Erweiterung eines holistischen statistischen Bilderkenners zur Verwendung von mehreren Merkmalen

David Rybach, Daniel Keysers, Hermann Ney
david.rybach@rwth-aachen.de
{keysers,ney}@cs.rwth-aachen.de

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany

Art der Arbeit: Studienarbeit
Betreuer der Arbeit: Dipl.-Inform. D. Keysers, Prof. Dr.-Ing H. Ney

Zusammenfassung
Statistische Bilderkenner, die mit einem holistischen Ansatz arbeiten, verwenden Wahrscheinlichkeitsverteilungen als Modelle für Objekte und deren Hintergründe. Die Modelle werden aus ortsabhängigen Merkmalen der Trainingsbilder berechnet, z.B. aus den Grauwerten. In dieser Arbeit wird ein Verfahren vorgestellt, dass mehrere Merkmale in die statistischen Modelle integriert.

1 Einleitung

Der erste Schritt in den meisten Verfahren zur Bilderkennung ist die Berechnung von Merkmalen aus den Bilddaten. Dies können ganz einfache Merkmale sein, zum Beispiel die Grau- oder Farbwerte des Bildes, oder Daten, die aus Berechnungen auf dem Bildmaterial hervorgehen. Die Merkmalsdaten werden analysiert und vom Klassifikator benutzt.

Bilderkenner, die mit einem holistisch statistischen Verfahren arbeiten, können komplexe Szenen analysieren und die abgebildeten Objekte klassifizieren. Sie „erklären“ das gesamte Bild mit statistischen Modellen. Es ist also notwendig, Modelle für Objekte und Hintergründe zu verwenden. Die Bestimmung der Position und der Größe des Objekts und dessen Klassifizierung geschehen in einem Schritt. Viele andere Verfahren müssen vor der Klassifikation eine Segmentierung des Bildes durchführen, die alleine schon fehleranfällig ist. Auch für das automatische Objekttraining werden unsegmentierte Daten benutzt, für die nur die Klasse des abgebildeten Objekts bekannt ist. Daher müssen die Methoden, die zur Klassifikation genutzt werden, auch für das Training angewendet werden. Die Verwendung von unsegmentierten Daten ist wünschenswert, um die manuelle Arbeit an den Daten zu minimieren.

In dieser Arbeit wird die Erweiterung eines holistischen Bilderkenners vorgestellt, die es ermöglicht, mehrere Merkmale eines Bildes zu verarbeiten. Um die Merkmale eines Bildes zu kombinieren, wird ein neues „Bild“ berechnet, das aus mehreren Schichten besteht. Die einzelnen Schichten repräsentieren Merkmale oder Teile eines Merkmals.

2 Verfahren

Das Verfahren basiert auf einem statistischen Ansatz zur Objekterkennung mit einem ganzheitlichen Modell. Dabei werden statistische Modelle für das Objekt und den Hintergrund verwendet, um alle Pixel eines Bildes zu erklären. Das Objekt wird als quadratischer Ausschnitt des Bildes angenommen. Als Modell für Objekte wird eine Gauß’sche Mischvertei-

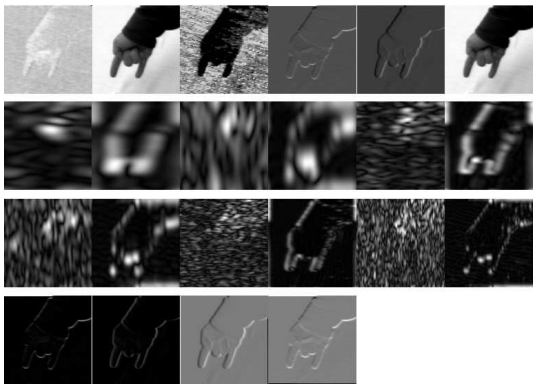

Abbildung 1: Features aus einem Bild: Farbe mit Komponenten S, V, H (3), Ableitung (2), Grauwert (1), Gabor-Transformation mit 3 Frequenzen und 2 Orientierungen jeweils für Farbe und Helligkeit getrennt (12), absolutwertige Ableitung (2), Sobel-Filter (2)

Abbildung 2: Matching eines Referenzmodells und einer Beobachtung: Die Referenz (links) wird skaliert (mitte) und die beste Position innerhalb der Beobachtung wird bestimmt (rechts).

lung verwendet. Die Menge von Pixeln, die nicht zum Objekt gehören, werden mit einer univariaten Gaußverteilung, dem Hintergrundmodell, beschrieben.

