Objektklassifikation mit Mischverteilungen

J. Dahmen, K. Beulen, H. Ney

Lehrstuhl für Informatik VI der RWTH Aachen,
Ahornstraße 55, D-52056 Aachen,
e-mail: {dahmen, beulen, ney}@informatik.rwth-aachen.de

Zusammenfassung. Ziel dieser Arbeit ist die Untersuchung der Fragestellung, inwieweit 'konventionelle' Klassifikatoren aus der statistischen Mustererkennung für die Objektklassifikation in der Bildverarbeitung einsetzbar sind. Dazu wurde ein auf Mischverteilungen basierender Klassifikator implementiert. Zur Merkmalsanalyse wurden die hochdimensionalen Bilddaten geeignet in einen niederdimensionalen Merkmalsraum projiziert. Die Klassifikation der so dimensionsreduzierten Merkmalsvektoren erfolgt über die Bayes'sche Entscheidungsregel. Auf der 'Chair-Image-Database' des Max-Planck-Instituts [1] erzielen wir eine Fehlerrate von 0.64%.

1 Einführung

In der vorliegenden Arbeit wird die Anwendbarkeit 'konventioneller' statistischer Klassifikatoren auf das Problem der Objektklassifikation in der Bildverarbeitung überprüft. Zur Evaluierung des von uns implementierten Klassifikators verwenden wir die 'Chair-Image-Database' des Max-Planck-Instituts, da für dieses Corpus bereits Resultate des Max-Planck-Instituts [1] und der Daimler-Benz-AG [2] vorliegen. Diese Gruppen setzen Verfahren wie neuronale Netze [1], Support-Vektoren [1] oder polynomiale Klassifikatoren [2] ein und erzielen so Fehlerraten zwischen 0.8% und 9% (siehe Kapitel 5). In dem von uns implementierten Klassifikator werden die Merkmalsvektoren der zu klassifizierenden Objekte (im folgenden auch Beobachtungen genannt) vereinfachend als gaußverteilt angesehen und die Parameter dieser Verteilung in einer Trainingsphase mittels des Expectation-Maximization-Algorithmus geschätzt. Die so gewonnenen Modelle werden dann in der Erkennung zur Realisierung der Bayes'schen Entscheidungsregel eingesetzt. Da die Merkmalsvektoren im Falle des von uns gewählten Corpus sehr hochdimensional sind, entwickeln wir zur Merkmalsreduktion die Beobachtungsvektoren nach einer geeignet zu wählenden Orthonormalbasis (siehe Kapitel 4). Unter Verwendung von Mischverteilungen und klassenspezifischen Varianzen erreichen wir eine Fehlerrate von 0.64%.

2 Datensammlung und Aufgabenstellung

Die von uns verwendete 'Chair-Image-Database' des Max-Planck-Instituts besteht aus computergenerierten Bildern von Bürostühlen aus 25 Klassen (`ftp://`

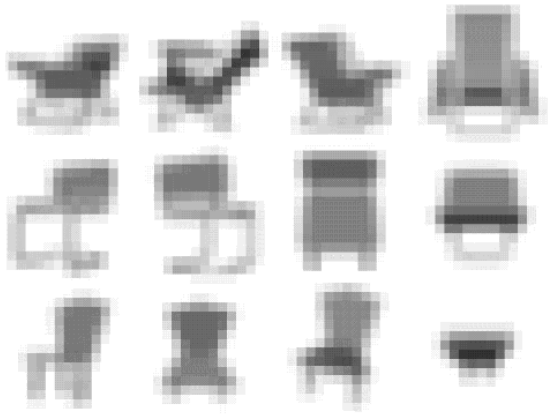

Abb. 1. Beispielbilder verschiedener Stuhlklassen ($16 \cdot 16$ Pixel, Bilder einer Zeile gehören zur gleichen Klasse)

