Bildsuche, Objekterkennung und Diskriminative Modelle

Thomas Deselaers

Institut für Bildverarbeitung – Departement Informationstechnologie und Elektrotechnik
ETH Zürich – Zürich – Schweiz
Email: thomas@deselaers.de – Web: http://thomas.deselaers.de

Abstract: In der hier vorgestellten Dissertation [Des08] werden drei Themenbereiche bearbeitet, die sich mit der automatischen Suche bzw. dem automatischen Erkennen von Bildern befassen. Im ersten Teil wird die Suche nach Bildern in einer Bilddatenbank untersucht. Dabei wird ein besonderer Schwerpunkt auf die inhaltsbasierte Bildsuche gelegt. Weiterhin wird untersucht, wie dies mit textuellen Bildannotationen kombiniert werden kann. Es wird ein System entwickelt, in dem sich beliebige Informationsquellen zur Bildsuche gemeinsam verarbeiten lassen. Im zweiten Teil werden verschiedene Modelle vorgestellt, um Objekte in Bildern anhand lokaler Eigenschaften zu erkennen. Insbesondere wird bei der Entwicklung der Methoden darauf geachtet, Heuristiken zu vermeiden und alle verfügbaren Informationsquellen einzubeziehen, um zu einem sauberen Modell zu gelangen. Das entwickelte Modell kommt fast vollständig ohne Heuristiken aus, ist kleiner und somit einfacher anzuwenden und zu trainieren als vergleichbare Modelle. Die erzielten Ergebnisse sind vergleichbar oder besser als der Stand der Forschung. Im dritten Teil werden einige Ideen aus der Modellierung in der Objekterkennung aufgegriffen, weitergehend untersucht und für die Erkennung handgeschriebener Ziffern verfeinert. Das daraus resultierende Modell erlaubt erstmals neben den üblichen Modellparametern, auch die Verformungsparameter geschlossen zu trainieren. In allen drei Bereichen werden die entwickelten Methoden quantitativ auf Standarddatensätzen evaluiert und mit dem Stand der Technik verglichen. Dabei zeigt sich, dass die erzielten Ergebnisse mit denen in der Literatur vergleichbar sind oder diese übertreffen. In dieser Arbeit werden einige der Methoden aus dem zweiten Bereich vorgestellt und ein kurzer Ausblick auf die beiden anderen Bereiche gegeben.

1 Einführung

Eine geeignete Modellierung ist eine der wichtigsten Komponenten eines automatischen Bilderkennungssystems. Systeme, die selbständig den Inhalt von Bildern erkennen können, haben viele praktische Anwendungen, wie z.B. der Zugriff auf große Bilddatenmengen, die nicht oder nur teilweise mit beschreibendem Text versehen sind oder in der automatischen Erkennung handgeschriebener Zeichen. Eine besonders schwierige Anwendung ist dabei die Erkennung von Objekten in natürlichen Bildern. Obwohl es für einen Menschen sehr einfach zu bestimmen ist, ob ein Bild ein Objekt einer bestimmten Kategorie enthält, muss ein Computer aufwändige Modelle und Verfahren anwenden, um dies zu erreichen.

In dieser Arbeit werden verschiedene Methoden entwickelt und miteinander verglichen, um Objekte in Bildern zu erkennen. Im Gegensatz zu vielen vergleichbaren Methoden aus der Literatur verzichten wir dabei weitgehend auf Heuristiken und versuchen die Annah-

men, die in dem Modell gemacht werden, soweit wie möglich zu reduzieren. Das daraus resultierende Modell benötigt im Vergleich zu anderen Modellen nur sehr wenige Parameter, ist effizient trainier- und anwendbar und erzielt hervorragende Ergebnisse.

