Statistische Objektlokalisation und -klassifikation

von Abdulla Elyas (Betreuer: Jörg Dahmen)

In dieser Arbeit wird eine Idee vorgestellt, die einen Standard-Ansatz der statistischen Objekterkennung verbessern möchte. Häufig wird vor der Trainingsphase der Objekterkennung eine Dimensionsreduktion durchgeführt, die die zu den gegebenen Beobachtungen gehörigen Merkmalsvektoren in einen niedrigdimensionalen Merkmalsraum projiziert. Dieser Schritt erfolgt in der Regel, um die Anzahl der zu schätzenden Modellparameter des statistischen Klassifikators zu reduzieren. Durch die alleinige Betrachtung dieses Merkmalsraums gehen jedoch Informationen verloren, die sich im unbeachteten 'Nicht-Merkmalsraum' befinden. Die hier vorgestellte Arbeit [1] stellt ein Verfahren vor, das die genannten Informationsverluste durch die Betrachtung des 'Nicht-Merkmalsraums' zu kompensieren sucht und zwar durch die Berücksichtigung des bei der Datenreduktion entstehenden Rekonstruktionsfehlers. Es wird ein effizientes Rechenverfahren vorgestellt, das den mit diesem Ansatz einhergehenden Rechenaufwand bis auf ein Minimum verringert. Eine Verbesserung der Klassifikatorleistung erscheint möglich, wird jedoch letztlich nicht überzeugend verifiziert.

Keywords: Klassifikation, Lokalisation, PCA, Mischverteilungen, Rekonstruktionsfehler

Inhaltsverzeichnis

1 Einleitung

Die Objekterkennung wird in vielen modernen Anwendungen zunehmend wichtiger. Die Herausforderung, Techniken zu entwickeln, die den hohen Anforderungen der Anwender entsprechen und mit dem aktuellen Stand der Technik kostendeckend realisierbar sind, treibt die Forschung immer weiter voran.

Bei der klassischen Objekterkennung steht man in der Regel vor dem Problem, über eine zu geringe Anzahl an Trainingsdaten zu verfügen, was dazu führt, das nicht beliebig viele Modellparameter aus diesen Daten gelernt werden können.

Es hat sich gezeigt, daß eine Merkmalsreduktion der Trainingsdaten, trotz des zwangsläufigen Informationsverlustes, bessere Ergebnisse liefert, da die Parameter des statistischen Klassifikators im reduzierten Merkmalsraum zuverlässiger gelernt werden können. Man betrachtet deswegen im Allgemeinen nur noch einen dimensionsreduzierten Merkmalsraum. Bei dieser Vorgehensweise ist ein Rekonstruktionsfehler unvermeidbar. Er entsteht durch die alleinige Betrachtung des Merkmalsraums - der sogenannte 'Nicht-Merkmalsraum' (der bei der Dimensionsreduktion wegfällt) wird völlig ignoriert.

Im Folgenden sei zunächst der klassische Ansatz eingeführt. Der Schwerpunkt liegt in der Vorstellung des erweiterten Ansatzes, der 1997 von Pentland und Moghaddam vorgestellt wurde [1]. Ihre Idee liegt darin, den genannten Informationsverlust durch die Betrachtung des Nicht-Merkmalsraums zu kompensieren. Durch das vorgestellte Verfahren wird es möglich, die hierzu notwendigen Berechnungen fast vollständig im niedrigdimensionalen Merkmalsraum durchzuführen. In Anbetracht der in den meisten Anwendungen hohen Dimension des Nicht-Merkmalsraumes, macht besonders diese Tatsache (im Hinblick auf die benötigte Rechenzeit) den vorgestellten Ansatz attraktiv. Im Folgenden sei deswegen besonderes Augenmerk auf die effizente Berechnung des Rekonstruktionsfehlers gelegt.

Abschließend wird die Leisungsfähigkeit des auf dieser Basis entwickelten Klassifikators durch verschiedene (von den Autoren durchgeführte) Experimente untersucht.

1.1 Einführung in die Objektklassifikation

Im Allgemeinen werden bei einer Klassifikation Objekte entsprechend ihrer Eigenschaften verschiedenen Objektgruppen bzw. Klassen zugeordnet. Niemann [2] definiert den Begriff der Klassifikation wie folgt:

> *Eine Klassifikation ist die Zuordnung eines Objekts/Musters zu genau einer von mehreren möglichen Klassen. Die Rückweisung eines Objekts/Musters, also die Zuordnung zu keiner der Klassen, ist zulässig.*

Diese Definition ist nur eine von vielen möglichen Umschreibungen für die Objektklassifikation. Die Zulässigkeit einer Rückweisung ist beispielsweise ein sehr umstrittener Punkt. Oft wird diese Vorgehensweise untersagt und eine Rückweisung als Fehlklassifikation interpretiert. Generell erfolgt die Klassifikation einer Beobachtung oder Musters durch einen sogenannten Klassifikator, ein Verfahren, das über die Zugehörigkeit des Objekts zu genau einer von K möglichen Klassen entscheidet.

$$x \longrightarrow \boxed{\text{Klassifikator}} \longrightarrow k \in \{1, \ldots, K\} \tag{1}$$

Als Input erhält der Klassifikator einen sogenannten Merkmalsvektor x, der die Eigenschaften bzw. Merkmale der Beobachtung repräsentiert. Anhand dieses Merkmalsvektors entscheidet das Klassifikationsverfahren darüber, zu welcher von K möglichen Klassen $1, \ldots, K$ das Objekt zugeordnet wird. Die Klassenzugehörigkeit einer Beobachtung kann vom Klassifikator entweder relativ oder absolut angegeben werden, d.h. der Klassifikator gibt entweder eine Wahrscheinlichkeit für die Zugehörigkeit eines Objekts zu jeder Klasse oder er liefert als Output die Klassenbezeichnung selbst.

Der Merkmalsvektor besteht aus verschiedenen Merkmalen bzw. Eigenschaften des Objekts. Diese Merkmale spannen den Merkmalsraum auf, in dem jedes Objekt durch einen Punkt beschrieben wird. Ein Klassifikator ist nun eine Diskriminantenfunktion, die einer Beobachtung bzw. ihrem zugehörigen Merkmalsvektor $x \in \mathbb{R}^D$ eine der möglichen Klassen $k = 1, \ldots, K$ zuordnet.

$$r : \mathbb{R}^D \longmapsto \{1, \ldots, K\} \tag{2}$$

$$x \longmapsto r(x) \tag{3}$$

Die geeignete Wahl der Entscheidungsregel $r(x)$ trägt zu einem Großteil zur Qualität des Klassifikators bei. Hierauf sei im folgenden Abschnitt näher eingegangen.