`ftp.mpik-tueb.mpg.de/pub/chair_dataset`). Jede Klasse enthält Aufnahmen eines unter verschiedenen Blickwinkeln betrachteten Stuhls (Abbildung 1). Alle Bilder sind auf ein Format von $16 \cdot 16$ Pixel und auf 256 Grauwerte normiert. Zusätzlich zum eigentlichen Grauwertbild werden noch vier richtungsabhängige Kantenbilder [1] verwendet, so daß die resultierenden Merkmalsvektoren 1280-dimensional sind. Die 'Chair-Image-Database' enthält verschiedene, unterschiedlich große Trainingscorpora und ein Testcorpus. Das Testcorpus besteht aus 2500 Bildern, zeigt also jeden Stuhl aus 100 verschiedenen Blickrichtungen, während die Anzahl der Blickrichtungen für die Trainingscorpora zwischen 25 und 400 liegt. In unseren Experimenten setzen wir sowohl das Corpus mit 89 Blickwinkeln (Train89) als auch das mit 400 Blickwinkeln (Train400) pro Klasse ein. Die Aufgabenstellung besteht darin, mit Hilfe der Trainingscorpora einen Klassifikator zu entwerfen, der dann mit dem Testcorpus evaluiert werden kann.

3 Bayes'sche Entscheidungsregel und Mischverteilungen

Für die Klassifikation der Daten verwenden wir die Bayes'sche Entscheidungsregel [3, S.10-43]. Sind sowohl die *a-priori* Wahrscheinlichkeit $p(k)$ als auch die klassenbedingte Wahrscheinlichkeit $p(x|k)$ bekannt, so läßt sich ein Beobachtungsvektor x mit folgender Regel klassifizieren:

$$\hat{k} = \underset{k}{\operatorname{argmax}} \{p(k)p(x|k)\} \tag{1}$$

Da weder die a-priori Klassenwahrscheinlichkeit $p(k)$ noch die klassenbedingte Wahrscheinlichkeit $p(x|k)$ bekannt sind, müssen diese aus den vorhandenen Trainingsdaten gelernt werden. Die a-priori Wahrscheinlichkeit wird über relative Häufigkeiten geschätzt, während die klassenbedingten Wahrscheinlichkeiten mit Gaußschen Mischverteilungen modelliert werden.

3.1 Gaußsche Mischverteilungen

Eine Mischverteilung (2) ist die gewichtete Summe von Einzelverteilungen mit Parametermenge ϑ_{ki}, wobei für die Gewichte die Bedingungen $\sum_i c_{ki} = 1$ sowie $c_{ki} > 0$ erfüllt sein müssen:

$$p(x|k) = \sum_{i=1}^{I_k} c_{ki} \cdot p(x|k, i, \vartheta_{ki}) \tag{2}$$

Wir verwenden für die Einzelverteilungen $p(x|k, i, \vartheta_{ki})$ die Gaußverteilung (3) und bezeichnen diese als Mischverteilungskomponenten oder Dichten:

$$p(x|k, i, \vartheta_{ki}) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_{ki})}} \exp\left[-\frac{1}{2}(x - \mu_{ki})^T \Sigma_{ki}^{-1}(x - \mu_{ki})\right] \tag{3}$$

Die Parametermenge einer Mischverteilung k umfaßt also neben den Parametern μ_{ki} und Σ_{ki} auch Gewichte c_{ki}, so daß im folgenden $\vartheta_{ki} = (c_{ki}, \mu_{ki}, \Sigma_{ki})$ gelte. Im Gegensatz zu Einzelverteilungen können Mischverteilungen auch multimodale Verteilungen beschreiben, so daß ihre Verwendung in der Regel zu besseren Resultaten führt. Das Schätzen von empirischen Kovarianzmatrizen erweist sich jedoch oftmals als schwierig. Zur Behebung dieses Problems gibt es drei Möglichkeiten:

a) Schätze nur eine Diagonalmatrix statt der vollen Kovarianzmatrix, betrachte also nur einen Varianzvektor.
b) *klassenspezifisches Pooling*: Bestimme für jede Klasse k nur ein Σ_k, also: $\Sigma_{ki} = \Sigma_k \; \forall \; i \in I_k$.
c) *globales Pooling*: Bestimme nur eine einzige Kovarianzmatrix Σ, also: $\Sigma_{ki} = \Sigma \; \forall \; k$ und $\forall \; i \in I_k$.