Neben den Grauwerten eines Bildes können auch andere Merkmale (*Features*) verwendet werden. Alle Features eines Bild werden dann zu einem ein Bild aus mehreren Schichten (*Layern*) zusammengefasst. In dieser Arbeit wurden neben einfachen Grau- und Farbwert-Features auch die Ableitung, ein Sobel-Filter und die Gabor-Transformation verwendet. Beispiele für alle Features sind in Abbildung 1 dargestellt.

Im Matching wird für eine Sammlung von Features eines Bildes, im folgenden Beobachtung genannt, die wahrscheinlichste Position und Skalierung des gegebenen Referenzmodells gesucht. Dazu werden aus dem Referenzmodell unterschiedlich skalierte Templates erzeugt und auf mehreren Positionen deren Distanz zur Beobachtung berechnet. Abbildung 2 zeigt, wie ein Objektmodell auf eine Beobachtung projiziert wird.

Im Training werden aus einer Menge von Trainingsbildern mit bekannten Klassen die Gauß'schen Mischverteilungen für die Objektmodelle und Gaußverteilungen für die Hintergrundmodelle berechnet. Das Training wird hier zunächst nicht diskriminativ durchgeführt, sondern für jede Klasse separat. Ausgehend von einem initialen Modell, das aus einer gegebenen Sammlung von Features und einer gegebenen Varianz berechnet werden kann, werden iterativ immer bessere Verteilungen berechnet, die die Trainingsdaten erklären. In jeder Iteration wird, nachdem für alle Trainingsdaten das Matching durchgeführt wurde, aus den besten berechneten Hypothesen ein neues Objektmodell berechnet. Die berechneten Verteilungen werden als Modell der jeweiligen Klasse für die Objektklassifikation verwendet. Abbildung 3 zeigt einen Teil der Trainingsbilder für eine Klasse und das daraus berechnete Objektmodell.

Zur Klassifikation eines Bildes wird die aus dem Bild berechnete Sammlung von Features verwendet. Diese Featuresammlung wird mit den Modellen gematcht, die im Training für die einzelnen Klassen erstellt wurden. Der Klassifikator ordnet dann dem zu klassifi-

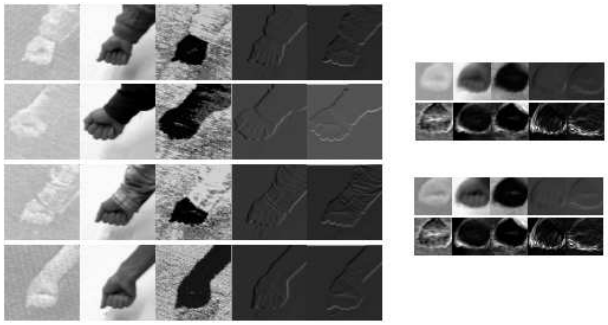

Abbildung 3: 4 von 18 Trainingsdaten (links) und das daraus berechnete Modell (rechts) mit 2 Dichten (jeweils oben der Mittelwert und darunter die Varianz).

Tabelle 1: Fehlerraten der Experimente auf der Bochum-Gestures Datenbank in Prozent. Angegeben sind die mittleren Fehlerraten der drei Testdatensätze.

Features	Fehlerrate
Grauwerte	8,6
Grauwerte, Sobel-Filter	10,8
Grauwerte, Gabor-Filter (3 Frequenzen, 2 Orientierungen)	6,4

zierenden Bild diejenige Klasse zu, deren Modell die geringste Distanz zum Bild hat. Die minimale Distanz eines Modells zum Bild impliziert die maximale Wahrscheinlichkeit, dass Objekt und Hintergrund vom Modell erzeugt werden [KMDN03].

3 Ergebnisse

Die Tests wurden auf der Bochum-Gestures- und der CALTECH-Faces-Datenbank durchgeführt. Die Bochum-Gestures Bilddatenbank besteht aus Fotos von 12 Handgesten. Insgesamt sind 1036 Farbbilder mit 128×128 Pixeln enthalten. Jede Geste liegt von unterschiedlichen Personen jeweils vor hellem, dunklem und komplexem Hintergrund vor. Die Bilddatenbank wird in [TvdM01] vorgestellt. Mit dem dort beschriebenen Elastic Graph Matching erzielen die Autoren Fehlerraten von 7,1% für Bilder mit einfachem Hintergrund und 14,2% für Bilder mit komplexem Hintergrund. Diese Ergebnisse sind nicht mit den in dieser Arbeit durchgeführten Experimenten zu vergleichen, da die Bilder dort manuell für das Training bearbeitet wurden.

In den Experimenten wurden die 345 Bilder mit hellem Hintergrund verwendet, die in 3 TestdatensAätze aufgeteilt wurden. Für die Startmodelle wurde aus allen Trainingsbildern einer Klasse der Mittelwert berechnet und daraus Feature-Sammlungen erstellt. Das Hintergrundmodell wurde als Gleichverteilung mit geringem Gewicht in die Distanzberechnung einbezogen. In Tabelle 1 sind ausgewählte Ergebnisse der Experimente dargestellt.

