Gesichtserkennung mit elastischer Graphenanpassung

Diplomarbeit von
Florian Hardt

26. Januar 2001

Hauptberichter : Prof. Dr. G. Wunner
Mitberichter : Dr. habil. S. Luding

Institut für Theoretische Physik I
Universität Stuttgart
Pfaffenwaldring 57, 70550 Stuttgart

The eyes are not responsible when the mind does the seeing.

-Publilius Syrus

Inhaltsverzeichnis

1 Automatische Bilderkennung

1.1 Wahrnehmungsleistung von Mensch und Maschine

Wir Menschen erkennen uns vertraute Objekte in unserer Umwelt mit an Selbstverständlichkeit grenzender Leichtigkeit. Die Buchstaben, die Sie gerade lesen, werden mühelos interpretiert, die Abbildungen in dieser Arbeit "mit einem Blick" erkannt.

Doch während Computer bei analytischen Aufgaben oder gemäßigt abstrakten wie Schachspiel menschliche Leistungen[1] übertreffen, hinken sie in der Bilderkennung weit hinterher. Dies wird zuweilen als Anzeichen dafür gesehen, daß der Maschine ein gewisser (schwer oder nicht beschreibbarer) menschlicher Faktor fehle. Übersehen wird dabei, daß die allein in jeder Sekunde vom menschlichen Sehsystem zu bewältigende Datenmenge immens ist. Auf der Netzhaut befinden sich grob 10^6 "Pixel", und an jedem wägen im Zehntelsekundentakt Dutzende von Neuronen die Hypothese ab, daß dort und dann eine bewegliche oder starre Grenzlinie zu sehen ist. Die zehn Millionen Neuronen des visuellen Kortex elaborieren diese Ergebnisse und schätzen in jedem Moment die mögliche Lage im Raum und die Farbe an allen Bildpunkten ein. Es bedürfte etwa einer Millionen derzeitiger PCs um allein diesen Teil des menschlichen Gehirns zu simulieren [Mor88,99].

Bei dieser Abschätzung erfüllt jeder PC die Aufgaben von ein paar hundert Neuronen. Will man die Gehirnprozesse schärfer abbilden (z.B. ein PC pro Neuron), steigt der Rechenbedarf, und die Menge der möglichen Realisationen der Simulation nimmt ab, da dann keine effizienten Algorithmen zur Beschreibung von Neuronengruppen gefunden werden müssen. Andersherum kann man versuchen, global effiziente Algorithmen zu finden und damit die Gehirnfunktionen in sehr viel größerem Maßstab abzubilden. Die Konstruktion eines solchen Systems ist weitaus schwieriger, doch der Rechenbedarf nimmt ab.

Beide Ansätze sind sowohl von theoretischer als auch praktischer Bedeutung für die Biologie, die Robotik und die Entwicklung künstlicher Intelligenz.

[1] Es kann spekuliert werden, ob abstraktes Denken "schwer" ist, weil die dafür verantwortlichen Gehirnfunktionen evolutionsgeschichtlich jung sind, während der Wahrnehmung unserer Umwelt alte und daher optimierte Gehirnfunktionen zugrunde liegen.

1.2 Gesichtserkennung

Diese Arbeit beschäftigt sich mit einem speziellen Bereich der Bilderkennung, der Gesichtsidentifikation. Dabei soll nicht eine Objektklasse "Gesicht" erkannt, sondern die Identität einer Person durch Vergleich mit bekannten Gesichtern in einer Datenbank ("Gallerie") festgestellt werden. Wenn im folgenden auch nachlässig von Gesichtserkennung gesprochen wird, so ist doch die Gesichtsidentifikation gemeint.

Praktische Anwendungen sind z.B. Zugangssicherung von Gebäuden oder Authentifikation eines Benutzers an einem Bankschalter. Gegenüber anderen biometrischen Verfahren (z.B. Identifikation anhand eines Fingerabdrucks oder der Iris) ist wenig oder keine aktive Mitwirkung der zu erkennenden Person notwendig. Dies kann bei den oben genannten Anwendungen zu einer höheren Akzeptanz von seiten der Benutzer führen, ermöglicht aber auch den Einsatz zur Personenfahndung.

Da Menschen einander vorrangig anhand des Gesichts erkennen sind die Ergebnisse des Systems leicht verifizierbar. Zudem können Rückschlüsse auf die menschliche Gesichtserkennung gezogen werden: Falls die vom Computer als ähnlich bewerteten Gesichter auch Menschen als ähnlich erscheinen, so liegt die Vermutung nahe, daß beide Erkennsysteme dieselben Merkmale zur Identifikation benutzen.

Schließlich dient der Gesichtsausdruck auch zur menschlichen Kommunikation: Ein System zur "Mimik-Erkennung" könnte die Interaktion von Mensch und Maschine verbessern.