Wir werden im folgenden stets Möglichkeit a) in Kombination mit b) oder c) wählen. Da eine diagonale Kovarianzmatrix nur für unkorrelierte Merkmale geeignet ist, werden die Daten vor der Parameterschätzung einer Whitening-Transformation [4, S.26-29] unterworfen. Die mittlere klassenbedingte Kovarianzmatrix

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{k_n})(x_n - \mu_{k_n})^t \tag{4}$$

der so transformierten Daten ist dann die Einheitsmatrix. Dabei sei N die Anzahl der Trainingsdaten, x_n das n. Trainingsdatum, k_n der zu x_n gehörige Klassenindex und μ_{k_n} der empirische Mittelwertsvektor der Klasse k_n.

4 Der Klassifikator

4.1 Systemarchitektur

Der zu entwerfende Klassifikator hat die in Abbildung 2 dargestellte Struktur. Die Transformationsmatrix U entspricht dabei der aus den Trainingsdaten gewonnenen Whitening-Transformationsmatrix (Kapitel 3). Da wir Eigenvektoren zum Eigenwert Null nicht berücksichtigen, erfolgt in diesem Schritt bereits eine Dimensionsreduktion auf 1231 Dimensionen. Zur Bestimmung der Transformationsmatrix T betrachten wir K Vektoren der Form $\mu_k - \mu$, wobei μ_k der empirische Mittelwertsvektor der Klasse k und μ der empirische Mittelwertsvektor aller Beobachtungen sei. Für den von diesen Vektoren aufgespannten Unterraum wird eine Orthonormalbasis bestimmt. Um die durch Rundungsfehler bedingten numerischen Probleme der üblichen *Gram-Schmidt-Orthogonalisierung* zu vermeiden, verwenden wir hierzu eine *Singular-Value-Decomposition* (SVD) [5, S.59-67]. Man erhält maximal $K - 1$ Basisvektoren [4, S.451]. Die lineare Transformation T ist nun die Projektion der Merkmalsvektoren in diesen Unterraum, der im Fall der von uns verwendeten Daten genau $K - 1$ dimensional ist.

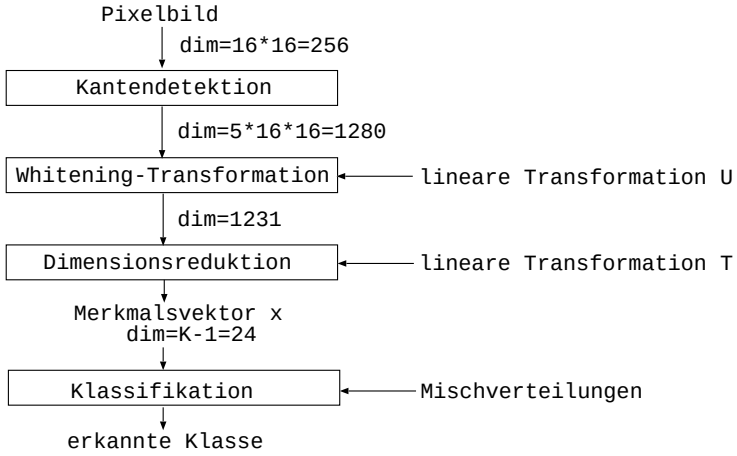

Abb. 2. Grundstruktur des Gesamtsystems

4.2 Training der Mischverteilungen

Die noch zu bestimmenden Verteilungsparameter werden mit dem EM-Algorithmus, einem Maximum-Likelihood-Schätzverfahren für Daten mit verborgenen Variablen, geschätzt. Wir geben Schätzformeln für die zu bestimmenden Parameter an, die sich aus der Anwendung des EM-Algorithmus auf unser Problem ergeben. Weiterhin beschreiben wir eine vereinfachte Form dieses Algorithmus, die auf einer Maximum-Approximation beruht.

