Technical Report

Dept. for Pattern Recognition and Image Processing
Institute for Automation
Technical University of Vienna
Treitlstr. 3/1832
A-1040 Vienna AUSTRIA
Phone: +43 (1) 58801-8161
Fax: +43 (1) 569697
E-mail: khy@prip.tuwien.ac.at

PRIP-TR-30 21. März 1994

Identifizieren von Gesichtern durch Steuerung der visuellen Aufmerksamkeit

Aufgabenstellung der Diplomarbeit

Karin Hraby

Abstract

Ziel der Diplomarbeit ist, ein System zu entwickeln, das die Ansätze der Eigenfaces von Turk und Pentland und des Mustervergleichs von Brunelli und Poggio anhand eines Mechanismus zur Steuerung der visuellen Aufmerksamkeit vereinigt. Dabei soll der Eigenface-Ansatz zu einer groben Vorklassifikation der Gesichter auf einer niedrigen Auflösung dienen. Verschiedene Teile des Gesichts sollen dann in höherer Auflösung betrachtet werden und als zusätzliche Information in die Klassifikation einfließen. Nach einer allgemeinen Einführung in das Themengebiet, werden die vorgeschlagenen Module des Systems dargestellt und ihr Zusammenwirken erläutert.

The goal of the diploma thesis is to develop a system that combines the eigenfaces introduced by Turk and Pentland and the template matching strategy of Brunelli and Poggio by a mechanism for focusing attention on different regions of the face. First the face is roughly classified at a coarse resolution, in a second step more information about the face is gained by focussing at different regions of the face at a higher resolution. The new information has to be included to get a new classification. After a general introduction into the field of face recognition the modules of the system and their interactions are explained.

Inhaltsverzeichnis

1 Einleitung

Eine der wichtigsten und am besten ausgebildeten Fähigkeiten des Menschen ist das Erkennen von Personen. Sie ist für den täglichen Umgang mit einer Vielzahl an verschiedenen Menschen von großer Bedeutung. So lernen zum Beispiel Säuglinge als erstes ihre Mutter von anderen Personen zu unterscheiden. Im Laufe unseres Lebens lernen wir hunderte verschiedene Personen kennen und erkennen und können vertraute Personen auch nach jahrelanger Trennung sofort wiedererkennen. Dabei können wir uns auf diese Fähigkeit so sehr verlassen, daß Verwechslungen nur selten auftreten.

Die wichtigste Informationsquelle für das Erkennen von Personen ist das Gesicht. Das Gesicht ist in seiner Einmaligkeit vergleichbar mit den Fingerabdrücken eines Menschen, wobei wir die Fähigkeit entwickelt haben, Menschen nur auf Grund ihres Gesichts zu unterscheiden. Das Gesicht ist natürlich nicht die einzige Information, die wir beim Erkennen von Personen verwenden; auch Merkmale wie Körperbau, Stimme, Gestik, Kleidung oder die Umgebung, in der wir eine Person treffen, dienen als Hinweise. Ein Beispiel für die zum Erkennen notwendige Umgebung wäre die Person an der Kassa meines Supermarktes, die ich in der Umgebung des Supermarktes problemlos wiedererkenne, auf einer Abendveranstaltung allerdings, also in einem ganz anderen Rahmen und in völlig anderer Kleidung hätte ich vielleicht nur das vage Gefühl, die Person von irgendwo zu kennen.

In sehr vielen Fällen ist aber das Gesicht alleine ausreichend, um eine Person wiederzuerkennen. Im weiteren werde ich mich daher auf Gesichter beschränken, genauer gesagt auf das Identifizieren von Gesichtern. Das **Erkennen** von Gesichtern läßt sich in zwei Teile zerlegen. Der erste Schritt ist das **Finden** eines Gesichts in einer natürlichen Szene. Der zweite Schritt ist dann das **Identifizieren** dieses Gesichts, wobei Identifizieren meist die Zuordnung eines Namens und damit einer Person zu diesem Gesicht bedeutet. Wie das Beispiel des Supermarkts aber zeigt, kann eine Person auch identifiziert werden, ohne ihren Namen zu kennen. Ich will mich der Einfachheit halber aber auf das Zuweisen eines Namens beschränken.

Eine weitere wichtige Fähigkeit des Menschen ist die Selektion von Information. Wir sind durch unsere Umwelt ständig mit einer Fülle an Informationen konfrontiert, wobei wir aber nur einen Teil dieser Informationen über unsere Sinnesorgane aufnehmen und davon wiederum nur einen Teil verarbeiten können. Der Vorgang der selektiven Aufnahme von Information heißt **Aufmerksamkeit** ([WGR80]). Es bedeutet, daß bestimmte Teile der Umgebung größere Bedeutung erhalten als andere. Ein Beispiel für visuelle Aufmerksamkeit ist das Autofahren. Der Verlauf der Fahrbahn ist von größerer Bedeutung als die Bäume am Rand der Straße und wird deshalb genauer beobachtet.

Ich werde in Kapitel 2 einen Überblick über die Bedeutung von Aufmerksamkeit im visuellen System geben. Kapitel 3 beschäftigt sich genauer mit den Fähigkeiten und Mechanismen des Menschen beim Identifizieren von Gesichtern. In Kapitel 4 werde ich einige bereits entwickelte Ansätze und Techniken zum automatischen Erkennen von Gesichtern vorstellen, um in Kapitel 5 ein System vorzustellen, das zwei der vorgestellten Ansätze

unter Einbezug eines Mechanismus zur Steuerung der visuellen Aufmerksamkeit vereinigt. Zuletzt wird in Kapitel 6 die Aufgabenstellung der Diplomarbeit konkret erläutert.

2 Visuelle Aufmerksamkeit

Ein Mechanismus zur Steuerung der visuellen Aufmerksamkeit hat seinen Ursprung in der Anatomie des menschlichen Auges mit seiner ungleichförmigen Verteilung der Photorezeptoren der Retina (für eine genauere Einführung siehe [Mil90]). Der hochauflösende Teil der Retina, die Sehgrube oder Fovea, ist sehr viel dichter mit lichtempfindlichen Zellen ausgestattet als der sie umgebende Rest der Netzhaut. Die Fovea kann daher einfallende Lichtmuster um vieles feiner analysieren als die Peripherie der Netzhaut. Da nur ein Teil der vom Auge wahrgenommenen Szene auf diesen hochauflösenden Teil der Retina fällt, wird durch Bewegung der Augen die Fovea auf die wichtigsten Regionen der Szene gerichtet. Diese Bewegungen der Augen, mit dem Ziel, Reize aus der Peripherie des Auges in die Fovea zu bringen, heißen Sakkaden und diese Art der Aufmerksamkeit wird **externe** Aufmerksamkeit ([Mil90]) genannt.

Die externe Aufmerksamkeit alleine ist allerdings nicht ausreichend, um die von der Retina empfangenen Lichtmuster zu verarbeiten. Eine Einschränkung ergibt sich durch die relative Langsamkeit der Sakkaden. Eine Sakkade dauert 20-60 msec, abhängig von der dabei zurückgelegten Entfernung, und es dauert weitere 200-300 msec bis die Muskeln der Augen für eine neuerliche Sakkade der Augen bereit sind. Einfache Experimente mit komplexen Bildern zeigen aber, daß es auch möglich ist, in einem Zeitintervall, das für eine Bewegung der Augen zu kurz ist, die Aufmerksamkeit auf mehr als ein Objekt eines Bildes zu richten. Diese Art der Aufmerksamkeit wird daher als **interne** Aufmerksamkeit bezeichnet.