Um die obigen Ausführungen über die Grundlagen der Objektklassifikation zu verdeutlichen sind nun einige Anwendungsbeispiele aufgeführt. Eine bekannte Aufgabenstellung der Mustererkennung im Zusammenhang mit der Klassifikation von Objekten ist die Erkennung von handgeschriebenen Adressen auf Postsendungen. Insbesondere die Verarbeitung der Postleitzahlen zeigt den Bezug zur Klassifikation. Hier sind die Objekte gerade die Ziffern 0 bis 9 und diese sollen in die Klasse 0 bis Klasse 9 klassifiziert werden. Weitere Anwendungsbereiche sind die Handschriftenerkennung ganz allgemein oder auch die Erkennung (isoliert) gesprochener Worte, wobei bei Letzterem die zu klassifizierenden Objekte die Audiosignale der Worte sind, und diese Klassen zugeordnet werden sollen, die gerade den Worten entsprechen. Im Bereich der Medizin kann man sagen, daß jede Diagnose, die ein Arzt stellt eine Klassifikation ist. Der Patient mit seinen Krankheitssymptomen, die den Merkmalen entsprechen, soll in eine der Klassen *krank* oder *gesund* eingeordnet werden, wobei die Klasse *krank* differenzierter betrachtet werden kann, also in verschiedene Krankheiten unterteilt. Eine konkrete Anwendung in der Medizin wäre beispielsweise ein *Diagnoseunterstützungssystem.*

1.2 Bayes'sche Entscheidungsregel

Für die Klassifikation der Daten verwenden wir die Bayes'sche Entscheidungsregel [3, S.10-43]:

$$x \longmapsto r(x) = \underset{k}{\operatorname{argmax}}\, p(k|x) \tag{4}$$

Die Funktion *argmax* liefert hier dasjenige k zurück, welches $p(k|x)$ maximiert. Demnach wird klassifiziert, indem die Funktion die a-posteriori Wahrscheinlichkeiten $p(k|x)$ vergleicht und die Klasse mit der höchsten Wahrscheinlichkeit zurückgibt.

Nach dem Satz von Bayes gilt:

$$p(k|x) = \frac{p(k) \cdot p(x|k)}{p(x)} \tag{5}$$

wobei $p(k)$ die a-priori Wahrscheinlichkeit von k und $p(x|k)$ die klassenbedingte Wahrscheinlichkeit von x gegeben k ist. Da der Nenner $p(x)$ nicht mehr vom maximierenden Argument k abhängt, kann dieser vernachlässigt werden, womit sich folgende Formel für die Bayes'sche Entscheidungsregel ergibt:

$$x \longmapsto r(x) = \underset{k}{\operatorname{argmax}} \{p(k) \cdot p(x|k)\} \tag{6}$$

Man kann zeigen, daß für den Fall bekannter *a-priori* Wahrscheinlichkeit $p(k)$ und klassenbedingter Wahrscheinlichkeit $p(x|k)$ die Baye'sche Entscheidungsregel optimal im Hinblick auf die erwartete Fehlerrate ist, es also keinen Klassifikator mit niedrigerer Fehlerrate gibt. Da diese Verteilungen jedoch in der Regel unbekannt sind, müssen sie aus den vorhandenen Trainingsdaten anhand von geeignet gewählten Modellen gelernt werden. Die a-priori Wahrscheinlichkeit wird meist über relative Häufigkeiten geschätzt, in der vorgestellten Arbeit jedoch mit einer Gleichverteilung aller a-priori Wahrscheinlichkeiten $p(k)$ modelliert:

$$p(k) = \frac{1}{K} \quad \forall \quad k \in \{1, \ldots, K\} \tag{7}$$

Die klassenbedingten Wahrscheinlichkeiten $p(x|k)$ werden mit Gauß'schen Mischverteilungen modelliert - ein Verfahren, was sich gegenüber der Modellierung mit Gauß'schen Einzelverteilungen in den meisten praktischen Anwendungen durchgesetzt hat, da eine multimodale Darstellung die Generierung besserer Modelle für die Realität möglich macht. Der Objekterkennung mit Mischverteilungen [4] sei deshalb das folgende Kapitel gewidmet.

2 Objekterkennung mit Mischverteilungen

Eine Mischverteilung (8) ist eine Linearkombination von Einzelverteilungen $\mathcal{N}(x|\mu_{ki}, \Sigma_{ki})$ (im folgenden auch kurz 'Dichte' genannt).

$$p(x|k) = \sum_{i=1}^{I_k} c_{ki} \cdot \mathcal{N}(x|\mu_{ki}, \Sigma_{ki}) \tag{8}$$

Dabei ist I_k die Anzahl der Dichten der Klasse k, c_{ki} das Gewicht, $\mu_{ki} \in \mathbb{R}^D$ der Mittelwertsvektor und $\Sigma_{ki} \in \mathbb{R}^{DxD}$ die Kovarianzmatrix der Dichte i der Klasse k. Für die Gewichte müssen die Bedingungen $\sum_{i=1}^{I_k} c_{ki} = 1$ sowie $c_{ki} > 0$ erfüllt sein. Wir verwenden für die Einzelverteilungen $\mathcal{N}(x|\mu_{ki}, \Sigma_{ki})$ die Gaußverteilung (9) und bezeichnen diese als Mischverteilungskomponenten oder Dichten:

$$\mathcal{N}(x|\mu_{ki}, \Sigma_{ki}) = \frac{1}{\sqrt{(2\pi)^D det(\Sigma_{ki})}} \exp\left[-\frac{1}{2}(x-\mu_{ki})^T \Sigma_{ki}^{-1}(x-\mu_{ki})\right] \tag{9}$$

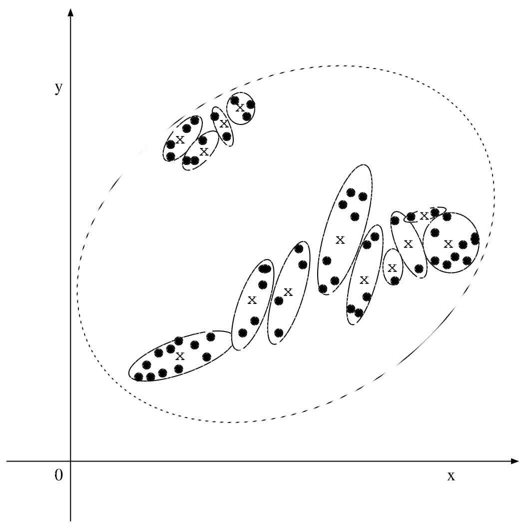

Abbildung 1: Modellierung der Verteilung einer Klasse k mit Einzel- bzw. Mischverteilungen

Im Gegensatz zu Einzelverteilungen können Mischverteilungen auch multimodale Verteilungen beschreiben, so daß ihre Verwendung, in der Praxis häufig zu besseren Ergebnissen führt. In Abbildung 1 ist dieser Vorteil modellhaft veranschaulicht. Die punktförmig dargestellten Merkmale der Klasse k liegen unregelmäßig im zweidimensionalen Merkmalsraum verstreut. Die gestrichelt eingezeichnete unimodale Modellierung weist einen sehr großen "Verschnitt" auf. Die multimodale Modellierung ist hingegen in der Lage, die vorliegende Klasse schärfer zu modellieren und liefert damit eine realitätsnähere Beschreibung der Beispielsklasse k. Nachdem nun der Vorteil der Modellierung mit Mischverteilungen klar geworden ist, stellt sich die Frage, wie die Parameter der Dichten $(c_{ki}, \mu_{ki}, \Sigma_{ki})$ ermittelt werden. Die Lösung dieses Problems liefert der Expectation-Maximization-Algorithmus (EM-Algorithmus), der während der Trainingsphase beim Schätzen der Parameter zum Einsatz kommt. Im folgenden Abschnitt sei näher darauf eingegangen.