Bei der Entwicklung unserer Methoden zur Klassifikation von Bildern beginnen wir üblicherweise mit Bayes' Entscheidungsregel und leiten daraus ein statistisch motiviertes System ab. Dabei versuchen wir, lokale und daher möglicherweise ungünstige Entscheidungen zu vermeiden, um stattdessen alle verfügbare Information in der Klassifikationsentscheidung zu verwenden. Eine typische lokale Entscheidung, die in vielen Arbeiten gemacht wird, ist die Segmentierung, bei der zuerst versucht wird festzustellen, wo im Bild sich ein Objekt befindet, bevor man versucht dieses zu identifizieren. Wenn in diesem Schritt ein Fehler gemacht wird, ist es oftmals nicht mehr möglich, das Objekt noch korrekt zu erkennen. Solche Fehler versuchen wir zu vermeiden, indem wir die Klassifikationsentscheidung unter Berücksichtigung des gesamten Bildes treffen.

Um die Qualität unserer Methoden zu ermitteln, verwenden wir grundsätzlich Standardaufgaben aus der Literatur mit einem wohldefinierten Evaluierungsmaß, wie z.B. der Klassifikationsfehlerrate.

Im folgenden Abschnitt stellen wir drei Ansätze zur Objekterkennung in natürlichen Bildern vor, vergleichen diese miteinander und diskutieren die Vor- und Nachteile. Dann geben wir noch einen kurzen Ausblick auf zwei andere Anwendungen, für die wir die vorgestellten Methoden anpassen und erweitern konnten.

2 Diskriminative Modellierung und Objekterkennung

Der erste Schritt zur Erstellung eines Objekterkennungssystems ist die Wahl eines geeigneten Merkmalsextraktionsverfahrens. Hierbei unterscheidet man im Allgemeinen zwischen zwei verschiedenen Möglichkeiten: Bei der Extraktion globaler Merkmale wird ein Merkmalsvektor aus dem ganzen Bild extrahiert, der versucht den Inhalt des gesamten Bildes zu beschreiben. Bei der Extraktion lokaler Merkmale werden viele Deskriptoren extrahiert, die aber nicht das gesamte Bild, sondern nur lokale Eigenschaften erfassen. Der Vorteil eines globalen Deskriptors ist, dass er oftmals einfach und schnell verarbeitet werden kann. Bei lokalen Deskriptoren ist die weitere Verarbeitung hingegen oftmals aufwändig. Lokale Merkmale haben jedoch den Vorteil, dass es sowohl möglich ist, Objekte zu erkennen, wenn sie teilweise überdeckt sind, als auch Variabilitäten in der geometrischen Anordnung von Objektteilen zu erfassen.

Da in der vorliegenden Arbeit der Effekt verschiedener Modellierungen untersucht wird, werden im Folgenden einfache, jedoch als gut funktionierend bekannte, lokale Merkmale benutzt. Dazu werden im Bild zuerst so genannte *Interest Points* detektiert. Dies sind Positionen im Bild, an denen z.B. eine hohe lokale Varianz ist. Dann wird jeder dieser Punkte durch die Pixelwerte einer rechteckigen Nachbarschaft beschrieben. Diese Pixelwerte werden direkt als Merkmale in den beschriebenen Verfahren verwendet. Um die Dimensionalität der Daten und somit die zu verarbeitende Datenmenge zu reduzieren, werden diese Merkmalsvektoren anhand der Eigenwertanalyse (PCA) auf 40-dimensionale Vektoren reduziert.

Bei der Erstellung von Mustererkennungs- oder maschinellen Lernsystemen wird oftmals zwischen den beiden Ansätzen *generativer* und *diskriminativer* Modellierung unterschieden. Beide Ansätze entstammen der Bayes'schen Entscheidungsregel und haben Vor- und Nachteile. Bayes' Entscheidungsregel ist gegeben als

$$r(x) \quad = \quad \arg\max_c\{p(x,c)\} = \arg\max_c\{p(c|x)\}, \tag{1}$$

bei der eine Beobachtung x derjenigen Klasse c zugewiesen wird, für die die gemeinsame Wahrscheinlichkeit $p(x,c)$ maximal ist. Durch Anwendung von Bayes' Theorem erhält man entweder $p(x,c) = p(c)p(x|c)$ oder $p(x,c) = p(x)p(c|x)$, die den beiden verschiedenen Ansätzen entsprechen.