Auf ein Computermodell angewandt bedeutet externe Aufmerksamkeit die Steuerung eines aktiven Sensors, wobei eine Verlagerung der Aufmerksamkeit in einem neuen Bild resultiert. Durch die Mechanismen der internen Aufmerksamkeit werden im Gegensatz dazu verschiedene Teile eines Bildes verarbeitet.

3 Identifizieren von Gesichtern durch den Menschen

Das Identifizieren von uns bekannten Gesichtern erfolgt beinahe augenblicklich und ist sehr robust. Wir können in sehr hohem Maß Gesichter unabhängig von den herrschenden Lichtverhältnissen, der Entfernung des Gesichts oder der Lage des Kopfes erkennen. Aber auch Änderungen des Aussehens durch eine neue Frisur, das Tragen einer Brille oder eines Bartes beeinträchtigen unsere Fähigkeiten nur wenig. Sogar Veränderungen durch Alterung stören im allgemeinen die Wiedererkennung nicht.

Im Laufe der letzten Jahrzehnte haben sich viele Wissenschafter aus den unterschiedlichsten Disziplinen mit der Analyse der Fähigkeiten des Menschen beim Erkennen von Gesichtern beschäftigt. Trotzdem ist man von einer Lösung der Aufgabe noch weit ent-

fernt. Ich möchte im folgenden nur einige, mir für das Thema der Diplomarbeit wichtig erschienene, Punkte aus der Fülle an erzielten Erkenntnissen herausgreifen.

Ein Weg, um Einflußfaktoren auf das Identifizieren von Gesichtern zu finden, ist das Messen der Zeitspanne, die für das Identifizieren von Gesichter gebraucht wird. Bruce [Bru88] stellt fest, daß das Identifizieren von um 180 Grad gedrehten Gesichtern signifikant länger dauert, als von Gesichtern in "natürlicher" Kopflage. Ebenso stellte sie fest, daß (von Menschen) als "außergewöhnlich" eingestufte Gesichter schneller identifiziert werden, als "durchschnittlich" eingestufte Gesichter. Abbildung 1 zeigt ein Beispiel dafür, wie die Wahrnehmung eines Gesichts durch Rotation beeinträchtigt wird.

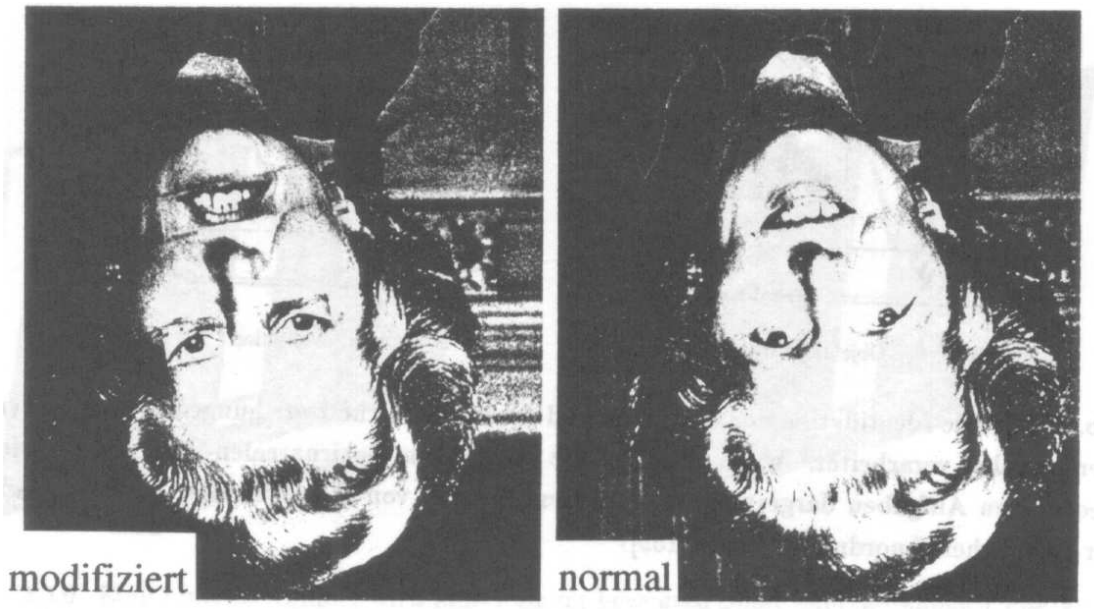

Abbildung 1: Beipiel (aus [Gie93]) für die eingeschränkte Rotationsinvarianz der menschlichen Wahrnehmung. Beide Bilder werden als ähnliche Gesichter wahrgenommen, bei einer Rotation um 180 Grad wird die Täuschung unmittelbar erkennbar.

Eine weitere Möglichkeit festzustellen, welche Faktoren für das Identifizieren von Gesichtern wichtig sind, ist zu untersuchen, welche Merkmale des Gesichts für die Identifikation von Bedeutung sind. Die Merkmale eines Objekts können meist leicht durch Beschreibung des Objekts erhalten werden. Aber so einfach es ist, ein Gesicht zu identifizieren, so schwer

fällt im allgemeinen die eindeutige Beschreibung eines Gesichts. So ist zum Beispiel die in Abbildung 2 unten gezeigte Person trotz der Darstellung aus einem anderen Blickwinkel, der geänderten Kleidung, Brille und anderer Lichtverhältnisse problemlos richtig als eine der beiden oben dargestellten Gesichter erkennbar. Eine eindeutige Unterscheidung der beiden oben dargestellten Gesichter durch Beschreibung der Merkmale ist im Gegensatz dazu sehr schwierig. Merkmale zur Beschreibung eines Gesichts wären die Farbe der Augen, Form und Dicke der Augenbrauen, Dicke der Lippen, Größe des Mundes u.a.m.. Goldstein et al. [GHL71] nehmen auf Grund von Experimenten an, daß die Wichtigkeit der Merkmale von oben nach unten abnimmt. Das heißt, daß bei einem Gesicht ohne herausragende Merkmale Haaransatz und Augen für die Identifikation wichtiger sind als etwa Mund und Kinn. Weiters stellen sie fest, daß die zur Unterscheidung verschiedener Gesichter benötigte Anzahl an unabhängigen Merkmalen logarithmisch mit der Anzahl der Gesichter wächst. Es würden daher zur Unterscheidung von 1000 Gesichtern ca. 7 Merkmale ausreichend sein, zur Unterscheidung von einer Million 12 Merkmale.