Der Expectation-Maximization-Algorithmus: Der EM-Algorithmus ist ein iteratives Schätzverfahren zur Bestimmung der Parameter einer Wahrscheinlichkeitsverteilung mit verborgenen Variablen. Seine Anwendung auf Mischverteilungen wird in [6] beschrieben, wobei der Index der zu einer Beobachtung gehörigen Dichte als verborgene Variable interpretiert wird. Diese Zuordnung wird als Wahrscheinlichkeit $p(i|x,k,\vartheta_{ki})$ ausgedrückt, wobei k der Klassenindex, i der Index der Mischverteilungskomponente und ϑ_{ki} deren Parameter sind. Die Schätzformeln für die Parameter der Verteilung $p(x|k,i,\vartheta_{ki})$ ergeben sich für den EM-Algorithmus zu:

$$p(i|x,k,\vartheta_{ki}) = \frac{c_{ki} \cdot p(x|k,i,\vartheta_{ki})}{\sum_{i'} c_{ki'} \cdot p(x|k,i',\vartheta_{ki'})} \tag{5}$$

$$\bar{\mu}_{ki} = \sum_{n=1}^{N_k} \frac{p(i|x_n,k,\vartheta_{ki})}{\sum_{n'} p(i|x_{n'},k,\vartheta_{ki})} x_n \tag{6}$$

$$\bar{\Sigma}_{ki} = \sum_{n=1}^{N_k} \frac{p(i|x_n,k,\vartheta_{ki})}{\sum_{n'} p(i|x_{n'},k,\vartheta_{ki})} [x_n - \bar{\mu}_{ki}][x_n - \bar{\mu}_{ki}]^T \tag{7}$$

$$\bar{c}_{ki} = \frac{1}{N_k} \sum_{n=1}^{N_k} p(i|x_n,k,\vartheta_{ki}) \tag{8}$$

wobei N_k die Anzahl der Beobachtungen der Klasse k und x_n der n. Merkmalsvektor dieser Klasse sei. Mit einer initialen Schätzung der Parameter (c_{ki}, μ_{ki}, Σ_{ki}) wird zunächst ein $p(i|x,k,\vartheta_{ki})$ bestimmt. Mit $p(i|x,k,\vartheta_{ki})$ können die Parameter ϑ_{ki} neu geschätzt werden, die wiederum eine weitere Schätzung von $p(i|x,k,\vartheta_{ki})$ erlauben. Dieser Vorgang wird nun bis zur Konvergenz wiederholt. Sowohl die Anzahl der Mischverteilungskomponenten als auch die initialen Werte ihrer Parameter werden über das sukzessive Spalten von Mischverteilungskomponenten bestimmt. Begonnen wird dabei mit dem Schätzen einer Einzelverteilung. Diese wird sukzessive gespalten, d.h. eine Einzelverteilung ki wird durch das Stören des Mittelwertsvektors μ_{ki} über einen Störvektor ϵ durch zwei Einzelverteilungen ersetzt. Dieses Verfahren hat sich für die Modellierung von Emissionswahrscheinlichkeiten in der Spracherkennung bewährt [7]. Man erhält zwei Mittelwertsvektoren $\mu_{ki}^+ = \mu_{ki} + \epsilon$ sowie $\mu_{ki}^- = \mu_{ki} - \epsilon$, also eine Mischverteilung mit den beiden Dichten $p(x|\mu_{ki}^+, \Sigma_{ki})$ und $p(x|\mu_{ki}^-, \Sigma_{ki})$. Nun werden die Parameter dieser Dichten mittels (5)-(8) neu geschätzt. Dieser Vorgang wird iteriert, bis die gewünschte Anzahl Dichten erreicht ist. Dabei werden alle Einzelverteilungen i der Klasse k gespalten, für die gilt:

$$\sum_{n=1}^{N_k} p(i|x_n,k,\vartheta_{ki}) \geq N_{Split} \tag{9}$$

Maximum-Approximation: Bei der Maximum-Approximation wird jeder Trainingsvektor nur einer Mischverteilungskomponente zugeordnet, d.h. $p(i|x,k,\vartheta_{ki})$ wird modelliert als:

$$p(i|x,k,\vartheta_{ki}) = \begin{cases} 1 & : \quad i = \text{'beste' Einzelverteilung} \\ 0 & : \quad \text{sonst} \end{cases} \tag{10}$$

Die Rechtfertigung dieser Vorgehensweise ergibt sich aus dem exponentiellen Abfall der Gaußverteilung, der zur Folge hat, daß in der Regel ein dominierendes $p(i|x,k,\vartheta_{ki})$ existiert. Vorteil der Maximum-Approximation ist, daß die relativ aufwendige Bestimmung der $p(i|x,k,\vartheta_{ki})$ ersetzt wird durch eine einfache Bestimmung des maximalen $p(x|k,i,\vartheta_{ki})$. Diese reduziert sich bei Logarithmierung von Gleichung (3) im wesentlichen auf die Bestimmung des quadrierten euklidischen Abstands zwischen Mittelwerts- und Beobachtungsvektor.