3 Training der Mischverteilungen

In diesem Kapitel beschäftigen wir uns mit dem Training der Mischverteilungsparameter. Dazu verwenden wir den Expectation-Maximization-Algorithmus (EM-Algorithmus), ein Maximum-Likelihood-Schätzverfahren für Daten mit verborgenen Variablen.

3.1 Der Expectation-Maximization-Algorithmus

Die Anwendung des EM-Algorithmus auf Mischverteilungen wird in [5] beschrieben. Als verborgene Parameter wird beim Schätzen von Mischverteilungen der Index der zu einer Beobachtung gehörigen Dichte interpretiert. Diese Zuordnung wird als Wahrscheinlichkeit $p(i|x_n, k, \lambda_{ki})$ ausgedrückt, wobei k der Klassenindex, i der Index der Mischverteilungskomponente und λ_{ki} deren Parameter $\{c_{ki}, \mu_{ki}$ und $\Sigma_{ki}\}$ sind. Die Reestimationsformeln für die Mischverteilungsparameter ergeben sich für den EM-Algorithmus zu:

$$p(i|x_n, k, \lambda_{ki}) = \frac{c_{ki} \cdot \mathcal{N}(x_n|k, i, \lambda_{ki})}{\sum_{i'} c_{ki'} \cdot \mathcal{N}(x_n|k, i', \lambda_{ki'})} \quad (10)$$

$$\gamma_{ki}(n) = \frac{p(i|x_n, k, \lambda_{ki})}{\sum_n' p(i|x_n', k, \lambda_{ki})} \quad (11)$$

$$\bar{c}_{ki} = \frac{1}{N_k} \sum_{n=1}^{N_k} p(i|x_n, k, \lambda_{ki}) \quad (12)$$

$$\bar{\mu}_{ki} = \sum_{n=1}^{N_k} \gamma_{ki}(n) \cdot x_n \quad (13)$$

$$\bar{\Sigma}_{ki} = \sum_{n=1}^{N_k} \gamma_{ki}(n) \cdot (x_n - \bar{\mu}_{ki})(x_n - \bar{\mu}_{ki})^T \quad (14)$$

wobei N_k die Anzahl der Traingsdaten der Klasse k und x_n der n. Merkmalsvektor dieser Klasse sei.

3.2 Anwendung des EM-Algorithmus

Eine Möglichkeit, diese aus den EM-Algorithmus gewonnenen Reestimationsgleichungen zum Training der Mischverteilungen zu verwenden, ist das 'Sukzessive Splitten'. Mit einer initialen Schätzung einer Gauß'schen Einzelverteilung wird zunächst das erste $p(i|x_n, k, \lambda_{ki})$ bestimmt. Mit $p(i|x_n, k, \lambda_{ki})$ können die Parameter λ_{ki} neu geschätzt werden, die wiederum eine weitere Schätzung von $p(i|x_n, k, \lambda_{ki})$ erlauben. Dieser Vorgang wird nun bis zur Konvergenz wiederholt. Anschließend werden die berechneten Dichten 'gespalten' (Split-Vorgang), und die Reestimation kann wieder bis zur Konvergenz durchgeführt werden. Dieses Verfahren hat sich für die Modellierung von Emissionswahrscheinlichkeiten in der Spracherkennung bewährt [6]. In Abbildung 2 ist es als Flussdiagramm dargestellt:

1. Initiale Schätzung der Parameter (Mittelwert, Gewicht, Kovarianz) einer Einzelverteilung.

2. Sukzessives Spalten der vorliegende Dichte, d.h. eine Einzelverteilung ki wird durch das Stören des Mittelwertsvektors μ_{ki} über die Störvektoren $\pm\ \epsilon$ durch zwei Einzelverteilungen ersetzt.

3. Man erhält zwei Mittelwertsvektoren $\mu_{ki}^+ = \mu_{ki} + \epsilon$ sowie $\mu_{ki}^- = \mu_{ki} - \epsilon$, also eine Mischverteilung mit den beiden Dichten $\mathcal{N}(x|\mu_{ki}^+, \Sigma_{ki})$ und $\mathcal{N}(x|\mu_{ki}^-, \Sigma_{ki})$. Nun werden die Parameter dieser Dichten mittels (10)-(14) neu geschätzt.

4. Schritt 2 und 3 werden iteriert, bis die gewünschte Anzahl an Dichten erreicht ist, um die Verteilung der Klasse k hinreichend genau zu beschreiben. Dabei werden alle Einzelverteilungen i der Klasse k gespalten, für die gilt: $\sum_{n=1}^{N_k} p(i|x_n, k, \lambda_{ki}) \geq N_{Split}$ (es werden also nur Dichten gespalten, in denen eine Mindestanzahl von N_{Split} Beobachtungen enthalten sind).

4 Zur Parameterschätzung

Wie in Kapitel 2 beschrieben, geht eine multimodale Beschreibung der Trainingsdaten mit einer größeren Anzahl an Einzelverteilungen pro Klasse einher. Dies hat große Auswirkungen auf die Anzahl der zu schätzenden Parameter. So müßte man beispielsweise schon bei einem nur 16x16 Pixel kleinem Bild 65.000 Einträge der Kovarianzmatrix berechnen - und zwar für jede Dichte. Dies wirft die Frage auf, wie man die hohe Anzahl an Parametern auf ein Minimum reduzieren kann, ohne die Qualität der Klassifikation zu beeinträchtigen. Besonders das Schätzen von empirischen Kovarianzmatrizen erweist sich oftmals als schwierig.