Die CALTECH-Faces-Datenbank des California Institute of Technology besteht aus Grauwert-Bildern von Gesichtern vor komplexen und vor einfachen Hintergründen [BLP96]. Für die Experimente wurden 534 dieser Fotos verwendet. Da die CALTECH-Faces-Datenbank keine Klasseninformationen beinhaltet, wurden nur Modelle für eine Klasse „Gesicht“ trainiert. Im Test wurden Bilder mit und ohne Gesicht klassifiziert. Anschließend wurde ein Grenzwert für die ermittelten Distanzen berechnet, mit dem entschieden werden kann, ob ein Bild ein Gesicht enthält oder nicht. Die erzielten Fehlerraten zeigt Tabelle 2.

Aus den Fehlerraten in Tabelle 1 ist ersichtlich, dass die Informationen des Gabor-Features die Erkennung verbessern. Die Integration der anderen Features kann die Fehlerrate bei Experimenten auf der Bochum-Gestures-Datenbank allerdings nicht senken. Die Experi-

Tabelle 2: Fehlerraten auf der CALTECH-Faces-Datenbank in Prozent.

Features	Fehlerrate
Grauwerte	47,7
Grauwerte, Ableitung	32,3
Grauwerte, Gabor-Filter (2 Frequenzen, 1 Orientierung)	42,2

mente auf den Gesichts-Bildern haben ein anderes Ergebnis. Dort senkt die Verwendung der Ableitung die Fehlerrate. Die Hinzunahme des Gabor-Features lässt die Fehlerrate wieder ansteigen, bleibt allerdings unter der Fehlerrate des Experiments, in dem nur Grauwerte benutzt werden.

5 Fazit und Ausblick

Das vorgestellte System ermöglicht die Integration mehrerer Features in einen holistischen Bilderkenner. Die erwartete Verbesserung der Erkennungsleistung bei steigender Anzahl von Features bestätigt sich, tritt aber nicht in allen Experimenten auf. Weitere Experimente sollen noch andere Kombinationen von Features untersuchen.

Offenbar ist der Einfluss der Features auf die Fehlerrate abhängig von der Art der untersuchten Bilder bzw. der Bilddatenbank. In der Bochum-Gestures Datenbank sind zum einen einige Gesten untereinander sehr ähnlich und somit schwer zu differenzieren. Zum anderen ist die Trainingsdatenmenge sehr klein. Trotzdem erzielt das Verfahren niedrige Fehlerraten. Man könnte die Trainingsdaten vergrößern, indem Variationen durch geringe Rotation und Skalierung erstellt werden.

Abhilfe für das Problem der teilweise fehlerhaften Segmentierung beziehungsweise der Abhängigkeit vom Hintergrund, könnten nicht-quadratische Prototypen schaffen. Dazu können zunächst quadratische Prototypen berechnet und in einem weiteren Schritt verfeinert werden. In [RPN01] wird ein Verfahren beschrieben, das Objekte auf heterogenem Hintergrund mit nicht-quadratischen Prototypen erkennt.

Bislang berücksichtigt das Verfahren noch keine Rotation der zu erkennenden Objekte. Bereits bei der Berechnung der Feature-Sammlungen, also im Vorverarbeitungsschritt, könnten für die Trainingsdaten Features aus rotierten Bildern berechnet werden. Eine Rotation der Features ist nicht direkt möglich, da einige Features von der Ausrichtung der Originaldaten abhängen.

Literatur

[BLP96] M.C. Burl, T.K. Leung und P. Perona. Face Localization via Shape Statistics. In *Proc. Int. Workshop on Automatic Face and Gesture Recognition*, Seiten 154–159, Juni 1996.

[KMDN03] D. Keysers, M. Motter, T. Deselaers und H. Ney. Training and Recognition of Complex Scenes using a Holistic Statistical Model. In *Proceedings of the 25th DAGM-Symposium Pattern Recognition*, Lecture Notes in Computer Science 2781, Seiten 52–59. Springer Verlag, September 2003.

[RPN01] M. Reinhold, D. Paulus und H. Niemann. Appearance-Based Statistical Object Recognition by Heterogeneous Background and Occlusions. In *Proceedings of the 23rd DAGM-Symposium Pattern Recognition*, Lecture Notes in Computer Science 2191, Seiten 254–261. Springer Verlag, September 2001.

[TvdM01] J. Triesch und C. von der Malsburg. A System for Person-Independent Hand Posture Recognition Against Complex Backgrounds. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 23(11):1449–1453, November 2001.