5 Ergebnisse

Den Beginn unserer Experimente bilden Gaußsche Einzelverteilungen, die mit den Originalmerkmalsvektoren, also ohne vorhergehende Dimensionsreduktion, trainiert werden. Als Problem erweisen sich dabei die synthetisch erstellten Bilddaten, in denen Merkmale mit einer empirischen Varianz gleich Null existieren. Dies sind in der Regel Pixel, die am Bildrand gelegen sind und in allen Bildklassen zum Bildhintergrund gehören. Um dieses Problem zu umgehen, setzen wir zwei Verfahren ein: Im ersten Fall werden Nullvarianzen durch extrem kleine Varianzen approximiert, alternativ werden solche Merkmale 'ausmaskiert', d.h. nicht zur Klassifikation verwendet. Wie die in Tabelle 1 dargestellten Ergebnisse zeigen, erweist sich das Ausmaskieren als die deutlich bessere Wahl. Die Tatsache, daß auf dem Train89-Corpus bessere Ergebnisse erzielt werden als auf dem Train400-Corpus ist dadurch zu erklären, daß die Blickwinkel für das kleine Trainingscorpus und das Testcorpus äquidistant, die für das Train400-Corpus jedoch zufällig gewählt sind. Eine vorhergehende Dimensionsreduktion auf 24 Merkmale mittels der in Kapitel 4 vorgestellten linearen Transformation T reduziert die Fehlerrate auf dem Train400-Corpus auf 3.36% (Einzelverteilung, klassenspezifisches Pooling). Abbildung 3 zeigt die Fehlerrate für die Verwendung von Mischverteilungen auf den dimensionsreduzierten Daten. Dabei werden

Tabelle 1. Fehlerraten mit verschiedenen Trainingscorpora und Varianzberechnungen (Einzelverteilungen, keine Dimensionsreduktion)

Trainingscorpus	Varianz-maskierung	Varianz-modellierung	Fehlerrate [%]
Train400	nein	klassenspezifisch	54.72
	ja	global	24.28
		klassenspezifisch	15.08
Train89	nein	klassenspezifisch	55.08
	ja	global	22.60
		klassenspezifisch	14.24

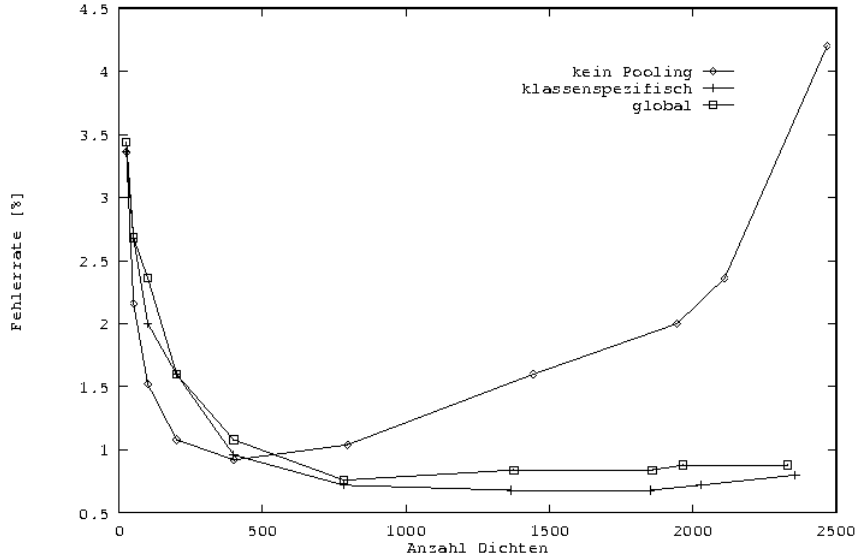

Abb. 3. Fehlerraten mit Train400 in Abhängigkeit von der Zahl der Dichten für verschiedene Arten des Varianzpoolings (EM-Algorithmus)