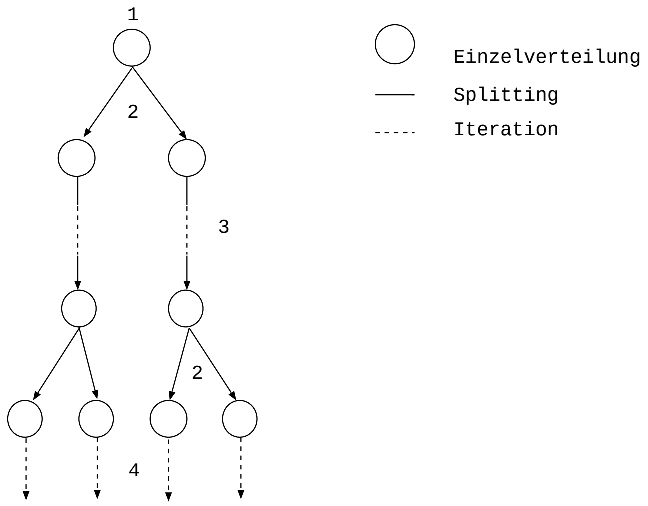

Abbildung 2: Sukzessives Splitten

4.1 Parameterreduktion

Zur Behebung dieses Problems gibt es verschiedene Möglichkeiten, die im Folgenden eingeführt werden. Um die einzelnen Verfahren zu motivieren, sei zunächst das zu schätzende Parametervolumen gegenübergestellt (K=Anzahl Klassen, I=mittlere Anzahl der Dichten je Klasse, D^2=Dimension der Kovarianzmatrix):

- Ohne Parameterreduktion $\Rightarrow K \cdot I \cdot D^2$ Parameter
- Klassenspezifisches Varianzpooling:
  Schätze für jede Klasse k nur eine einzige Kovarianzmatrix: $\Sigma_{ki} = \Sigma_k \ \forall \ i \in \{1, \ldots, I_k\}$
  $\Rightarrow K \cdot D^2$ Parameter
- Globales Pooling:
  Schätze für alle Klassen eine einzige Kovarianzmatrix: $\Sigma_{ki} = \Sigma \ \forall \ k \in \{1, \ldots, K\}, \forall \ i \in \{1, \ldots, I_k\}$
  $\Rightarrow D^2$ Parameter
- Schätze nur Diagonaleinträge der Kovarianzmatrix
  $\Rightarrow D$ Parameter pro Σ (statt D^2)
- Dimensionsreduktion der Merkmalsvektoren: $x \in \mathbb{R}^D \longmapsto y \in \mathbb{R}^d$
  $\Rightarrow$ statt $D \longmapsto d$ Parameter

Da eine diagonale Kovarianzmatrix nur für unkorrelierte Merkmale geeignet ist, können die Daten vor der Parameterschätzung einer Whitening-Transformation [7, S.26-29] unterworfen werden. Die mittlere klassenbedingte Kovarianzmatrix

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{k_n})(x_n - \mu_{k_n})^t \tag{15}$$

der so transformierten Daten ist dann die Einheitsmatrix. Dabei sei N die Anzahl der Trainingsdaten, x_n das n-te Trainingsdatum, k_n der zu x_n gehörige Klassenindex und μ_{k_n} der empirische Mittelwertsvektor der Klasse k_n. Besonders die Dimensionsreduktion ($x \in \mathbb{R}^D \longmapsto y \in \mathbb{R}^d$) erweist sich als eine gute Möglichkeit, die hochdimensionale Kovarianzmatrix robuster zu schätzen. Zu den unterschiedlichen Möglichkeiten der Dimensionsreduktion gehört unter anderem die Principal Components Analysis oder kurz PCA, auf die wir nun eingehen möchten.

4.2 Dimensionsreduktion mit der Principal Components Analysis (PCA)

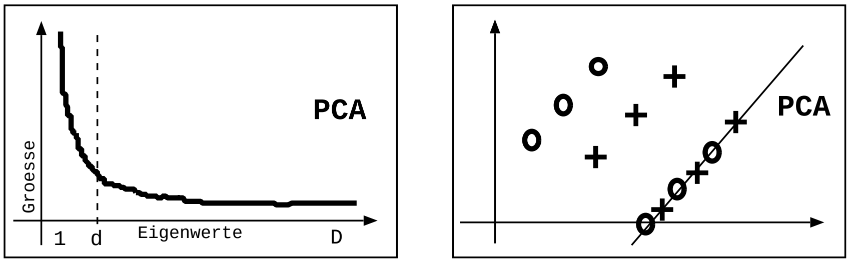

Abbildung 3: PCA - Eigenwertverteilung und geometrische Interpretation

Pentland und Moghaddam verwenden bei der Vorverarbeitung der Trainingsdaten die Principal Components Analysis, um die Merkmalsvektoren auf einen niedrigdimensionalen Merkmalsraum abzubilden. Grundidee der PCA ist es, charakteristische Eigenschaften der hochdimensionalen Merkmalsvektoren (Dimension D) durch eine niedrigdimensionale Abbildung (Dimension d) zu erfassen. Man kann zeigen, daß dies auf die Berechnung der Eigenvektoren der Kovarianzmatrix Σ der gegebenen Daten hinausläuft. Der Merkmalsvektor x wird in einer linearen Abbildung auf den entsprechenden d-dimensionalen Merkmalsvektor y projiziert.

$$x \in \mathbb{R}^D \overset{PCA}{\longmapsto} y \in \mathbb{R}^d \tag{16}$$

Die Merkmale, die sich im $(D-d)$-dimensionalen 'Nicht-Merkmalsraum' befinden werden vernachlässigt. Deshalb bedeutet die Abbildung mit Hilfe der PCA immer auch einen Informationsverlust, der dann zum Tragen kommt, wenn x durch y rekonstruiert werden soll, indem y in den D-dimensionalen Raum zurückprojiziert wird.

$$y \in \mathbb{R}^d \overset{Rek.}{\longmapsto} \bar{x} \in \mathbb{R}^D \tag{17}$$

Ziel der PCA ist es, den bei der Projektion in den Merkmalsraum entstehenden Rekonstruktionsfehler $e^2(x)$ zu minimieren.

$$e^2(x) = \|(x-\bar{x})\|^2 = \sum_{i=d+1}^{D} {y_i}^2 \tag{18}$$

Hierzu bestimmt man zunächst die empirische Kovarianzmatrix Σ (y_i entspricht dem Merkmalsvektor x multipliziert mit dem i-ten Eigenvektor der Kovarianzmatrix Σ)

$$\Sigma = \frac{1}{N} \cdot \sum_{n=1}^{N} (x_n - \mu) \cdot (x_n - \mu)^t \tag{19}$$

und berechnet deren Eigenvektoren $\Phi_i,\ i = 1, \ldots, D$. Diese werden nun entsprechend ihrer zugehörigen Eigenwerte absteigend sortiert. Die gesuchte Transformationsmatrix wird dann aus den ersten d Eigenvektoren gebildet. Die PCA projiziert also die ursprünglichen Merkmalsvektoren in den von den ersten d Eigenvektoren aufgespannten Unterraum. Abbildung 3 zeigt links eine typische Verteilung der Eigenwerte (nach Größe geordnet) und rechts eine geometrische Interpretation der Abbildung durch die PCA. Man sieht, daß bei der PCA die Merkmale der Klasse x und die der Klasse o in ihrer Varianz erhalten bleiben (die Klassentrennung allerdings nicht mehr gelingt!). Eine weitere nützliche Eigenschaft der PCA liegt in der Dekorrelation der Daten. Dies führt dazu, daß die Kovarianzmatrix Σ in diagonalisierter Form vorliegt.