Im generativen Ansatz werden die klassenbedingten Wahrscheinlichkeiten $p(x|c)$ sowie die Priorwahrscheinlichkeiten $p(c)$ modelliert und die Posteriorwahrscheinlichkeiten für jede Klasse gemäß

$$p(c|x) \quad = \quad \frac{p(c)p(x|c)}{\sum_{c'} p(c')p(x|c')} \tag{2}$$

bestimmt. Im diskriminativen Ansatz wird $p(c|x)$ direkt modelliert.

Welcher dieser Ansätze zu bevorzugen ist, lässt sich nicht eindeutig beantworten, da beide korrekt sind. Jedoch kann man in beiden Ansätzen Vor- und Nachteile identifizieren:

Der generative Ansatz versucht, die Trainingsdaten optimal zu repräsentieren und es ist möglich mit nicht oder nur teilweise klassifizierten Daten zu trainieren. Ein Vorteil des diskriminativen Ansatzes ist, dass er direkt die Posteriorwahrscheinlichkeit $p(c|x)$ und damit die Klassifikationsentscheidung modelliert. Jedoch benötigt man zum Training vollständig klassifizierte Trainingsdaten. Außerdem besteht die Gefahr der Überanpassung, die in vielen diskriminativen Modellen mit hohem Aufwand zu verhindern versucht wird.

Verschiedene Arbeiten versuchen die Vorteile dieser beiden Ansätze direkt zu verbinden. Dazu wurde z.B. ein Modell erstellt, welches es ermöglicht, nahtlos zwischen den verschiedenen Ansätzen zu wählen [LBM06]. Außerdem wurden diskriminative Modelle um generative Eigenschaften erweitert [GRB07].

Im Folgenden stellen wir drei verschiedene Modelle vor, die die beiden Ansätze auf unterschiedliche Arten kombinieren: Der erste Ansatz benutzt ein generatives Modell, um aus einer Menge von lokalen Merkmalen einen Merkmalsvektor zu erstellen, der dann mit einem diskriminativen Modell klassifiziert wird. Im zweiten Ansatz wird der generative Ansatz direkt verwendet, jedoch mit einem diskriminativen Trainingsverfahren verfeinert. Im dritten Ansatz wird ein diskriminatives Modell erstellt, zu dem ein äquivalentes generatives Modell existiert, und wir zeigen, wie man vom einen ins andere umformen kann.

2.1 Histogramme von lokalen Merkmalen

Das erste hier vorgestellte Verfahren basiert auf den Methoden, die wir in [DKN05a, DKN05b] bzw. [Des08, Abschnitt 4.4] vorgestellt haben und ist in der Literatur unter dem Namen *Bag-of-Visual-Words*-Ansatz bekannt. Dieses Verfahren besteht aus zwei Phasen:

In der ersten Phase werden aus einem Bild X eine Menge lokaler Merkmale $\{x_1, \ldots, x_L\}$ extrahiert. Dann wird anhand eines Clustering-Verfahrens für alle extrahierten lokalen Merkmale einer gegebenen Bildmenge $\{X_1, \ldots, X_n\}$ ein *visuelles Vokabular* erstellt. Die lokalen Merkmale der einzelnen Bilder werden dann den jeweiligen Clustern, die auch als *visuelle Wörter* bezeichnet werden, zugeordnet. Dabei wird die Erscheinungsinformation der einzelnen Merkmale verworfen. Es wird also nur noch gespeichert, welches visuelle Wort wie häufig vorkommt.

In unserem Ansatz verwenden wir für die Erstellung des visuellen Vokabulars den *Expectation-Maximization*-Algorithmus für Gauß'sche Mischverteilungen. Wir beginnen mit einer einzigen Normalverteilung, die dann iterativ aufgeteilt wird, bis die gewünschte Anzahl Cluster erreicht ist. Dann wird für jedes Bild das Histogramm über die Clusterzugehörigkeiten der lokalen Merkmale erstellt. Diese Histogramme geben an, welches visuelle Wort wie häufig in den entsprechenden Bildern vorkommt. Der dabei entstehende Merkmalsvektor ist zwar ein globaler Bilddeskriptor, jedoch enthält er aus Information über lokale, aussagekräftige Merkmale.