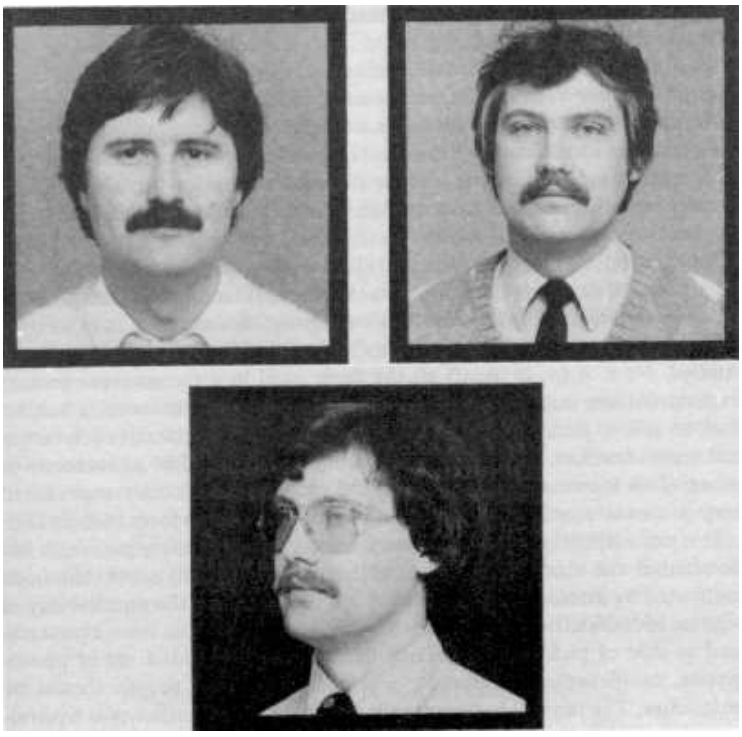

Abbildung 2: Die Identifikation des Mannes in der unteren Reihe als einer der beiden Männer in der ersten Reihe, fällt leichter als eine Unterscheidung der beiden Männer durch Beschreibung der Merkmale (aus [Bru88]).

Ein in den U.S.A. für Augenzeugen eines Verbrechens verwendetes Hilfsmittel ist Identikit bzw. Photokit ([Bru88]). Beide Systeme gehen von der Annahme aus, daß Menschen ein Gesicht in seine Einzelteile wie Augenpartie, Nase, Mund zerlegen kann bzw. eine Gesicht durch Zusammensetzen dieser einzelnen Teile rekonstruieren kann. Identikit und Photokit stellen daher eine Datenbank an unterschiedlichen Augen, Nasen u.s.w. zur Verfügung, die beliebig kombiniert werden können, um ein bestimmtes Gesicht darzustellen. Bei Identikit sind diese Teile gezeichnet, bei Photokit beinhaltet die Datenbank Photos. Als Gründe für die bei der Evaluierung dieser beiden Systeme erzielten unbefriedigenden Ergebnisse gibt Bruce [Bru88] neben der zu geringen Variationsmöglichkeit der einzelnen Bausteine an, daß vor allem die zur Rekonstruktion notwendige Zerlegung des Gesichts kein geeignetes Hilfsmittel darstellt, da Menschen ein Gesicht als Gesamtheit wahrnehmen und nicht als eine Ansammlung von Merkmalen. Das Vorhandensein und die räumliche Relation der Merkmale eines Gesichts scheinen wichtiger zu sein als die Details der Merkmale.

In [Sam91] analysieren die Autoren, welche Auflösung (Anzahl der Pixel und Grauwertstufen) zur Identifikation eines Gesichts minimal notwendig ist und schließen auf Grund von Experimenten, daß eine Auflösung von $32 \times 32 \times 4$ ppb, d.h. eine Array von 32×32 Pixeln mit 4 Bit pro Pixel, zur Identifikation eines Gesichts ausreichend ist. Harmon [Har79] erzielte durch eine geeignete Wahl des Reduktionsfensters bei einer Auflösung von 16×16 Pixel, 16 Grauwerten und eines Testset von 14 verschiedenen Gesichtern einen Identifikationsgrad von 95%. Bemerkenswert ist, daß Tsotsos [Tso90] durch komplexitätstheoretische Überlegungen ca. 30×30 Pixel als die Auflösung erhält, die vom visuellen System auf einmal verarbeitet werden kann.

Experimente aus der Gehirnforschung [WGR80] können Aufschluß über die Regionen des Gesichts geben, die mehr Aufmerksamkeit erhalten, als andere. Aufzeichnungen der Augenbewegungen von Rhesusaffen bei der Betrachtung von Bildern mit Gesichtern von anderen Affen lassen eine Bevorzugung der Augen- und Mundregion erkennen.
Abbildung 3 zeigt das Bild eines Affengesichts und die bei der Betrachtung durch zwei verschiedene Affen aufgezeichneten Augenbewegungen.

4 Automatisches Identifizieren von Gesichtern

Systeme zum automatischen Erkennen von Gesichtern sind in vielen Bereichen (Identifikation von Verbrechern, verschieden Arten von Sicherheitssystemen u.v.m.) einsetzbar. Das Interesse an diesem Gebiet geht bis in das letzte Jahrhundert zurück [Gal88], wobei die ersten Ansätze zur Lösung des Problems mit Hilfe des Computers in die frühen 70er Jahre zurückreichen [FE73]. In den letzten Jahren ist das Interesse an diesem Gebiet wieder gestiegen und es besteht eine starke Zusammenarbeit von Wissenschaftern aus Psychologie, Neurophysiologie und Bildverarbeitung ([EJNY86, YE89]). Einen Überblick über die Arbeiten der letzten Jahren geben die Beiträge von [BB89] und [SI92].

Ich werde im folgenden an Hand von einigen Arbeiten verschiedene Techniken vorstellen, die sich mit dem Identifizieren von frontal aufgenommenen Gesichtern beschäftigen.

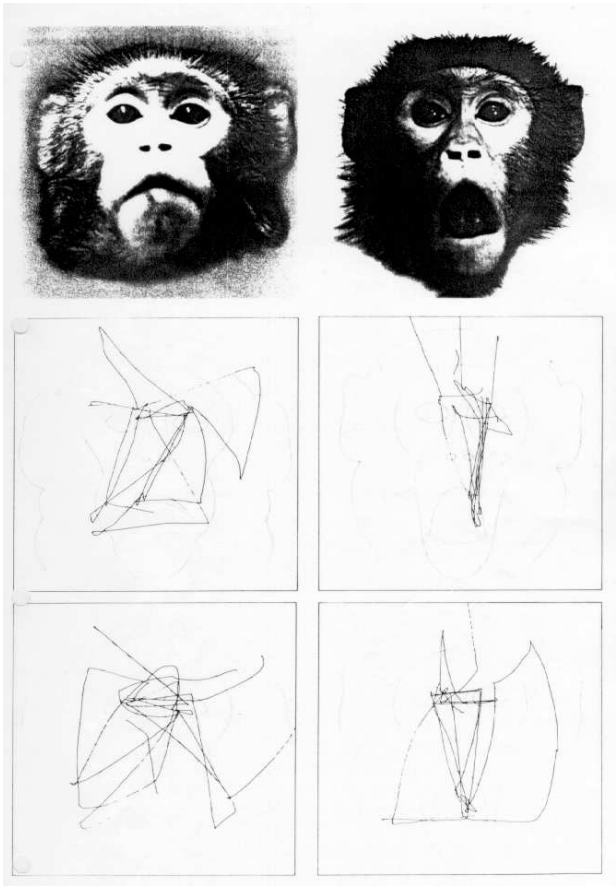

Abbildung 3: Visuelle Aufmerksamkeit. Die beiden Bilder von Affengesichtern in der obersten Reihe wurden verschiedenen Affen gezeigt. Die beiden unteren Reihen zeigen die Aufzeichnung der Augenbewegungen von zwei verschiedenen Affen bei der Betrachtung dieser Bilder (aus [WGR80]) .