5 Erweiterung des 'Standard-Ansatzes'

Nachdem nun der Standard-Ansatz in groben Zügen eingeführt ist, kann die Idee zur Erweiterung des Ansatzes motiviert werden.

5.1 Motivation

Wie beschrieben, wird nach der Dimensionsreduktion (PCA) in der Regel während der Trainingsphase nur noch der d-dimensionale Merkmalsraum betrachtet. Dies rührt vor allem daher, daß man bei der Mustererkennung oftmals vor dem Problem steht, über zu wenige Trainingsdaten zu verfügen, als daß man alle Parameter des Systems zuverlässig lernen könnte. Die Parameter des reduzierten Merkmalsraums hingegen lassen sich leichter zuverlässig schätzen und führen somit oftmals zu besseren Klassifikationsergebnissen.

$$x \in \mathbb{R}^D \overset{PCA}{\longmapsto} y \in \mathbb{R}^d \quad (F - Raum, \text{`}Feature - Space') \tag{20}$$

Problematisch bei dieser Vorgehensweise ist, daß Informationen, die sich im $(D-d)$-dimensionalen Nicht-Merkmalsraum *($\hat{F}$-Raum, 'Not-Feature-Space')* befinden, nicht berücksichtigt werden. Dieser Informationsverlust kann sich kritisch auf die Klassifikationsergebnisse auswirken, wenn die Merkmalsvektoren zweier unterschiedlicher Klassen im projizierten Raum plötzlich so nah beieinander liegen, daß sie nicht mehr unterscheidbar sind (obwohl sie im Originalraum problemlos unterscheidbar waren). Beispielhaft ist ein solcher Fall in Abbildung 4 demonstriert. Die Merkmale der Klassen "Dreieck" und "Kreis" differieren bezüglich ihrer Y-Koordinate, sie liegen in $\mathbb{R}^D$ (links) weit von einander entfernt. Nach der PCA liegen sie in $\mathbb{R}^d$ (rechts) jedoch wegen der Projektionsrichtung nah bei einander. Eine solche Projektion könnte zu unerwünschten Klassifikationsfehlern führen.

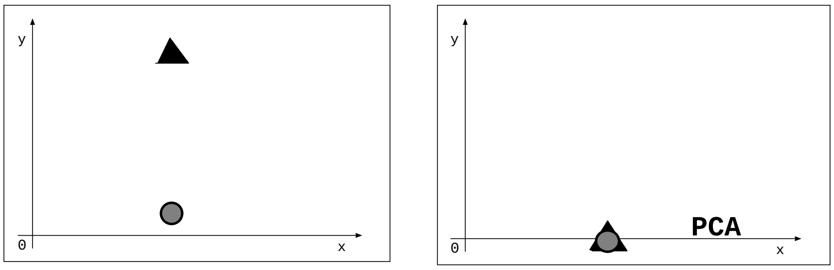

Abbildung 4: Informationsverlust bei der PCA

Abhilfe für dieses Problem sehen Pentland und Moghaddam in der Idee, auch den Nicht-Merkmalsraum zu betrachten. Interessant hierbei ist vor allem ihr vorgestelltes Verfahren, um die hierzu notwendigen Berechnungen fast vollständig im niedrigdimensionalen Merkmalsraum durchzuführen. In Anbetracht der in den meisten Anwendungen hohen Dimension des Nicht-Merkmalsraums, macht besonders diese Tatsache den vorgestellten Ansatz interessant.

5.2 Betrachtung des 'Nicht-Merkmalsraums' ($\bar{F}$-Raum)

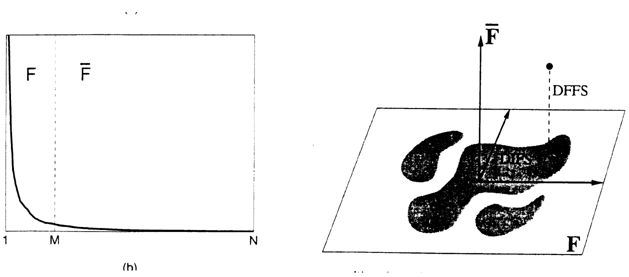

Abbildung 5: Geometrische Interpretation DIFS/DFFS

In Abbildung 5 ist die geometrische Interpretation für die Betrachtung des $\bar{F}$-Raums zu sehen. Während im F-Raum die einfache Distanz innerhalb des Raumes berechnet wird (Distance In Feature

Space (DIFS)), betrachtet man im $\bar{F}$-Raum den orthogonalen Abstand zum Merkmal (Distance From Feature Space (DFFS)). Ziel es ist nun, diese orthogonale Zerlegung auf eine neue Interpretation der klassenspezifischen Wahrscheinlichkeit zu übertragen:

$$p(x|k) \quad = \quad \underbrace{p_F(x|k)}_{PCA^d} \quad \cdot \quad \underbrace{p_{\bar{F}}(x|k)}_{PCA^{D-d}} \tag{21}$$

Vor dem Hintergrund der Objektlokalisation stellt uns die Hochdimensionalität des $(D-d)$-dimensionalen $\bar{F}$-Raums in (21) jedoch vor große Probleme. Am Beispiel in Abbildung 6 sei dies modellhaft demonstriert:

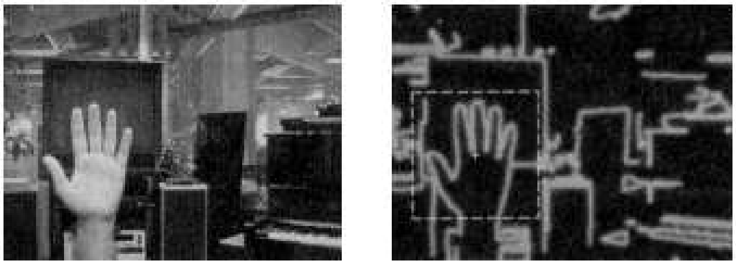

Abbildung 6: Objektlokalisation - Berechnung von $p(x|k)$ per Template

Bei dem zum Einsatz kommenden Maximum-Likelihood-Klassifikator wird zur Objektlokalisation und Objektklassifikation ein Template der Größe des zu suchenden Objekts Bit für Bit über das Testbild verschoben und die Wahrscheinlichkeit $p(x|k)$ in *jedem* betrachteten Pixel berechnet. Wenn das Beispielbild $M \times N$ Pixel groß ist, hieße dies eine etwa $M \times N$-fache Berechnung von $p(x|k)$, um das gesuchte Objekt zu lokalisieren und zu klassifizieren. Eine effiziente Berechnung besonders des $\bar{F}$-Raums ist somit entscheidend für die effiziente Berechnung der klassenbedingten Wahrscheinlichkeit (im Hinblick auf einen zumutbaren Rechenaufwand des ML-Klassifikators).