In der zweiten Phase werden die erzeugten Histogramme klassifiziert. Auch wenn es möglich ist, diese mit einem generativen Modell zu klassifizieren, hat sich gezeigt, dass es oft zu besseren Ergebnissen führt, wenn hierzu ein diskriminatives Verfahren verwendet wird. In der Literatur werden dazu regelmäßig Support-Vektor-Maschinen verwendet. Wir verwenden ein log-lineares Modell der Posteriorwahrscheinlichkeiten:

$$p(c|x) \quad = \quad \frac{\exp(\alpha_c + \sum_i \lambda_{ci} h_i(X))}{\sum_{c'} \exp(\alpha_{c'} + \sum_i \lambda_{c'i} h_i(X))}, \tag{3}$$

wobei die α_c und λ_{ci} Parameter des Modells sind und $h_i(X)$ der i-te Eintrag des Histogramms für das Bild X ist.

Nachdem das Modell trainiert wurde, lässt sich anhand der trainierten Parameter λ_{ci} ablesen, welche visuellen Wörter besonders wichtig für die Klassifikation sind. Das Ergebnis einer solchen Analyse wird in Abbildung 1 gezeigt. Die Abbildung zeigt die wichtigsten visuellen Wörter für die Klassifikation von Gesichtern, Flugzeugen und Motorrädern. Es zeigt sich, dass das System im Stande ist, semantisch bedeutungsvolle Information zu lernen. So kann man erkennen, dass z.B. das wichtigste visuelle Wort zur Erkennung von Gesichtern ein Auge ist, dass Flugzeuge viele horizontale Strukturen und Motorräder viele diagonale Strukturen aufweisen.

Abbildung 1: **Visuelle Wörter.** Die drei wichtigsten visuellen Wörter der drei Klassen Gesichter, Flugzeuge und Motorräder (v.l.n.r.).

2.2 Gauß'sche Mischverteilungen und diskriminatives Training

Gauß'sche Mischverteilungen sind ein Standardansatz zur Mustererkennung und wurden im vorherigen Verfahren bereits verwendet, um das visuelle Vokabular zu erstellen. Jedoch wurde im vorherigen Verfahren in der ersten Phase bei der Erstellung der Histogramme eine harte Entscheidung getroffen und viel potentiell relevante Information verworfen. Hier wird dies vermieden und die gesamte Bildinformation wird direkt in der Entscheidungsregel benutzt.

Der hier vorgestellte Ansatz orientiert sich weitgehend an den Methoden, die wir in [Des08, Abschnitt 4.7] bzw. [HDN06] vorgestellt haben. Dazu starten wir wieder mit Bayes' Entscheidungsregel und binden die Merkmalsvektoren $\{x_1, \ldots, x_L\}$ direkt ein:

$$r(\{x_1, \ldots, x_L\}) = \arg\max_c \{p(c|x_1, \ldots, x_L)\} = \arg\max_c \{p(c)p(\{x_1, \ldots, x_L\}|c)\} \quad (4)$$

$$= \arg\max_c \{p(c) \prod_l p(x_l|c)\}. \quad (5)$$

Die klassenbedingten Wahrscheinlichkeiten $p(x_l|c)$ werden durch Gauß'sche Mischverteilungen modelliert:

$$p(x_l|c) \quad = \quad \sum_i p(i|c)p(x_l|i,c) = \sum_i p(i|c)\mathcal{N}(x_l|\mu_{ci}, \Sigma_{ci}). \quad (6)$$

Durch eine Erweiterung des Modells ist es außerdem möglich, Positionsinformation in das Modell zu integrieren und damit das Modell weiter zu verbessern.

Normalerweise werden Gauß'sche Mischverteilungen als generatives Modell anhand des *Maximum-Likelihood Kriteriums* trainiert, welches die Repräsentation der Trainingsdaten optimiert. Da diskriminative Modelle oft bessere Klassifikationsergebnisse erzielen, verfeinern wir unser Mischverteilungsmodell durch diskriminatives Training. Dazu wird das austrainierte generative Modell als Startpunkt gewählt und dann anhand des *Maximum Mutual Information Kriteriums* weitertrainiert.