4.1 Normalisierung

Um eine relative Unabhängigkeit der Identifikation gegenüber der Größe, Position und Lage des Gesichts zu erhalten, müssen im allgemeinen die zu identifizierenden Gesichter normalisiert werden. Bei den meisten Ansätzen erfolgt die Normalisierung der Größe des Gesichts durch Festsetzen des Augenabstands. Eine Translation wird durch Festlegung der Position der Augen im Bild und eine Rotation durch Waagerechtdrehen der Augenachse ausgeglichen. Da das Vorhandensein und die Position der Augen auch beim Finden von Gesichtern in einer Szene eine große Rolle spielt, ist diese Normalisierung oft das Ergebnis dieses Prozesses.

Ein Beispiel für einen Algorithmus zum Finden von Gesichtern, der als Ergebnis die normalisierten Gesichter liefert, ist der von Chetverikov und Lerch [CL92] beschriebene.

4.2 Geometrische Merkmale

Ein Ansatz zur automatischen Identifikation von Gesichtern basiert auf der Berechnung einer Menge geometrischer Merkmale. Dafür werden aus einem Bild die relative Position und Form von Merkmalen des Gesichts wie Augen, Nase, Mund oder Kinn berechnet. Einer der ersten Versuche zum automatischen Identifizieren von Gesichtern mit Hilfe von geometrischen Eigenschaften stammt von Kanade [Kan73]. In dem System von Kanade repräsentiert ein Vektor von numerischen Daten die Position und Größe der wichtigsten Merkmale eines Gesichts. Das System kann durch automatisches Extrahieren von 13 unabhängigen Merkmalen 75% der Gesichter einer Menge von 20 verschiedenen Personen identifizieren.

Die Arbeiten von Brunelli und Poggio [BP92, BP93] bauen auf dieser Arbeit von Kanade auf. Das für die Normalisierung der Gesichter notwendige Finden der Augen wird durch eine normalisierte Korrelation mit einer in 5 verschiedenen Größen vorliegenden Abbildung der Augen (Template) durchgeführt, wobei der Rechenaufwand durch eine hierarchische Korrelation verringert wird. Die Autoren erreichen dadurch eine Robustheit des Systems gegen Rotationen bis zu 15 Grad. Sind die Augen gefunden und das Gesicht normalisiert, verwenden die Autoren allgemeines Wissen über die Lage der Merkmale eines Gesichts, um diese Merkmale automatisch zu finden.

Die Berechnung der horizontalen und vertikalen Gradienten bringt Information über die Lage verschiedener Merkmale: die vertikalen Gradienten dienen der Lokalisation von Haaransatz, Mund, Nasenende und Augen; die horizontalen Gradienten zeigen den linken und rechten Rand des Gesichts und der Nase. Die exakte Position der Nase ist in der Folge durch eine Spitze in der horizontalen Projektion der vertikalen Gradienten gegeben, der Mund durch ein Tal in der horizontalen Projektion des Grauwertbildes.
Auf diese Weise wird ein Vektor mit den folgenden 35 verschiedenen Merkmalen (siehe auch Abbildung 4) erzeugt:

- Dichte der Augenbrauen;
- Vertikale Position der Augenbrauen an der Stelle des Mittelpunktes der Augen;
- Grobe Beschreibung des linken Augenbrauenbogens durch 11 Werte;
- Vertikale Position der Nase und Breite der Nase;
- Vertikale Position, Weite und Dicke der Ober- und Unterlippe des Mundes;
- 11 Radien, die die Form des Kinns beschreiben;
- Breite des Gesichts auf der Höhe der Nase und in der Mitte zwischen Nasenspitze und Augen.

Das Erkennen eines Gesichts aus diesen Daten erfolgt über einen Bayes-Klassifizierer.

Bei der Untersuchung der Robustheit der Merkmale stellte sich heraus, daß mit steigender Anzahl der zu unterscheidenden Klassen die Leistung des Systems sinkt und die Varianz

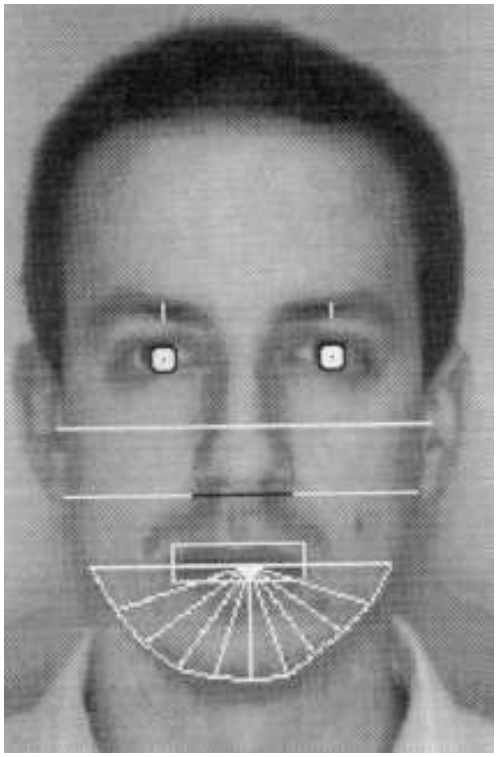

Abbildung 4: Geometrische Merkmale, aus [BP93]

zwischen den Klassen von der gleichen Größenordnung ist, wie die Varianz innerhalb der Klassen.

4.3 Mustervergleich

Ein weiterer Ansatz basiert auf einem Mustervergleich. Im einfachsten Fall wird das zu identifizierende Gesicht (ein zweidimensionales Array von Grauwerten) unter Verwendung einer entsprechenden Metrik mit verschiedenen Vorlagen (Templates) verglichen. Um Robustheit für diese Technik zu erreichen, werden verschiedene Erweiterungen verwendet, wie z.B. Vorverarbeitungen des Bildes, Verwendung von mehreren Templates für jedes Gesicht (z.B. verschiedenen Kopflagen, Auflösungen, etc.) oder verformbare Templates.

Eine erste Arbeit, in der Gesichter durch Vergleich verschiedener Teile des Gesichts identifiziert werden, stammt von Baron [Bar81]. Baron verwendet Frontalbilder mit einer Auflösung von 128×120 Pixel und verschiedenem Hintergrund. In den Bildern werden zuerst die Augen mit Hilfe einer Serie von Augen-Templates in verschiedenen Größen und Kreuz-Korrelation gesucht. Die Normalisierung der Gesichter erfolgt durch Setzen des Augenabstands auf einen festgelegten Wert. Die Speicherung der Gesichter erfolgt durch die folgenden fünf verschiedenen Regionen: gesamtes Gesicht, Augenpaare, Mund, Kinn und Haare. Diese zur Unterscheidung dienenden Regionen werden "händisch" ausgewählt

und auf eine Größe von 15×16 Pixel reduziert. Dadurch ergeben sich für verschiedene Teile des Gesichts verschiedene Auflösungsstufen, eine niedrige Auflösung für das ganze Gesicht und eine höhere Auflösung für die Einzelteile.

