schrittweise Hinzunahme von neuen Merkmalen optimiert. Dabei zeigt sich, daß diese Methode keinen gültigen Pfad durch den Konfigurationsraum hin zu einem guten Modell findet, da die Wertelandschaft viele suboptimale Nebenmaxima aufweist. Es wird daher nach weiteren Abhängigkeiten in den Parametern gesucht, die den Trainingsprozeß leiten können. Eine experimentelle Untersuchung der Parameter ς und ϑ zeigt eine starke gegenseitige Abhängigkeit sowie kompaktere und besser lokalisierte Trefferbereiche für hohe Schwellen. Ein Vergleich der detektierten Bildpositionen kleiner Beispielobjekte zeigt eine hohe Übereinstimmung von 44% bis 65% für eine rein auf Geometrie ausgelegte Modellkonfiguration im Vergleich zu einer gemischt auf Geometrie und Merkmalsausprägung ausgelegte Konfiguration. Da dies erneut einen hohen Einfluß der Geometrie anzeigt, werden in einem weiteren Versuch die Verbundwahrscheinlichkeiten aller geometrischen Konstellationen von Paaren von Merkmalen berechnet. Die resultierenden Histogramme zeigen oft räumliche Cluster, die möglicherweise wichtige Positionen für zu modellierende Teile darstellen. Aufgrund der oft spärlichen Statistik erscheint eine entsprechende automatische Optimierung aber nicht angemessen. Bei sehr häufigen Merkmalskombinationen dominiert der Abstand der Merkmale klar jede feinere räumliche Aufteilung. Messungen der Positionsabweichungen bei der Suche nach bestimmten Merkmalskonstellationen identifizieren den räumlichen Abstand als entscheidenden (und praktisch linearen) Einfluß auf die Stabilität bei der Detektion von Merkmalsverbünden.

Auf Basis dieser statistischen Eigenschaften können Gruppen von Merkmalen durch räumliche Clusterung zusammengestellt werden. Sollen so erzeugte Teile zu größeren Objekten zusammengesetzt werden, müssen jedoch weitere Abhängigkeiten zwischen der Teilegröße, der Objektgröße und der Stichprobenabdeckung beachtet werden.

5.3 Erzeugung eines visuellen Alphabets von Teilen

In Anlehnung an Erkenntnisse zur Funktionsweise des Sehzentrums im Gehirn [Tan96] wird ein *visuelles Alphabet* von Teilen erzeugt, welches unter den im vorigen Abschnitt genannten Randbedingungen parametrisiert wird. Die Idee ist, mehrere Teile des Alphabets

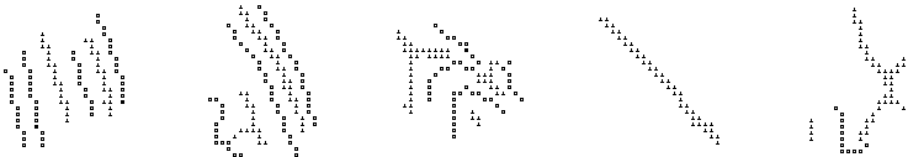

Abbildung 2: Beispiele aus dem Teilealphabet, hier mit einem Durchmesser von etwa 20 Pixeln. Die linken zwei Muster sind Beispiele des gleichen Clusters. Quadrate stehen für Kantenmerkmale, Dreiecke für Skelettlinien.

zu komplexeren Objektansichten zu kombinieren. Kandidaten für die Teile des Alphabets ergeben sich aus einer hierarchischen Clusterung der Basismerkmale nach ihren Bildpositionen. Die Merkmale eines Clusters werden wie in Abb. 1 (rechts) dargestellt als Unterteile eines abstrakteren Teils zusammengefaßt. Aus rechentechnischen Gründen müssen die sich aus der verwendeten Stichprobe ergebenden etwa 3 Millionen Kandidaten auf eine rechentechnisch behandelbare, dabei aber auch visuell ausreichend vielfältige Untermenge verdichtet werden. Hierzu wird erneut ein hierarchisches Clusterungsverfahren eingesetzt. 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. Verschiedene Arten visueller Ähnlichkeit lassen sich an der hierarchischen Clusterstruktur ablesen. Zur Festlegung einer bestimmten Clusterung wird die Homogenität der Cluster automatisch analysiert. Das visuelle Alphabet wird aus den Clusterprototypen erstellt. Das resultierende Alphabet enthält 5000 Teile mit einer Größe von 10 bis 60 Pixeln und Ortstoleranzen zwischen 2 und 10 Pixeln. Abbildung 2 zeigt einige Beispiele.