5.3 Effiziente Berechnung der klassenbedingten Wahrscheinlichkeit

Die Berechnung der klassenbedingten Wahrscheinlichkeit umfaßt im wesentlichen die Berechnung der Mahalanobis-Distanz:

$$d(x,\mu) = \tilde{x}^T \Sigma^{-1} \tilde{x} \qquad mit \qquad \tilde{x} = x - \mu \tag{22}$$

Innerhalb der in Kapitel 4.2 vorgestellten Principal Components Analysis erfolgt durch die sogenannte Karhuen-Loève-Transformation (KLT) eine Diagonalisierung der Kovarianzmatrix Σ. Es folgt dann für die Mahalanobis-Distanz:

$$d(x,\mu) = \sum_{i=1}^{D} \frac{{y_i}^2}{\sigma_i^2} \qquad mit \qquad y_i = (x-\mu)^T \Phi_i \tag{23}$$

Wie jedoch an den Summationsgrenzen schnell zu erkennen ist, bewegen wir uns immer noch im D-dimensionalen Raum, was bedeutet, daß die Rechnung wenig effizient ist. Um die Mahalanobis-Distanz effizient zu berechnen, wählt man den Ansatz, die Summe in zwei unabhängige Teile F und $\bar{F}$ aufzusplitten:

$$d(x,\mu) = \underbrace{\sum_{i=1}^{d} \frac{{y_i}^2}{\sigma_i^2}}_{F} + \underbrace{\sum_{i=d+1}^{D} \frac{{y_i}^2}{\sigma_i^2}}_{\bar{F}} \tag{24}$$

Zumindest die erste Summe entspricht nun einer Projektion im d-dimensionalen F-Raum und kann somit schnell berechnet werden. Die Hochdimensionalität des zweiten Summanden kann man durch eine geschickte Umformulierung des Rekonstruktionsfehlers $e^2(x)$ in den Griff bekommen.

5.4 Effiziente Berechnung des Rekonstruktionsfehlers

Wie im vorangegangenen Kapitel gezeigt, benötigen wir für die effiziente Berechnung der klassenbedingten Wahrscheinlichkeit $p(x|k)$ noch eine schnelle Berechnung der Summe $\sum_{i=d+1}^{D} \frac{{y_i}^2}{\sigma_i^2}$. Der in Kapitel 4.2 eingeführte Rekonstruktionsfehler wird uns hierbei in einer abgeänderten Formulierung weiterhelfen, da er identisch mit dem aufsummierten Zähler der zweiten Teilsumme von (24) ist:

$$e^2(x) = \sum_{i=d+1}^{D} {y_i}^2 \tag{25}$$

Die Formulierung des Rekonstruktionsfehlers, die wir mit (25) angegeben haben ist nicht unmittelbar einsichtig und sei deswegen zunächst in den äquivalenten Umformungen (26) bis (30) kurz hergeleitet. Als Ausgangspunkt der Herleitung soll der Rekonstruktionsfehler in seiner ursprünglichen Darstellung dienen:

$$e^2(x) = \|x - \bar{x}\|^2 \tag{26}$$

$$e^2(x) = \|\sum_{i=1}^{D} y_i\Phi_i - \sum_{i=1}^{d} y_i\Phi_i\|^2 \tag{27}$$

$$e^2(x) = \|\sum_{i=d+1}^{D} y_i\Phi_i\|^2 \tag{28}$$

$$e^2(x) = \sum_{i=d+1}^{D} {y_i}^2 \cdot \underbrace{\|\Phi_i\|^2}_{1} \tag{29}$$

$$\boxed{e^2(x) = \sum_{i=d+1}^{D} {y_i}^2} \tag{30}$$

Die neue Formulierung für $e^2(x)$ ergibt sich aus den äquivalenten Umformungen (31) bis (34):

$$e^2(x) = \sum_{i=d+1}^{D} {y_i}^2 \tag{31}$$

$$e^2(x) = \sum_{i=1}^{D} {y_i}^2 - \sum_{i=1}^{d} {y_i}^2 \tag{32}$$

$$e^2(x) = \|y\|^2 - \sum_{i=1}^{d} {y_i}^2 \tag{33}$$

Da $\tilde{x}$ mittelwertfrei ist und die KLT nur eine Rotation um den Koordinatenursprung darstellt, sind die Längen von y und $\tilde{x}$ identisch. Damit folgt:

$$\boxed{e^2(x) = \|\tilde{x}\|^2 - \sum_{i=1}^{d} {y_i}^2} \tag{34}$$

Die Mahalanobis-Distanz ist demnach fast im d-dimensionalen Raum berechenbar. Bleibt nur noch die Frage nach der effizienten Berechnung von $\|\tilde{x}\|^2$. Obwohl sich Pentland und Moghaddam zu dieser noch offenen Frage in ihrer Arbeit nicht äußern, ist unserer Meinung auch dieses Problem zu lösen. Hierzu sollte man sich die Vorgehensweise des ML-Klassifikators noch einmal vor Augen führen. Anhand des eindimensionalen Beispiels in Abbildung 7 sei ein möglicher Lösungsansatz demonstriert.

Der Klassifikator berechnet an jedem Pixel $k \in \{1, \ldots, P - N\}$ des Bildes:

$$d_k = \sum_{n=1}^{N} (x_{n+k} - \mu_n)^2 = \underbrace{\sum_{n=1}^{N} x_{n+k}^2}_{A} + \underbrace{\sum_{n=1}^{N} \mu_n^2}_{B} - 2\underbrace{\sum_{n=1}^{N} x_{n+k}\mu_n}_{C} \tag{35}$$

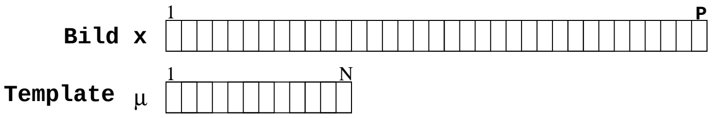

Abbildung 7: Eindimensionales Beispiel für die Berechnung von $\|\tilde{x}\|^2$

Die Teilsumme A kann 'geschickt' berechnet werden, indem man davon absieht, für jeden Index die komplette Summe neu zu berechnen, sondern einfach das erste Pixel (quadriert) subtrahiert und das nächste Pixel (quadriert) addiert:

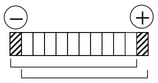

Abbildung 8: Berechnung von $\sum\limits_{n=1}^{N} x_{n+k}^2$

Teilsumme B kann einmal vorab berechnet werden. Die Teilsumme C stellt eine Korrelation zwischen den Vektoren x und μ dar und ist daher effizient unter Verwendung des Faltungssatzes der Fouriertransformation berechenbar. Damit erhalten wir für die unter (23) eingeführte Beschreibung der Mahalanobis-Distanz die approximierte Neuformulierung:

$$\hat{d}(x,\mu) = \sum_{i=1}^{d} \frac{y_i}{\sigma_i^2} + \frac{1}{\rho} \left[\sum_{i=d}^{D} {y_i}^2 \right] \tag{36}$$

$$\boxed{\hat{d}(x,\mu) = \sum_{i=1}^{d} \frac{y_i}{\sigma_i^2} + \frac{e(x)^2}{\rho}} \tag{37}$$

Der aufsummierte Nenner von (23) wird mit ρ approximiert. Dies entspricht einer gepoolten Varianz für $\bar{F}$. Man kann zeigen, daß der optimale Wert für ρ im Mittelwert gefunden ist:

$$\rho_{opt} = \frac{1}{D-d} \sum_{i=d+1}^{D} \sigma_i^2 \tag{38}$$

Die resultierende klassenbedingte Wahrscheinlichkeit $p(x|k)$ ergibt sich dann zu:

$$p(x|k) = \left[\frac{exp(-\frac{1}{2} \sum\limits_{i=1}^{d} \frac{{y_i}^2}{\sigma_i^2})}{2\pi^{\frac{d}{2}} \prod\limits_{i=1}^{d} \sigma_i} \right] \cdot \left[\frac{exp(-\frac{1}{2} \frac{e^2(x)}{\rho})}{(2\pi\rho)^{\frac{D-d}{2}}} \right] \tag{39}$$

$$= p_F(x|k) \qquad \cdot \qquad p_{\bar{F}}(x|k) \tag{40}$$

Wobei $p_F(x|k)$ mit Mischverteilungen modelliert und $p_{\bar{F}}(x|k)$ anhand des Rekonstruktionsfehlers berechnet wird (wie oben beschrieben).

6 Anwendung und Ergebnisse

Pentland und Moghaddam führten zur Demonstration ihres ML-Klassifikators zahlreiche Experimente mit Ergebnissen vor, aus denen die folgenden exemplarisch vorgestellt werden sollen.

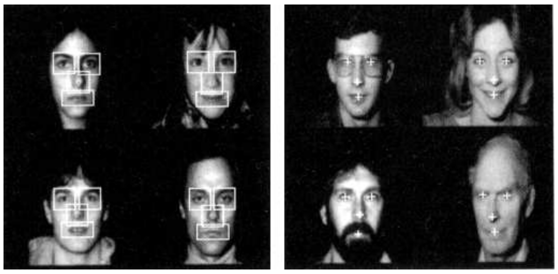

Abbildung 9: Lokalisation von Gesichtsbereichen

6.1 Lokalisation von Gesichtsbereichen

Ziel dieses Experiments ist, die Position von Augen, Nase und Mund in Gesichtsaufnahmen zu bestimmen. Das zu Grunde liegende Datenmaterial umfasst 7562 Bilder der MIT-Datenbank (Media Laboratory's) von 'Fahndungsfotos'. 7000 dieser Fahndungsfotos wurden für den Test benutzt. Wegen der immer gleichbleibenden Orientierung der unterschiedlichen Gesichter, variierten die Trainingsdaten einer Klasse nicht sehr stark. Es genügte deswegen laut Autoren beim Training des ML-Klassifikators Einzelverteilungen zu verwenden. Um die Klassifikatorleistung in Bezug zu anderen Klassifikatoren zu setzen, wurden zusätzlich zum ML-Klassifikator auch ein Square-Sum-Distance- (SSD) und ein Distance-In-Feature-Space-Klassifikator (DIFS) auf die Aufgabenstellung angesetzt. Die Klassifikationskriterien waren

- SSD: 'Nearest Neighbour', d.h. es wird über dem euklidischen Abstand zwischen dem Test- und dem Referenzbild auf Grauwertebene klassifiziert.
- DFFS: $\bar{F}$-Raum, d.h. es wird über den Rekonstruktionsfehler klassifiziert, der F-Raum ist hier 5 dimensional.
- ML: F- und $\bar{F}$-Raum wie beschrieben, die klassenbedingte Wahrscheinlichkeit wird mit Einzelverteilung modelliert, der F-Raum ist hier 5 dimensional.

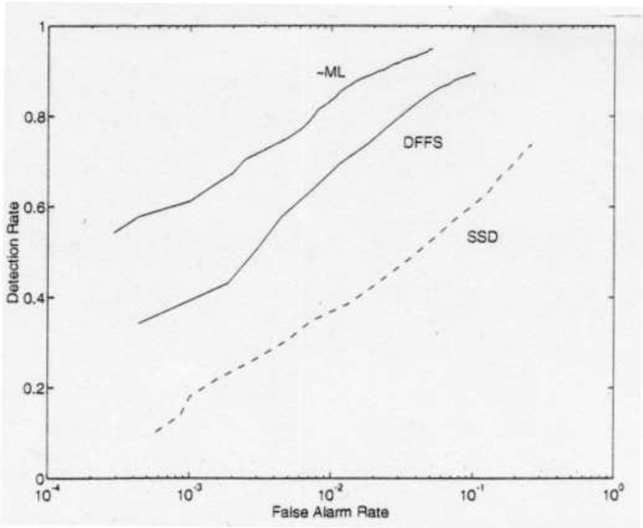

Abbildung 10: Testergebnisse Lokalisation von Gesichtsbereichen

Beim Versuch erzielte der ML-Klassifikator mit einer 5%igen Fehlerrate (DFFS ca. 10%, SSD ca. 25%) die besten Klassifikationsergebnisse. In der Grafik in Abbildung 10 wird die Detektionsrate der

Falschalarmrate gegenübergestellt. Falschalarme treten auf, wenn an einer Stelle ein Gesichtsbereich erkannt wird, an dem sich weder Augen, Nase noch Mund befinden. Wie in der Abbildung deutlich wird, treten bei fester Detektionsrate aller Klassifikatoren die wenigsten Falschalarme beim ML-Klassifikator auf. Auch bei fester Falschalarmrate der Klassifikatoren überzeugt der ML-Klassifikator mit der höchsten Detektionsrate - teilweise mit doppelt so guten Ergebnissen wie der SSD-Klassifikator.