Die einzelnen Templates für die Regionen des Gesichts werden durch Mittelung über verschiedenen Aufnahmen des Gesichts einer Person erzeugt. Um ein gewisses Maß an Robustheit zu erreichen, können pro Gesichtsteil bis zu fünf verschiedene Templates gespeichert werden. Zur Identifikation werden ebenfalls zuerst die Augen gesucht, das Gesicht normalisiert und das Bild auf eine Auflösung von 15×16 Pixel reduziert. In der Folge wird zuerst das gesamte Gesicht mit den Gesichtern der Datenbank verglichen. Nur bei den Gesichtern, die über einem bestimmten Korrelationswert liegen, werden auch noch die anderen Teile des Gesichts miteinander verglichen. Um das Gesicht als eine bestimmte Person zu identifizieren, müssen 3/4 der gespeicherten Templates übereinstimmen. Baron konnte mit diesem Ansatz alle 42 gelernten Gesichter aus einer Menge von 150 Testbildern identifizieren und alle unbekannten Gesichter richtig ablehnen. Er berichtet, daß eine Vergrößerung der Auflösung der Templates die Identifikation nicht verbesserte. Das System ist nicht robust gegenüber starker Variation des Lichts, es kann Gesichter nur bis zu einer Rotation von 20 Grad identifizieren und ist nicht robust gegenüber größeren Änderungen der Größe des Gesichts.

Aufbauend auf dem Ansatz von Baron wird in dem von Brunelli und Poggio [BP93] vorgestellten Ansatz ein Gesicht durch 4 verschiedenen Regionen (Augenpartie, Nase, Mund und den Bereich von den Augenbrauen bis zum Kinn) repräsentiert. Diese Teile des Gesichts werden automatisch und relativ zur Position der Augen gefunden. Um ein Gesicht zu identifizieren, werden alle Regionen des Gesichts mit den in der Datenbank gespeicherten Personen verglichen (Kreuzkorrelation). Dieser Vergleich liefert einen Vektor der erzielten Übereinstimmungen der Regionen und das Gesicht mit der höchsten Gesamtübereinstimmung ist das Erkannte. Die Größe der Templates beträgt für alle Regionen 36×36 Pixeln. Untersuchungen über den Beitrag der einzelnen Merkmale zur Unterscheidung ergaben folgende Reihenfolge der Wichtigkeit: Augen, Nase, Mund, Augenbrauen-Kinn-Bereich.

Da die errechnete Korrelation auf Beleuchtungsänderungen empfindlich ist, wurden verschiedene Arten der Vorverarbeitung des Bildes miteinander verglichen: keine Vorverarbeitung; Normalisierung der Grauwerte durch Festlegung des Verhältnisses eines Grauwertes gegenüber der durchschnittlichen Helligkeit seiner Nachbarschaft; Bestimmung des Intensitätsgradienten des Gauß-gefilterten Bildes; Laplace des Grauwertbildes. Die besten Ergebnisse wurden mit Gradientenbildern erzielt. Eine Variation der Größe der Templates erbrachte gute Ergebnisse bei einer Größe der Templates von 36×36 Pixeln.

4.4 Eigenfaces

Einen informationstheoretischen Ansatz zur Identifikation von Gesichtern haben Turk & Pentland [TP91] gewählt. Relevante Information wird automatisch aus den Gesichtern extrahiert, möglichst effizient kodiert und zur Unterscheidung der verschiedenen Gesichter verwendet. Die extrahierte Information kann den vorhin beschriebenen Merkmalen des

Gesichts, wie Augen, Nase oder Mund entsprechen, muß aber nicht.

Die von Turk & Pentland verwendete Methode, um die in einem Bild (eines Gesichts) vorhandene Information zu extrahieren, ist eine Hauptkomponentenanalyse der Verteilung einer Menge verschiedener (Gesichts-)Bilder. Ein Bild eines Gesichts wird dabei als Punkt in einem sehr hochdimensionalen Raum gesehen. Ein Bild von der Größe $N \times N$ Pixel entspricht somit einem Vektor mit N^2 Komponenten und stellt einen Punkt im N^2-dimensionalen Raum dar. Die verschiedenen Gesichter sind allerdings auf Grund von vielen Gemeinsamkeiten nicht gleichförmig in diesem Raum verteilt. Die durch eine Hauptkomponentenanalyse erhaltenen Eigenvektoren sind die Vektoren, die diese Verteilung im Raum am besten darstellen.

Zur Berechnung der Eigenfaces werden aus einer Menge von Gesichtsbildern $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$,..., $\mathbf{\Gamma}_M$ das Durchschnittsgesicht $\mathbf{\Psi} = \frac{1}{M}\sum_{n=1}^{M}\mathbf{\Gamma}_n$ und die "Differenzgesichter" $\mathbf{\Phi}_i = \mathbf{\Gamma}_i - \mathbf{\Psi}$ erzeugt. Diese Differenzgesichter bilden die Ausgangsbasis für die Hauptkomponentenanalyse, bei der eine Menge von M orthonormalen Vektoren $\mathbf{u}_n$ gesucht wird, die die Verteilung der Daten (linear) beschreibt. Hierbei wird der k-te Vektor $\mathbf{u}_k$ so gewählt, daß die Varianz

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(\mathbf{u_k^T}\mathbf{\Phi}_n)^2 \tag{1}$$

maximal ist, unter der Bedingung

$$\mathbf{u}_l^T\mathbf{u}_k = \delta_{lk} = \begin{cases} 1 & : \quad l = k \\ 0 & : \quad sonst \end{cases} \tag{2}$$

Die Vektoren $\mathbf{u}_k$ und die Skalare λ_k sind die Eigenvektoren und deren Eigenwerte der Kovarianzmatrix[1] C.

$$C = \frac{1}{M}\sum_{n=1}^{M}\mathbf{\Phi}_n\mathbf{\Phi}_n^T \tag{3}$$

Die Darstellung dieser N^2-dimensionalen Eigenvektoren als Bild bezeichnen Turk & Pentland als **"Eigenfaces"** oder "Eigengesichter". Jedes dieser zur Hauptkomponentenanalyse verwendeten Gesichter $\mathbf{\Gamma}_i$ kann dann als Linearkombination der Eigenvektoren (oder Eigenfaces) exakt dargestellt werden. Da aber meist eine exakte Darstellung nicht notwendig ist, werden nur die M' "besten" Eigenfaces verwendet, d.h. die Vektoren, die am meisten zur Unterscheidung in der Menge der Gesichter beitragen, das sind die Eigenvektoren mit den größten Eigenwerten. Der von den M' besten Eigenfaces aufgespannte M'-dimensionale Unterraum heißt dann **"Facespace"** oder "Gesichterraum". Turk und Pentland verwendeten nach der Hauptkomponentenanalyse von 16 Gesichtern der Größe 512×512 Pixel zum Identifizieren dieser Gesichter die 7 besten Eigenfaces (M' = 7). Abbildung 5 zeigt einige durch Hauptkomponentenanalyse erhaltene Eigenfaces.

Ein Gesicht $\mathbf{\Gamma}$ wird durch $\omega_k = \mathbf{u}_k^T(\mathbf{\Gamma} - \mathbf{\Psi})$ für k = 1,..., M' in seine Eigenface-Teile transformiert (d.h. in den Face-Space projiziert). Der Vektor $\mathbf{\Omega}^T = [\omega_1\omega_2...\omega_{M'}]$ der Gewichte zur

[1] Da die $\mathbf{\Phi}_i$ den Mittelwert 0 haben, ist die Korrelationsmatrix gleich der Kovarianzmatrix.