Auch in diesem Fall lassen sich die Mittelpunkte der Dichten als visuelle Wörter visualisieren. Diese sind mit den in Abbildung 1 Gezeigten vergleichbar.

2.3 Log-Lineare Mischverteilungen

Ein Nachteil des Ansatzes mit Gauß'schen Mischverteilungen ist, dass das diskriminative Training als Verfeinerung des ursprünglichen, generativen Modells zwar zu einer Kombination der beiden Ansätze führt, dass er aber schwer zu interpretieren ist, da es keine geschlossene Formulierung des Trainingskriteriums gibt.

Ausgehend von der Posteriorwahrscheinlichkeit des Gauß'schen Mischverteilungsmodells, formulieren wir dieses in ein log-lineares Mischverteilungsmodell um. Dies ist ein geschlossenes diskriminatives Modell, lässt sich direkt als solches trainieren und ist effizienter und numerisch stabiler auszuwerten als das ursprüngliche Gauß-Modell.

Dazu folgen wir [Des08, Abschnitt 4.8] bzw. [Wey08]. Ausgehend von der Posteriorwahrscheinlichkeit zu Gleichung (4), formen wir das Modell um und gelangen zu folgender Form:

$$p(c|x_1,\ldots,x_L) \quad = \quad \frac{\sum_i \exp(\sum_l \alpha_{ci} + \lambda_{ci}^T x_l)}{\sum_{\tilde{c}} \sum_{\tilde{i}} \exp(\sum_l \alpha_{\tilde{c}\tilde{i}} + \lambda_{\tilde{c}\tilde{i}}^T x_l)}. \tag{7}$$

Diese beiden Modelle sind äquivalent und die Parameter des einen können aus denen des jeweilig anderen bestimmt werden.

Weiterhin haben wir gezeigt, dass das Training eines solchen log-linearen Mischverteilungsmodells durch Anwendung der Maximumapproximation semi-konvex wird und somit garantiert ist, dass das Training zu einem (lokalen) Optimum konvergiert. Dazu werden mit alternierender Optimierung abwechselnd die Modellparameter $\{\alpha_{ci}, \lambda_{ci}\}$ und die Alinierung von Trainingsbeobachtungen zu Modellkomponenten aktualisiert. Man kann zeigen, dass kein Schritt in diesem Verfahren jemals das nach oben beschränkte Trainingskriterium verschlechtert.

Ein weiterer Vorteil dieses Modells ist, dass weitere Informationsquellen, wie z.B. Positionsinformation, durch eine einfache Erweiterung des Modells integriert werden können.

Die Visualisierung der Modellparameter ist in diesem Fall nicht so einfach möglich wie in den beiden vorherigen Fällen. Jedoch ist es möglich, das Modell in eine Gauß-Mischverteilung zu transformieren, um dann die Mittelwerte darzustellen.

2.4 Gegenüberstellung und Vergleich der drei Methoden

Die drei Methoden haben viele Gemeinsamkeiten. Alle Methoden benutzen lokale Bilddeskriptoren und basieren auf Bayes' Entscheidungsregel.

Die erste, histogrammbasierte Methode hat den Vorteil, dass sie einfach zu implementieren ist und trotzdem oftmals sehr gute Ergebnisse erzielt. Ein typisches Problem bei diesem Verfahren ist, dass das Vokabular aufwändig zu erstellen ist und die Qualität der Ergebnisse stark von der Größe des Vokabulars abhängt. Ein weiteres Problem dieses Ansatzes ist, dass es sehr schwierig ist, Positionsinformation in das Modell zu integrieren. Die meisten Arbeiten, die diesem Ansatz folgen, benutzen entweder gar keine Positionsinformation oder integrieren diese in einem Nachverarbeitungsschritt.