5.4 Modellierung mehrerer Ansichten eines Objekts

Da bei verformbaren Objekten die visuelle Erscheinung stark variiert, ist es zweckmäßig, die verschiedenen in einer Stichprobe auftretenden Darstellungen in möglichst homogene Gruppen zu unterteilen und diese getrennt zu modellieren. So entsteht ein neues visuelles Alphabet, dessen Elemente jeweils Gruppen von Beispielansichten repräsentieren. Die Gruppen werden durch eine dritte hierarchische Clusterung ermittelt, welche ein Dendrogramm der Stichprobenelemente erzeugt. Um Distanzen zwischen den Elementen zu berechnen, werden diese durch binäre Vektoren dargestellt, welche für jeden Eintrag des Alphabets angeben, ob das entsprechende Teil in dem Bild detektiert wird. Für eine Stichprobe von 800 Trainingsbildern ergibt sich ein Dendrogramm der Tiefe 15. Um eine Gruppe von Stichprobenelementen als Ansicht zu modellieren, werden alle Teile kombiniert, die sich bei allen Elementen an den gleichen Positionen detektieren lassen. Eine Unterabtastung im Rahmen der Ortstolerant ς begrenzt die Modellgröße auf ein praktikables Maß.

Experimente auf einzelnen Ansichtsmodellen zeigen starke Abhängigkeiten der optimalen Parameter von den strukturellen Eigenschaften des Dendrogramms, insbesondere starke

Korrelationen von 70 bis 92 Prozent zwischen den Parametern ς, ϑ und der Höhe bzw. Tiefe eines Knotens im Baum. Auch für die Güte der Erkennung und die Modellgröße existiert so ein Zusammenhang. Für den rechentechnisch behandelbaren Parameterbereich läßt sich ein lineares Modell für die Ortstoleranz finden, das eine Minimierung falsch Positiver Treffer über die Wahl der Schwelle ϑ zuläßt. Es ergeben sich Ortstoleranzen zwischen 15 und 75 Pixeln.

5.5 Training auf der Kategorieebene

Auf der höchsten Ebene des Modells werden jeweils mehrere Ansichtsmodelle unter einen Knoten zusammengefaßt, um die gesamte Stichprobe abzudecken. Auftretende Redundanzen werden zur Rauschfilterung und zur Anpassung des Gesamtsystems an ein mögliches Anwendungsziel genutzt. Da die Ansichtsmodelle bereits vollständige Objekte darstellen, werden die Knoten auf dieser Ebene als Bag-of-Features eingestellt. Außerdem kann den Knoten auf dieser Ebene schon eine Klasse zugeordnet werden, sodaß der Bottom-up-Erkennungsprozeß hier endet.

Redundanzen werden ermittelt, indem für jedes Element der Trainingsstichprobe die erkennenden Ansichtsmodelle ermittelt und kombiniert werden. Das Modell wird auf einen hohen positiven Vorhersagewert optimiert, indem die Schwelle ϑ des neuen Knotens höher als das auftretende Rauschen bei der Detektion eingestellt wird. Eine sensitivere Konfiguration ergibt sich dagegen, wenn die Schwelle auf einen hohe Korrektklassifikationsrate über alle Trainingselemente optimiert wird.

6 Test auf der Cartoon-Stichprobe

Die Leistungsfähigkeit der entwickelten Methoden wird am Beispiel einer Cartoon-Datenbank gezeigt. 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 [SK06] und Legosteine [CAC04], oder eher ungeordnete Bildersammlungen aus Suchmaschinenabfragen [GHP07].

Abbildung 3 zeigt Beispiele aus der Cartoon-Datenbank. Die 1600 Positivbeispiele zeigen Bilder des Kopfes der Disney-Figur *Donald*, welche aus der Buchvorlage [Dis] gescannt wurden. Die 1600 Negativbeispiele sind zufällige, quadratische Bildausschnitte ohne Überlappung mit Donald-Köpfen. Die Größe der Negativbeispiele spiegelt die Verteilung der Größe der Donald-Köpfe wieder.