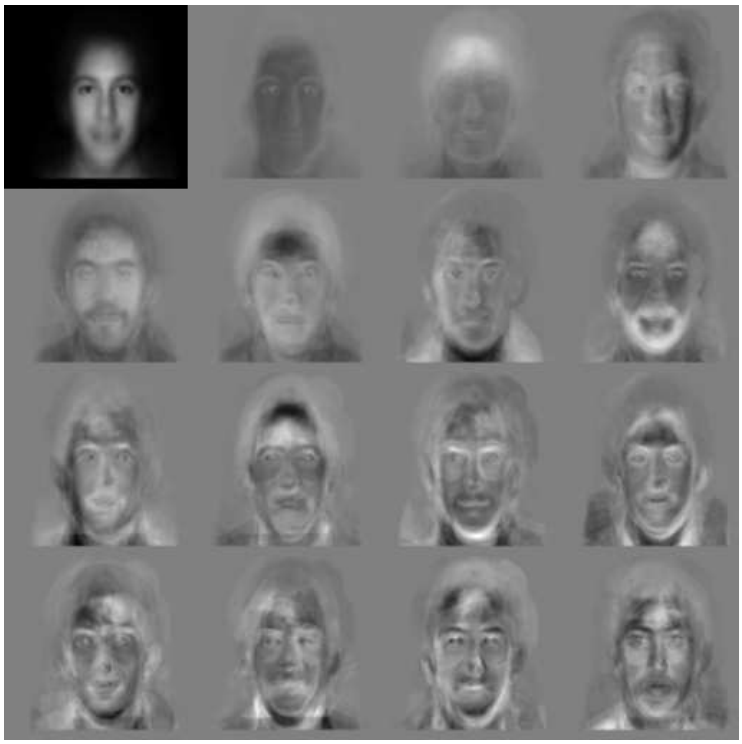

Abbildung 5: Eigenfaces

Darstellung eines Gesichtes als Linearkombination der Eigenfaces ist somit die gewünschte effiziente Darstellung der Information eines Gesichts und wird zur Klassifikation der Gesichter verwendet. Bei der Klassifikation wird die euklidische Distanz E dieses Vektors zu den Vektoren aller anderen gespeicherten Gesichtern berechnet. Liegt die minimale Distanz E_{min} unter einem bestimmten Schwellwert S, so wird das Gesicht dieser Klasse zu gewiesen. Andernfalls wird das Gesicht als unbekannt zurückgewiesen.

Für die durchgeführten Experimente verwendeten die Autoren eine Datenbank mit 2500 Bildern. Dafür wurden 16 Personen in je 3 verschiedenen Kopflagen, Kopfgrößen und Lichtverhältnissen aufgenommen und auf diesen 512×512 Pixel großen Bildern eine Gauß-Pyramide in 6 Ebenen konstruiert. Für verschiedene Tests wurden mehrmals Gruppen von 16 Bildern ausgewählt. Eine Gruppe enthält ein Bild pro Person mit der gleiche Kopflage, Kopfgröße und Lichtverhältnissen. Bei diesen Tests wurde die Robustheit des Systems gegenüber der Variation der einzelnen Darstellungsformen getestet. Als besonders sensibel stellte sich das System gegenüber Änderungen der Orientierung und Größe des Kopfes heraus. Ein weiterer wichtiger Parameter ist der Schwellwert S, der bestimmt, wie ähnlich sich zwei Gesichter sein müssen, um in eine Klasse zu fallen. Ist dieser Schwellwert unendlich (das heißt, das Gesicht wird auf alle Fälle der Klasse mit dem kleinsten E-Wert zugewiesen), so konnten im Durchschnitt 96% der Gesichter bei Änderung der Lichtverhältnisse, 85% bei Änderung der Lage des Kopfes und 64% bei Änderung der Größe richtig identifiziert

werden. Durch Senkung des Schwellwertes S steigt der Prozentsatz an richtigen Klassifikationen auf 100%, 94% bzw. 74%, wobei dann allerdings 20% der eigentlich bekannten Gesichter als unbekannt abgelehnt werden.

5 Identifizierung durch Aufmerksamkeit

Den Ausgangspunkt für meine Diplomarbeit bildet der in Kapitel 4.4 erläuterte Ansatz der Eigenfaces. Das in der Folge vorgestellte System zur Identifikation von Gesichtern stellt eine Verbindung dieses Ansatzes mit dem Ansatz von Brunelli & Poggio aus Kapitel 4.3 dar, die zur Identifikation verschiedene Teile des Gesichts in unterschiedlichen Auflösungen verwenden. Das System setzt sich aus folgenden Teilen zusammen:

1. Auflösungsreduktion
2. Eigenface-Projektion
3. Integration & Klassifikation
4. Sakkaden-Generation

Im folgenden werde ich die Aufgaben der einzelnen Teile des Systems vorstellen, um im Anschluß daran das Zusammenwirken der Teile und die weiteren Ziele der Diplomarbeit genauer zu erläutern.

5.1 Auflösungsreduktion

Ausgehend von der in Kapitel 3 dargelegten Annahme, daß zur Identifikation eines Gesichts eine Auflösung in der Größenordnung von 32×32 Pixel ausreichend ist, wird in diesem Modul die Auflösung der Gesichtsbilder (bzw. der Ausschnitte) mit Hilfe einer Bildpyramide auf 32×32 Pixel reduziert. Um nur einen Ausschnitt des Originalbildes in derselben Weise zu verarbeiten, verwenden wir eine Erweiterung der Bildpyramide um sogenannte Shifter Circuits [OA92]. Dieses Modell erlaubt es, sowohl einen Ausschnitt als auch das gesamte Bild auf eine entsprechende Pyramidenebene zu projizieren. Das "Routing" der Daten in der Pyramide übernehmen sogenannte "Control Units", die vom Sakkaden-Generator gesteuert werden.

5.2 Eigenface-Projektion

Dieses Modul entspricht der von Turk & Pentland vorgeschlagenen Klassifikation mit Hilfe einer Hauptkomponentenanalyse.

Ein Klasse von neuralen Netzwerken, die eine Hauptkomponentenanalyse durchführen kann, sind die autoassoziativen Netzwerke. Ein autoassoziatives Netzwerk besteht aus N Input- und N Output-Units (N ist die Anzahl der Pixel des Bildes) und $M << N$ Hidden-Units. Dieses vollverbundene 3-Layer-feed-forward Netz wird meist mit einem

Back-Propagation-Algorithmus so trainiert, daß das Inputmuster am Output reproduziert wird und der Fehler

$$E = \sum_{i=1}^{L} (I_i - O_i)^2 \tag{4}$$

minimiert wird (L ist die Anzahl der zum Trainieren des Netzes verwendeten Gesichtsbilder). Die Aktivierung der Units des Output-Layers ist gegeben durch

$$O_i = \mathbf{W}_2 \mathbf{W}_1 I_i \tag{5}$$

wobei die Matrizen $\mathbf{W}_1 \in \mathbb{R}_{M \times N}$ und $\mathbf{W}_2 \in \mathbb{R}_{N \times M}$ die Gewichte zwischen Input- und Hidden-Layer bzw. Hidden- und Output-Layer darstellen.