sowohl die Anzahl der Dichten als auch die Art des Varianzpoolings variiert. Das optimale Ergebnis von 0.64% (Tabelle 2) wird mit insgesamt 2292 Dichten und klassenspezifischer Varianz erreicht. Auf dem Train89-Corpus erzielen wir eine Fehlerrate von 1.72% (456 Dichten). Wie Tabelle 2 zeigt, gibt es auf den von uns verwendeten Daten keine nennenswerten Unterschiede zwischen dem Einsatz von EM-Algorithmus und Maximum-Approximation. Tabelle 3 zeigt den Vergleich unserer Ergebnisse mit Resultaten anderer Arbeitsgruppen, die auf denselben Daten erzielt wurden.

6 Zusammenfassung und Ausblick

Wir haben in dieser Arbeit einen auf Mischverteilungen basierenden Ansatz zur Objektklassifikation vorgestellt. Obwohl dieser auf Maximum-Likelihood-Training beruht, im Gegensatz zu den Verfahren der anderen Arbeitsgruppen also nicht diskriminativ ist, sind unsere Ergebnisse mit diesen vergleichbar bzw. überlegen. Die beste Fehlerrate von 0.64% erzielen wir mit Gaußschen Mischverteilungen und klassenspezifischen Varianzen. Neben einem Einsatz des Verfahrens auf realen Daten sind für die Zukunft weitere Verbesserungen des Algorithmus geplant, beispielsweise die Realisierung eines diskriminativen Trainings oder eine weitere Vorverarbeitung der Merkmalsvektoren (Invarianz gegenüber Rotation, Translation und Skalierung, Normierung der Daten, etc.).

Tabelle 2. Fehlerraten für Maximum-Approximation und EM-Algorithmus mit Train400 (klassenspezifische Varianz)

EM-Algorithmus		Maximum-Approximation	
Dichten	Fehlerrate [%]	Dichten	Fehlerrate [%]
25	3.36	25	3.36
50	2.68	50	2.68
100	2.00	100	1.92
200	1.60	200	1.40
399	0.96	400	0.88
782	0.72	780	0.84
1365	0.68	1406	0.76
1853	0.68	2292	0.64
2358	0.80	2373	0.80

Tabelle 3. Gegenüberstellung der Ergebnisse verschiedener Gruppen

Gruppe	Methode	Trainingscorpus	Fehlerrate [%]
Max-Planck-Institut [1]	Support-Vektoren	Train89	1.0
	Neuronales Netz	Train89	9.0
Daimler-Benz AG [2]	polynom. Klassifikator	Train89	1.68
		Train400	0.80
diese Arbeit	Mischverteilungen	Train89	1.72
		Train400	0.64

Literatur

1. V. Blanz, B. Schölkopf, H. Bülthoff, C. Burges, V. Vapnik, T. Vetter, "Comparison of View–Based Object Recognition Algorithms using Realistic 3D Models," C. von der Malsburg, W. von Seelen, J.C. Vorbrüggen, B. Sendhoff (eds.): *Artificial Neural Networks - ICANN'96* , Springer Lecture Notes in Computer Science Vol. 1112, S.251-256, Berlin, 1996.
2. U. Kressel, persönliche Mitteilung, Daimler-Benz-AG Research and Technology, Abteilung F3M/T, Ulm, März 1998.
3. R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, 1990.
5. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C*, University Press, Cambridge, 1992.
6. A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39(B), 1-38, 1977.
7. H. Ney, "Acoustic Modelling of Phoneme Units for Continuous Speech Recognition," L. Torres, E. Masgrau, M.A. Lagunas (eds.): *Signal Processing V: Theories and Applications*, Elsevier Science Publishers B.V., 1990.

Dieser Artikel wurde mit dem LaTeX Makro-Paket und dem LLNCS-Style formatiert.