Dieses Ergebnis muß jedoch kritisch durchleuchtet werden. Ein Klassifikator der, wie der DFFS-Klassifikator, im Nicht-Merkmalsraum klassifiziert, gehört nicht unbedingt zum Standard in der Objekterkennung. Wünschenswert wäre deswegen ein Vergleich des ML-Klassifikators mit einem *DIFS*-Klassifikator, der auf dem F-Raum klassifiziert. Die oben genannten guten Ergebnisse des ML-Klassifikators müssen deswegen relativiert werden, da der Gewinn gegenüber einem Klassifikator, der nur auf dem PCA-Merkmalsraum arbeitet, nicht ersichtlich ist.

6.2 Erkennung von Handgesten

In einem weiteren Experiment sollten Handgesten vor unregelmäßigem Hintergrund erlernt werden. Als Datenmaterial dienten 375 Bilder einer Videosequenz mit statischen Handgesten. Die erste Hälfte der Bilder wurde vor schwarzem Hintergrund, die zweite vor gemischtem Hintergrund gelernt. Getestet wurden die Bilder vor gemischtem Hintergrund.

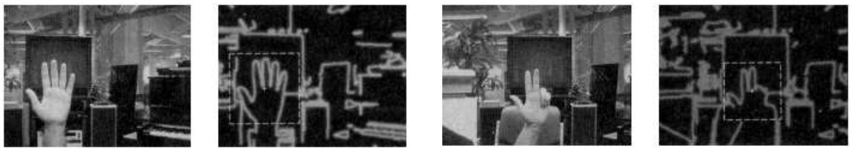

Abbildung 11: Erkennung von Handgesten

Bei diesem Experiment wurden die Testergebnisse des ML-Klassifikators mit denen eines SSD-Klassifikators verglichen. Eine Hand galt als korrekt detektiert, wenn sie in einer Umgebung von $\pm$ 5 Pixel der Originalhand lokalisiert und richtig klassifiziert wurde. Dem ML-Klassifikator lag ein 10-dimensionaler F-Raum zu Grunde. Die Modellierung der klassenbedingten Wahrscheinlichkeit erfolgte unter Verwendung von Mischverteilungen mit 5 Dichten pro Klasse.

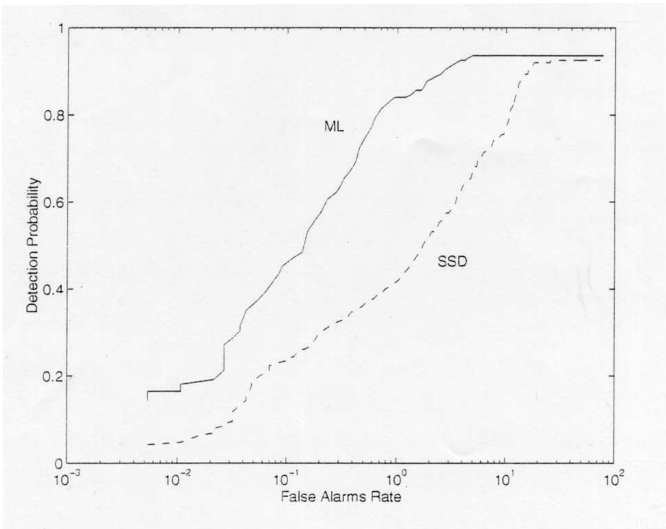

Abbildung 12: Testergebnisse Erkennung von Handgesten

Erstaunlicherweise erreichten beide Klassifikatoren eine etwa gleichwertige Fehlerrate von 10%, wobei der ML-Klassifikator eine geringere Falschalarmrate erzielte (Abbildung 12). Allerdings muß kritisiert werden, daß die vorliegende Falschalarmrate mit knapp unter 10 Falschalarmen pro Bild doch recht hoch ist. Bei einer geringeren Detektionsrate konnte der ML-Klassifikator auf Grund weniger Falschalarme bessere Ergebnisse vorweisen. Dem kritischen Beobachter fällt bei den Beispielbildern des Versuchs in Abbildung 11 auf, daß die Hände, wie zufällig, vor einem schwarzen Kasten liegen. Da bei diesem Experiment gezielt vor gemischtem Hintergrund gelernt wurde, dürfte dies eigentlich nicht der Fall sein.

7 Zusammenfassung

In dieser Seminararbeit wurde der klassische Ansatz bei der Objekterkennung im Allgemeinen und eine Erweiterung des Ansatzes im Besonderen vorgestellt. Zur Klassifikation wurde die Bayes'sche Entscheidungsregel herangezogen. Die Modellierung der Wahrscheinlichkeiten erfolgte mit Hilfe von Gauß'schen Mischverteilungen. Bei der Betrachtung der Parameterschätzung haben wir zudem den EM-Algorithmus und die 'Principal Components Analysis' beschrieben.

Der erweiterte Ansatz von Pentland und Moghaddam wurde durch die zu Recht kritisierten Informationsverluste bei der PCA begründet und versucht im Wesentlichen diese zu kompensieren. So wird, zusätzlich zum Merkmalsraum (F-Raum), auch der sonst wegen seiner Hochdimensionalität ignorierte Nicht-Merkmalsraum ($\bar{F}$-Raum) betrachtet. Für die hierzu notwendigen Berechnungen wurden effiziente Verfahren eingeführt, die den Ansatz praktikabel machen.

Trotz der an sich vernünftigen und in ihrem Rechenansatz gut nachvollziehbaren Idee, entäuscht die 1997 in [1] vorgestellte Arbeit meines Erachtens durch die nicht ausreichende Validierung des tatsächlichen Qualitätsgewinns der Klassifikation, da der ML-Klassifikator leider nicht mit üblichen Klassifikatoren verglichen wurde. Ein Vergleich mit einem DIFS-Klassifikator wäre hier wünschenswert. Darüber hinaus fehlt eine detallierte Beschreibung zur effizienten Berechnung von $||\tilde{x}||^2$ für die klassenspezifischen Wahrscheinlichkeit. Diese erschien uns für das Gesamtverständnis jedoch so wichtig, daß wir hierzu eigene Überlegungen in Kapitel 5.4 ergänzt haben.

Literatur

[1] B. Moghaddam, A. Pentland, *Probabilistic Visual Learning for Object Representation*, IEEE Transactions on pattern analysis and machine intelligence, VOL. 19, NO. 7, JULY 1997.

[2] Niemann H.: *Klassifikation von Mustern*, Springer Verlag, Berlin 1983.

[3] R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[4] J. Dahmen, K. Beulen, M. Güld, H. Ney, "A Mixture Density Based Approach to Object Recognition for Image Retrieval", Proceedings of the 6th International RIAO Conference on Content-Based Multimedia Information Access, Paris, France, pp. 1632-1647 April 2000.

[5] 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), pp. 1-38, 1977.

[6] H. Ney, "Acoustic Moelling 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.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego CA, 1990.