Boulard & Kamp [BK88] haben für diese Art von Netzen gezeigt, daß die Hidden-Units den Input auf den von den ersten M Hauptkomponenten der Verteilung der Inputmuster aufgespannten Raum projizieren. Dabei entsprechen die gelernten Gewichte zwischen Input-Layer und Hidden-Layer den Eigenfaces, die Aktivierung des Output-Layers der Projektion des Bildes am Input-Layer auf den nieder-dimensionalen Facespace und die Aktivierung des Hidden-Units dem Beitrag der einzelnen Eigenfaces zur Darstellung des Bildes am Input-Layer. Die Aktivierung der Hidden-Units ist somit die gewünschte effiziente Darstellung der relevanten Information eines Gesichts und bildet den Input für das Integrations-Modul.

5.3 Integration & Klassifikation

Die Aufgabe dieses Moduls ist es, die durch das Fokussieren auf Details des Gesichts neu gewonnene Information in die bereits vorhandene Information zu integrieren und durch eine Klassifikation der Input-Daten das Gesicht zu identifizeren.

Eine Integration von zeitlich aufeinander folgender Information kann durch ein rekurrentes Netzwerk, wie es Jordan und Elman [Jor86, Elm88] vorgestellt haben, erreicht werden. Ein derartiges Netzwerk entspricht einem Assoziationsnetzwerk, das um einen Layer sogenannter "State-Units" erweitert ist. Diese State-Units erhalten ihren Input sowohl von den Hidden-Units als auch von sich selbst und zeichnen somit den Zustand des Netzwerkes auf. Die "State-Units" können als statische Assoziation mit der gesamten Mustersequenz interpretieren werden, das entspricht damit der gewünschten Integration der Information der verschiedenen Gesichtsteile.

Das Netzwerk kann trainiert werden, Sequenzen von Mustern zu erkennen und den Input zu klassifizieren. In unserem Fall wäre es das Erkennen der vom Sakkaden-Generator erzeugten Folge (ganzes Gesicht, Augen, Nase und Mund) in der Darstellungen des Facespace und die Zuweisung eines Namens zu dem Gesicht.

Die Wahl des Schwellwertes S der Klassifikation bestimmt dabei, wie ähnlich sich die Repräsentationen von zwei Gesichtern sein müssen, um als gleich erkannt zu werden. Wie in Kapitel 4.4 erläutert, bewirkt eine Variation des Schwellwertes entweder eine Erhöhung des Prozentsatzes an falschen Klassifikationen oder des Prozentsatzes an als unbekannt abgelehnten Gesichtern. Für die Erweiterung des System wird dieser Schwellwert so gesetzt, daß die Anzahl an falschen Klassifikationen niedrig ist.

5.4 Sakkaden-Generation

Der Sakkaden-Generator ist jenes Modul, in dem das Fokussieren auf Details des Gesichts erfolgt. Dafür werden Merkmale des Gesichts ausgewählt, die für die weitere Unterscheidung zwischen den verschiedenen Gesichtern wichtig sind. Die Auswahl dieser Merkmale kann dabei entweder datengetrieben oder fix erfolgen. In einem ersten Ansatz wird eine im vorhinein festgelegte Reihenfolge von Regionen bestimmt. Das entspricht dem Ansatz von Brunelli & Poggio, die ebenfalls Augen, Nase und Mund zur Unterscheidung verwenden und den in Kapitel 3 beschriebenen Aufzeichnungen der Augenbewegungen bei Affen. Da als Ergebnis des Findens von Gesichtern in einer Szene die Augen bereits lokalisiert und die Gesichter normalisiert sind, können diese Regionen auf Grund von allgemeinem Wissen über den Aufbau eines Gesichts leicht automatisch gefunden werden.

5.5 Ablauf-Schema

Aus diesen vier Modulen ergibt sich das in Abbildung 6 dargestellte System. Die Auflösung des Originalbildes wird mit der Hilfe einer Bildpyramide auf eine Größe von 32×32 Pixel reduziert. Nach der Projektion des Bildes auf den Facespace dienen die Aktivierungen der Hidden-Units dem Integrations-Modul zur Klassifikation des Gesichts. Ist das Gesicht keinem der gespeicherten Gesichter ähnlich genug (d.h. die Ähnlichkeit ist kleiner als ein festgelegter Schwellwert S), wird die Aufmerksamkeit auf einen Teil des Gesichts fokussiert. Die Reihenfolgen der Regionen, auf die die Aufmerksamkeit gelenkt wird, sind Augenpartie, Nase und Mund. Wobei diese Gesichtsteile durch das Routing ebenfalls eine Auflösung von 32×32 Pixel haben und ebenfalls in den Facespace projiziert werden. Die dadurch zusätzlich erhaltene Information wird in die bereits vorhandene Information integriert (Integrations-Modul) und gleichzeitig eine neuerliche Klassifikation durchgeführt. Dieser Vorgang wird so lange wiederholt, bis das Gesicht identifiziert werden kann oder keine weitere Information verfügbar ist und das Gesicht somit als unbekannt abgelehnt wird.