Die auffälligsten Merkmale der Cartoon-Stichprobe sind die starken geometrischen Verformungen des dargestellten Objekts sowie die Vielfalt an Objektposen und Perspektiven. 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

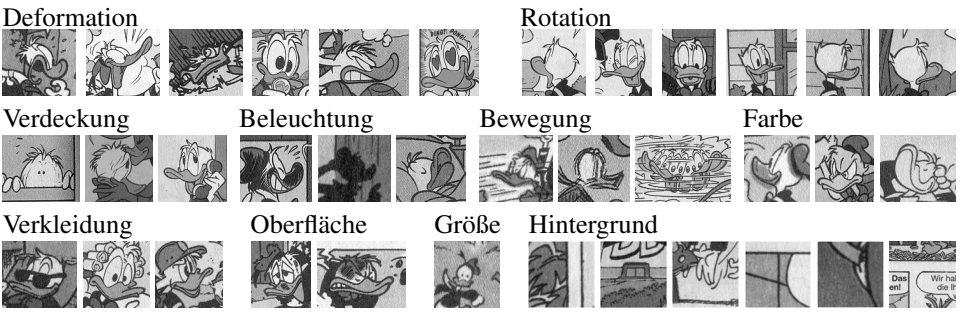

Abbildung 3: Beispiele aus der Cartoon-Stichprobe [Dis]

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.

Wie bei Aufnahmen natürlicher Objekte treten Verdeckungen, Beleuchtungseffekte, Bewegungsunschärfen und seltener Spiegelungen auf, welche z.B. durch Schraffuren, schwarze Flächen oder quer durch das Objekt verlaufende Striche angedeutet werden. Die Objektoberflächen sind vereinfacht einfarbig und meistens matt dargestellt. Erschwernisse für die Bildverarbeitung ergeben sich dagegen aus der Zeichen- und Reproduktionstechnik. Die Farbe einer Fläche muß über einen gewissen Bereich des Halbtonrasters gemittelt werden. Die Farbfüllungen der Flächen stimmen nicht immer mit den schwarzen Strichen überein. Die Farbe wird auch gelegentlich variiert, um emotionale Nebenbedeutungen zu illustrieren.

Für die Experimente wird die Stichprobe randomisiert und in gleich große Trainings- und Testmengen unterteilt. Die erste in Abschnitt 5.5 beschriebene Modellkonfiguration erreicht im Test einen positiven Vorhersagewert von 97% bei einer Korrektklassifikationsrate von 78%. Die Erkennung ist daher sehr zuverlässig aber auch recht selektiv. Die zweite Konfiguration erreicht eine Sensitivität von 89% bei einer Korrektklassifikationsrate von 77%. Das Modell ist daher ähnlich leistungsfähig wie das erste aber weniger konservativ.

7 Fazit

In dieser Arbeit wird ein System zur visuellen Erkennung verformbarer Objekte vorgestellt, das neuartig ist bezüglich des Trainings von visuellen Alphabeten auf mehreren Abstraktionsebenen. Ein besonderes Kennzeichen ist, daß für alle Teile des Modells eine individuelle Gewichtung von räumlicher Teile-Anordnung und Merkmalscharakteristik einstellbar ist. Dadurch konnte u.a. die Frage nach der Bedeutung der Geometrie [CH06, FPZ06] bezüglich ihrer Abhängigkeit von der Abstraktionsstufe präzisiert und beantwortet werden. Für eine Klasse von stark in der Erscheinung variierenden Objekten konnte so eine hohe Erkennungsrate erzielt werden.

Literatur

[BB97] M. Burge und W. Burger. Learning Visual Ideals. *Proc. of the 9th ICIAP, Florence, Italy*, Seiten 316–323, 1997.

[BBM96] M. Burge, W. Burger und W. Mayr. Recognition and learning with polymorphic structural components. *Proc. of the 13th ICPR, Vienna, Austria*, 1:19–28, 1996.

[BETvG06] H. Bay, A. Ess, T. Tuytelaars und L. van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2006.

[CAC04] L. Cole, D. Austin und L. Cole2. Visual Object Recognition using Template Matching. *Australasian Conference on Robotics and Automation ACRA*, 2004.

[CH06] D. Crandall und D. Huttenlocher. Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition. In *Proc. of European Conf. on Computer Vision (ECCV)*, Seiten 16–29, 2006.

[Dis] W. Disney. *Lustiges Taschenbuch*, Jgg. 204, 320, 323, 327, 328, 336, 357, 367, Spezial 13, Enten Edition 7, 20, Sonderband 12. Egmont Ehapa, Berlin.