Bei dem zweiten Modell mit Gauß'schen Mischverteilungen ist zwar die Anzahl der Dichten in dem Modell auch ein Parameter, aber dessen Einfluss auf die Qualität der Ergebnisse ist deutlich geringer. Eine Erklärung dafür ist, dass die Anzahl der Dichten einen deutlich geringeren Einfluss hat, da die Erscheinungsinformation der lokalen Bildmerkmale direkt in die Entscheidung einfließt. Ein weiterer Vorteil dieses Ansatzes ist, dass es einfach möglich ist, die gesamte Bildinformation komplett in die Entscheidung einfließen zu lassen, und dass man zusätzlich weitere Informationsquellen, wie z.B. Positionsinformation integrieren kann. Wenn in der diskriminativen Verfeinerung dieses Modells nicht alle Parameter, sondern nur die Gewichte $p(i|c)$ diskriminativ trainiert werden, entsprechen diese direkt den λ-Parametern des log-linearen Modells der zweiten Phase des histogrammba-

Abbildung 2: **Beispielbilder aus der Caltech Datensammlung.** Bei dieser Aufgabe gilt es zu entscheiden, ob ein Bild eines der Objekte (Flugzeug, Gesicht, Motorrad) zeigt oder nicht (rechtes Bild).

sierten Modells.

Im letzten Modell werden alle Vorteile des Gauß'schen Modells beibehalten. Außerdem ist es ein vollständig diskriminatives Modell, welches auch als solches trainiert werden kann. Dies reduziert den Einfluss der Anzahl der Dichten weiterhin, und somit können schon mit einer sehr geringen Anzahl von Dichten hervorragende Modelle trainiert werden.

In [Des08] werden vier weitere Modelle zur Objekterkennung vorgestellt.

3 Experimentelle Auswertung

Zur experimentellen Analyse der Verfahren haben wir verschiedene Bilddatensätze verwendet. Der kleinste und am einfachsten zu verarbeitende Korpus wurde von Caltech vorgestellt und besteht aus mehreren Aufgaben, in denen zu entscheiden ist, ob eine bestimmte Objektkategorie in einem Bild zu sehen ist [FPZ03]. Hier betrachten wir nur die Teilaufgaben "Flugzeuge", "Gesichter" und "Motorräder". Für diese Aufgaben stehen jeweils ca. 500-800 Trainings- und Testbilder zur Verfügung, wobei die Hälfte der Bilder das Objekt zeigt und die andere Hälfte nicht. Ein Beispielbild für jede Objektkategorie ist in Abbildung 2 zu sehen. Generell ist diese Aufgabe relativ einfach und es ist möglich, sehr gute Erkennungsraten zu erzielen. In [Des08, Abschnitt 4.13] werden zusätzlich Experimente auf medizinischen Daten und auf schwierigeren Datensammlungen vorgestellt.

Ein Überblick über die Ergebnisse, die mit den vorgestellten Methoden auf dieser Datensammlung erzielt worden sind, wird in Tabelle 1 gegeben. Weiterhin sind in dieser Tabelle zwei der besten Vergleichsergebnisse aus der Literatur aufgeführt. Alle Ergebnisse sind als Fehlerraten in Prozent angegeben.

An diesen Ergebnissen sieht man, dass die Gauß'schen Mischverteilungen und die Log-Linearen Mischverteilungen in etwa gleich gut abschneiden während die histogrammbasierte Methode etwas schlechter ist. Die Positionsinformation und das diskriminative Training führen zu einer deutlichen Verbesserung der Ergebnisse im Gauß-Modell. Für die Histogramme ist ein Wörterbuch mit 2048 visuellen Wörtern erstellt worden. Die Gauß-Modelle wurden mit jeweils 256 Dichten pro Klasse trainiert. Eine weitere Vergrößerung dieser Modelle bringt kaum noch Verbesserungen. Bei den log-linearen Modellen reichen in diesem Fall sogar vier Komponenten pro Klasse und auch hier führt eine weitere Vergrößerung kaum zu Verbesserungen. Insgesamt lässt sich feststellen, dass die Anzahl der Modellkomponenten in dem log-linearen Modell nur einen sehr geringen Einfluss hat und dass auch das Gauß-Modell in Bezug auf diesen Parameter robuster ist als die histogramm-

Tabelle 1: Ergebnisse für die drei Caltech-Aufgaben “Flugzeuge”, “Gesichter” und “Motorräder”.