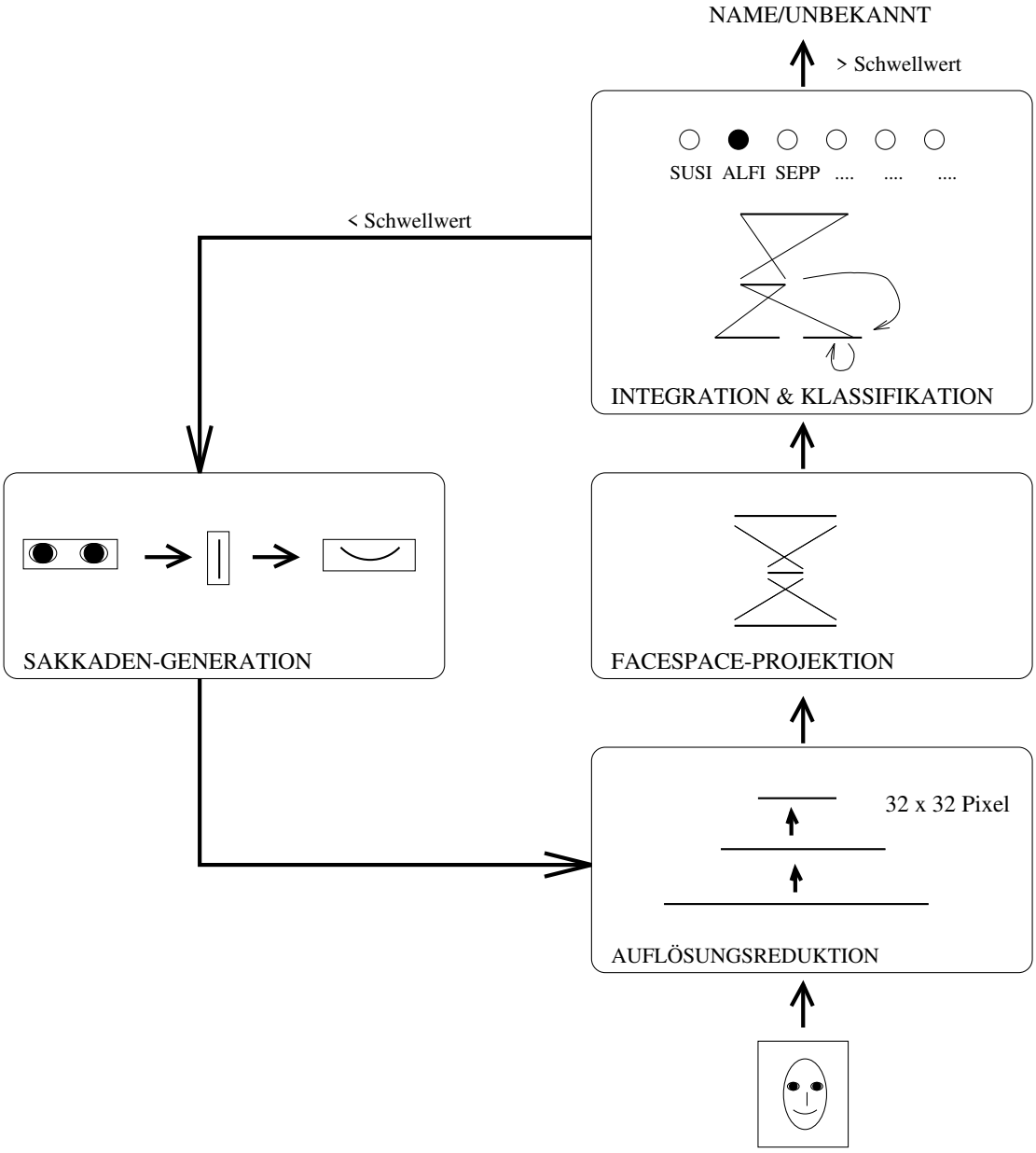

Abbildung 6: Identifikation durch Steuerung der Aufmerksamkeit

6 Aufgabenstellung und Ziele

Ich habe zunächst die Fähigkeiten und Arbeitsweisen des Menschen beim Erkennen von Gesichtern, sowie den Begriff der visuellen Aufmerksamkeit erläutert. Im weiteren wurden einige Ansätze zur automatischen Identifikation von Gesichtern aus der Literatur vorgestellt. Auf Grund von Schwachstellen dieser Techniken stellt eine Verbindung der beiden Ansätzen aus den Kapiteln 4.3 und 4.4, erweitert um einen Mechanismus zur Steuerung der Aufmerksamkeit auf Teile des Gesichts, einen kognitiv plausiblen und effizienten Weg zur Lösung einiger Probleme dar. In einem ersten Schritt habe ich ein System vorgestellt, das diese Verbindung der Ansätze herstellt. Es ist bisher ein theoretisches Modell, das erst in der Praxis getestet und konkretisiert werden muß.

Ich werde an Hand der einzelnen Module die weiteren Schritte der Diplomarbeit näher erläutern.

Auflösungsreduktion Für die Auflösungsreduktion soll eine Gauß-Pyramide verwendet werden, die Software für die Erweiterung der Bildpyramide um die sogenannten Shifter Circuits (wie in [OA92]) steht zur Verfügung und muß den gegebenen Anforderungen angepaßt werden.

Eigenface-Projektion Die von Turk & Pentland entwickelte Sofware zur Hauptkomponentenanalyse steht ebenfalls zur Verfügung.

Integration & Klassifikation Die Tauglichkeit der vorgeschlagenen Netzwerkstruktur zur Integration der Information aus den Gesichtsteilen muß überprüft werden (es ist ein Praktikum dafür ausgeschrieben). Gegebenenfalls müssen andere Netzwerkstrukturen zur Integration von zeitlich aufeinander folgender Information untersucht werden.

Sakkaden-Generation Eines der wichtigsten Ziele der weiteren Arbeit ist es, einen Mechanismus zur Steuerung der Aufmerksamkeit zu entwickeln, der (ähnlich der Hauptkomponentenanalyse) die zur weiteren Unterscheidung wichtigen Regionen des Gesichts automatisch d.h. datengetrieben findet. Er soll die explizite Steuerung der Aufmerksamkeit des in dieser Arbeit vorgestellten Ansatzes ersetzen.

Ablauf-Schema Eine enge Verbindung der Module untereinander soll bewirken, daß das System als eine Einheit trainiert werden kann, wobei als "Teaching Input" nur die Klassifikation des Gesichts (d.h. Zuweisung eines Namens) dienen soll. Zur Evaluierung der Ergebnisse dient die von Turk & Pentland verwendete Datenbank an Gesichtern.

Literatur

[Bar81] R.J. Baron. Mechanisms of human facial recognition. *Int. J. Man Machine Studies*, 15:137–178, 1981.

[BB89] V. Bruce and M. Burt. *Computer recognition of faces*, pages 487–506. In Young and Ellis [YE89], 1989.

[BK88] H. Bourlard and Y. Kamp. Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biological Cybernetics*, 59:291–294, 1988.

[BP92] R. Brunelli and T. Poggio. Face recognition through geometrical features. In *European Conference on Computer Vision*, pages 792–800, 1992.

[BP93] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, October 1993.

[Bru88] Vicky Bruce. *Recognising Faces.* Lawrence Erlbaum Associates, Publishers, 1988.

[CL92] D. Chetverikov and A. Lerch. Multiresolution face detection. In *Theobild*, pages 131–140, 1992.

[EJNY86] H.D. Ellis, M.A. Jeeves, F. Newcombe, and A. Young, editors. *Aspects of Face Processing.* Martinus Nijhoff Publishers, 1986. Proc. of the NATO ARW, Aberdeen, 1985.

[Elm88] J.L. Elman. Finding structure in time. Technical Report 8801, UCSD, CRL, 1988.

[FE73] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, (22):7–93, 1973.

[Gal88] Sir F. Galton. Personal identification and description. *Nature*, pages 173–177, June 1888.

[GHL71] A. Goldstein, L. Harmon, and B. Lesk. Identification of human faces. *Proceedings of the IEEE*, 59:748–760, 1971.

[Gie93] G.-J. Giefing. *Foveales Bildverarbeitungssystem zur verhaltensorientierten Szenenanalyse.* Number 245 in Reihe 10: Informatik/Kommunikationstechnik. VDI Verlag, Düsseldorf, 1993.

[Har79] L. D. Harmon. The recognition of faces. *Scientific American*, (229):71–82, October 1979.

[Jor86] M.I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proc. of 8th Annual Conf. of the Cognitive Science Society*, pages 531–546, 1986.

[Kan73] T. Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Kyoto Univ., Dept. Inform. Sci., 1973.

[Mil90] R. Milanese. Focus of attention in human vision: a survey. Technical Report 90.03, University of Geneva, August 1990.

[OA92] B. Olshausen and D. Anderson, C. Van Essen. A Neural model of visual Attention and invariant pattern recognition. Technical Report CNS Memo 18, Caltech Institute of Technology, Computational and Neural Systems Program, 1992.

[Sam91] A. Samal. Minimum resolution for human face detection and identification. In *SPIE/SPSE Symp. Electronic Imaging*, January 1991.

[SI92] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expressions. *Pattern Recognition*, 25(1):65–77, 1992.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Tso90] J.K. Tsotsos. Analyzing Vision at the Complexity Level. *Behaviral and Brain Sciences*, 13(3):423–469, 1990.

[WGR80] R.H. Wurtz, M.E. Goldberg, and D.L. Robinson. Behavioral modulation of visual responses in the monkey: stimulus selection for attention and movement. *Progress in Psychobiology and Physiological Psychology 9*, pages 43–83, 1980.

[YE89] A.W. Young and H.D. Ellis, editors. *Handbook of Research on Face Processing.* Elsevier Science Publishers B.V. (North-Holland), 1989.