[FPZ06] R. Fergus, P. Perona und A. Zisserman. A Sparse Object Category Model for Efficient Learning and Complete Recognition. In J. Ponce, M. Hebert, C. Schmid und A. Zisserman, Hrsg., *Toward Category-Level Object Recognition*, Jgg. 4170 of *LNCS*, Seiten 443–461. Springer, 2006.

[FPZ07] R. Fergus, P. Perona und A. Zisserman. Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition. In *International Journal of Computer Vision*, Jgg. 71, Seiten 273–303, Marz 2007.

[GHP07] G. Griffin, A. Holub und P. Perona. Caltech-256 Object Category Dataset. Bericht 7694, California Institute of Technology, 2007.

[KS04] Y. Ke und R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. *Computer Vision and Pattern Recognition (CVPR)*, 2:506–513, 2004.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of International Conference on Computer Vision (ICCV)*, Seiten 1150–1157, 1999.

[LS03] B. Leibe und B. Schiele. Analyzing Contour and Appearance Based Methods for Object Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[MLS05] K. Mikolajczyk, B. Leibe, und B. Schiele. Local Features for Object Class Recognition. In *International Conference on Computer Vision (ICCV'05)*, October 2005.

[Naw01] R. Nawrath. Industrielle Bildverarbeitung in Schleswig-Holstein. Bericht, Technologiestiftung Schleswig-Holstein (TSH), 2001.

[NMN96] S. Nayar, H. Murase und S. Nene. *Parametric appearance representation. In Early Visual Learning.* Oxford University Press, February 1996.

[OB06] B. Ommer und J. M. Buhmann. Learning Compositional Categorization Models. In *ECCV'06*. LNCS 3953, Springer, 2006.

[PLRS04] J. Ponce, S. Lazebnik, F. Rothganger und C. Schmid. Toward True 3D Object Recognition. *Congrès de Reconnaissance des Formes et Intelligence Artificielle, Toulouse, France*, 2004.

[Ros69] A. Rosenfeld. Picture Processing by Computer. *ACM Computing Surveys (CSUR)*, 1(3):147–176, 1969.

[SANM96] S. K. Nayar S. A. Nene und H. Murase. Columbia Object Image Library (COIL-100). Bericht CUCS-006-96, Columbia University, NY, February 1996.

[Sel01] A. Selinger. *Analysis and Applications of Feature-Based Object Recognition*. Dissertation, University of Rochester, Computer Science Department, Rochester, NewYork, USA, July 2001.

[SK06] M. Stommel und 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.

[SS07] M. Stark und B. Schiele. How Good are Local Features for Classes of Geometric Objects. *IEEE 11th International Conference on Computer Vision ICCV*, Seiten 1–8, 2007.

[Sto10] M. Stommel. *Zur Erkennung verformbarer Objekte anhand ihrer Teile*. 2010. Dissertation, Fachbereich Elektrotechnik und Informatik, Universität Siegen, http://dokumentix.ub.uni-siegen.de/opus/volltexte/2010/470/index.html.

[SWP05] T. Serre, L. Wolf und T. Poggio. A new biologically motivated framework for robust object recognition. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[Tan96] K. Tanaka. Inferotemporal cortex and object vision. *Annual Reviews of Neuroscience*, 19:109–139, 1996.

[VJ04] P. Viola und M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[WWP00] M. Weber, M. Welling und P. Perona. Unsupervised Learning of Models for Recognition. In *European Conference on Computer Vision (ECCV)*, Seiten 18–32, 2000.

Martin Stommel erforscht seit 2008 als wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Künstliche Intelligenz der Universität Bremen die automatische Erkennung von Personen und Objekten in Videos. Er arbeitete von 2002 bis 2008 als wissenschaftlicher Mitarbeiter in der Fachgruppe Echtzeit Lernsysteme der Universität Siegen, wo er über das vorliegende Thema promovierte. Im Rahmen von Forschungs-, Lehr- und Industriearbeiten arbeitete er ferner auf den Gebieten der Stereobildverarbeitung, Dokumentaufbereitung und visuellen Robotersteuerung. Nach dem Studium der Technischen Informatik an der Universität Siegen von 1997 bis 2002 untersuchte er 2002 am Institut für Automatische Regelung der Technischen Hochschule Schlesien in Gliwice, Polen, statistische Ansätze zur Rauschfilterung von Bildern und zur Detektion von Personen an der Farbe.