Methode	Flugzeuge	Gesichter	Motorräder
Histogramme lokaler Merkmale	2.6	5.8	1.5
Gauß-Mischverteilungen	1.5	3.2	3.5
+ Positionsinformation	0.5	0.0	0.8
+ diskriminatives Training	0.5	0.0	0.3
Log-Lineare Mischverteilungen	2.3	2.3	2.3
+ Positionsinformation	1.5	2.0	1.0
+ verbesserte Merkmale	1.3	1.8	1.0
Konstellationsmodell [FPZ03]	9.8	3.6	7.5
Biologisch inspiriertes Modell [SWP05]	3.3	1.8	2.0

basierte Methode.

Im Vergleich zu den beiden Ergebnissen aus der Literatur sieht man, dass die vorgestellten Modelle hervorragend abschneiden. Das Konstellationsmodell ist ein generatives Modell, welches die zu erkennenden Objekte durch drei bis sieben Komponenten darstellt [FPZ03]. Das biologisch inspirierte Modell versucht die menschliche Wahrnehmung zu imitieren und verwendet dazu Strukturen, die dem menschlichen visuellen Cortex nachempfunden wurden [SWP05].

4 Andere Anwendungen

Die oben beschriebenen Verfahren und verwandte Methoden wurden auch in weiteren Anwendungen eingesetzt. Diese werden hier kurz vorgestellt.

4.1 Bildsuche

Die Suche nach Bildern in einer großen Bilddatenmenge ist eng mit der oben vorgestellten Objekterkennung verwandt und teilweise können ähnliche Methoden angewandt werden.

Die wichtigsten Komponenten eines Bildsuchsystems sind:

Merkmalsextraktion. In der Merkmalsextraktion hat sich gezeigt, dass die Patch–Histogramme (vgl. Abschnitt 2.1) hervorragend geeignet sind, um ähnliche Bilder zu finden [DKN08].

Ähnlichkeitsmaße. Die aus den Bildern extrahierten Merkmalsvektoren müssen mit geeigneten Ähnlichkeitsmaßen verglichen werden.

Merkmalskombination. Um geeignete Kombinationen von Merkmalen zu finden, sind diskriminative Methoden hervorragend geeignet [Des08, Abschnitt 3.9]. Dabei kann man sowohl visuelle und textuelle Merkmale als auch verschiedene Arten visueller Merkmale miteinander kombinieren.

Benutzerinteraktion. Analog zum Lernen von Merkmalskombinationen kann auch Benutzerinteraktion benutzt werden, um die Parameter eines Systems zu ermitteln. Auch hierbei sind diskriminative Techniken eine geeignete Wahl [Des08, Abschnitt 3.11].

Quantitative Evaluierung. Die quantitative Evaluierung ist unverzichtbar, um verschiedene Systeme zu vergleichen. Im Rahmen der Arbeit wurden verschiedene öffentliche Evaluierungsmaßnahmen in der ImageCLEF-Kampagne organisiert, in denen verschiedene Systeme verglichen wurden.

Alle in diesem Bereich entwickelten Methoden und Ansätze sind Teil des als Open-Source verfügbaren Image Retrieval Systems FIRE.

4.2 Handschrifterkennung

Die Erkennung handgeschriebener Zeichen ist eine der am besten untersuchten Mustererkennungsaufgaben. In [Des08, Kapitel 5] bzw. [Gas08] werden einige der Ideen der log-linearen Modellierung aus der Objekterkennung abgewandelt, um Verformungen von handgeschriebenen Ziffern zu modellieren. Anstelle der Variable, die eine Beobachtung einer Modellkomponente zuweist, wird in diesem Fall die Verformung eines Bildes bzw. die Anpassung des Modells an eine zu klassifizierende Beobachtung als latente Variable modelliert.

Im Gegensatz zu vergleichbaren Ansätzen ist es dabei möglich, außer den üblichen Modellparametern auch die Verformungsparameter zu trainieren. Wie in der Objekterkennung können auch hier sehr effiziente kleine Modelle, die ausgesprochen gute Ergebnisse auf Standarddatensätzen erzielen, trainiert werden.

5 Zusammenfassung

Ausgehend von einem gut funktionierenden Standardmodell zur Objekterkennung haben wir gezeigt, wie durch Reduktion der verwendeten Annahmen und Heuristiken ein sauberes statistisches Modell hergeleitet werden kann. Das resultierende Modell ist effizient trainier- und anwendbar. Es erzielt sehr gute Ergebnisse und benötigt nur sehr wenige Parameter, was darauf schließen lässt, dass es gut auf andere Aufgaben generalisiert.

Ähnliche Methoden wurden auch für die Bildsuche und die Handschrifterkennung eingesetzt. In diesen Bereichen konnten ebenfalls hervorragende Ergebnisse erzielt werden.

Danksagung. Diese Doktorarbeit ist durch die fachliche Unterstützung vieler Leute möglich geworden. An dieser Stelle danke ich denjenigen, die hier besonders hervorzuheben sind: Hermann Ney für die Betreuung und Ermöglichung der Arbeit, Bernt Schiele für viele hilfreiche Diskussionen, Daniel Keysers und Philippe Dreuw für verschiedene Beiträge, sowie Andre Hegerath, Tobias Gass, Tobias Weyand, Henning Müller, und Roberto Paredes.

Literatur

[Des08] Thomas Deselaers. *Image Retrieval, Object Recognition, and Discriminative Models*. Dissertation, RWTH Aachen, Aachen, Germany, Dezember 2008.

[DKN05a] Thomas Deselaers, Daniel Keysers und Hermann Ney. *Discriminative Training for Object Recognition using Image Patches.* In *IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2005.

[DKN05b] Thomas Deselaers, Daniel Keysers und Hermann Ney. *Improving a Discriminative Approach to Object Recognition using Image Patches.* In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, *LNCS* 3663, August 2005.

[DKN08] Thomas Deselaers, Daniel Keysers und Hermann Ney. *Features for Image Retrieval: An Experimental Comparison. Information Retrieval*, 11(2):77–107, 2008.

[FPZ03] R. Fergus, P. Perona und A. Zissermann. *Object Class Recognition by Unsupervised Scale-Invariant Learning.* In *IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2003.

[Gas08] Tobias Gass. *Deformations and Discriminative Models for Image Recognition.* Diplomarbeit, RWTH Aachen, Juli 2008.

[GRB07] Helmut Grabner, Peter M. Roth und Horst Bischof. *Eigenboosting: Combining Discriminative and Generative Information.* In *IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2007.

[HDN06] Andre Hegerath, Thomas Deselaers und Hermann Ney. *Patch-based Object Recognition Using Discriminatively Trained Gaussian Mixtures.* In *British Machine Vision Conference*, September 2006.

[LBM06] Julia A Lasserre, Christopher M Bishop und Thomas P Minka. *Principled Hybrids of Generative and Discriminative Models.* In *IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2006.

[SWP05] Thomas Serre, Lior Wolf und Tomaso Poggio. *Object Recognition with Features Inspired by Visual Cortex.* In *IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2005.

[Wey08] Tobias Weyand. *Log-Linear Mixture Models for Patch-Based Object Recognition.* Diplomarbeit, RWTH Aachen, September 2008.

Thomas Deselaers hat an der RWTH Aachen Informatik studiert und war von März 2004 bis Dezember 2008 wissenschaftlicher Mitarbeiter am Lehrstuhl für Sprachverarbeitung und Mustererkennung. Im Dezember 2008 schloss er dort seine Dissertation ab. Seit Januar 2009 arbeitet er als Researcher am Institut für Bildverarbeitung der ETH Zürich. Im Jahr 2006 verbrachte er einen Forschungsaufenthalt bei Microsoft Research in Cambridge (GB). In den Jahren 2002 und 2008 war er zu Forschungsaufenthalten am Instituto Tecnológico de Informática an der Universidad Politécnica de Valencia (ES). Er ist Vice-Chair des Technischen Kommittees 5 der IAPR und Organisator einiger internationaler Evaluationen im Bereich Bildsuche im Rahmen des ImageCLEF-Programms. Seine Forschungsinteressen liegen in den Bereichen Computer Vision und maschinelles Lernen.