1.3 Problematik der Gesichtserkennung

Objekterkennung mit dem Computer ist ein schwieriges und rechenzeitintensives Unterfangen. Variationen in Größe, Beleuchtung und Perspektive müssen ausgeglichen werden, zudem muß das Objekt vom Bildhintergrund getrennt werden. Bei der Gesichtserkennung kommen weitere Probleme hinzu:

- Gesichter sind eine Klasse ähnlicher Objekte.
- Gesichter sind nicht rigide (Mimik).
- Für die meisten sinnvollen Anwendungen müssen Veränderungen wie Frisur, Bartwuchs oder Brillen berücksichtigt werden.

In dieser Arbeit werden vor allem diese drei speziellen Punkte behandelt, während durch die Verwendung normierter Gesichter die Probleme mit Beleuchtung, Größe und Hintergrund weitgehend wegfallen.

1.4 Das menschliche Wahrnehmungssystem

Das Studium des menschlichen Gesichtswahrnehmungssystems ist ein interdisziplinäres Forschungsgebiet, in dem Psychologen, Neurobiologen und Informatiker vor allem an folgenden Kernfragen arbeiten [Ger97]:

- Handelt es sich bei Gesichtserkennung um eine erlernte oder angeborene Fähigkeit?
- Handelt es sich um einen spezialisierten Prozeß oder ist er im wesentlichen identisch mit der allgemeinen visuellen Wahrnehmung?
- Ist der Prozeß featureorientiert (in einzelne Schritte untergliedert) oder holistisch?
- An welchen Merkmalen wird ein Gesicht erkannt?
- Wie wird ein Gesicht intern repräsentiert und wie erfolgt der Vergleich?

Psychologische Experimente haben gezeigt, daß Gesichtswahrnehmung eine Kombination von erlerntem und angeborenem Können ist. Säuglinge haben bereits mit der Geburt eine interne Gesichtsrepräsentation, die sich im Laufe der Entwicklung verbessert.

Kopfüber präsentierte Gesichter sind schwerer zu erkennen, was darauf hindeutet, daß zugunsten der Effizienz hochspezialisierte statt flexibler (allgemeiner) Prozesse verwendet werden. Zudem können Kinder leichter durch Verkleidungen (Hut, Brille, eine vertraute Person mimt den Weihnachstmann) getäuscht werden, da ihre Gesichtserkennung stärker an einzelnen Merkmalen (z.B. buschige Augenbrauen) orientiert ist [Dia77].

Untersuchungen von Prosopagnosia-Patienten[2] haben gezeigt, daß diese zwar keine Gesichter identifizieren, wohl aber deren Mimik erkennen können [Ley79]. Dies gibt Anlaß zu der Vermutung, daß Identität und Gesichtsausdruck unabhängig voneinander verarbeitet werden.

Tachioskopische Halbfeld-Präsentationen[3] lassen eindeutig erkennen, daß sich beide Hirnhälften bei der Gesichtserkennung komplementär ergänzen, somit also spezifische Prozesse eingesetzt werden, allerdings ist offen, ob dieselben Prozesse nicht auch für andere visuelle Aufgaben verwendet werden [Hay82].

Die seit den 80er Jahren gewonnenen neurophysiologischen Erkenntnisse zeigen klar, daß ein Gesicht zunächst in seine Raumfrequenzen zerlegt wird. Die

[2] Prosopagnosia ist eine hirnorganische Störung im rechten Temporallappen, die den betroffenen Personen die Erkennung von Gesichtern unmöglich macht.

[3] Tachioskopische Halbfeld-Präsentationen werden eingesetzt, um die Aufgabenteilung von linker und rechter Gehirnhemisphäre zu untersuchen. Dabei wird der visuelle Stimulus jeweils nur auf ein Auge gerichtet. Dieser Reiz wird dann vorwiegend in der dem Auge diagonal gegenüberliegenden Hirnhälfte verarbeitet.

tiefpaßgefilterten Daten ermöglichen es, die grobe Zusammensetzung des Gesichts zu erkennen, zeitlich versetzt wird das Bild von den höheren Frequenzen ergänzt. Im temporalen Kortex existiert eine spezielle Zellhierarchie, die nur auf Gesichter reagiert, wobei von den untersten Schichten (empfindlich für die Anwesenheit bestimmter Gesichtsteile wie Augen oder Mund) über die mittleren (deren Zellen die Konfiguration überprüfen) zu den oberen Schichten hin (empfindlich für festgelegte Ansichten wie frontal oder von hinten) generalisiert wird. Ein Teil dieser Zellen reagiert am stärksten auf ein bestimmtes bekanntes Gesicht.

1.5 Überblick

Diese Arbeit beschreibt die Konstruktion eines Systems zur Gesichtserkennung. In diesem Kapitel wurde die allgemeine Problematik erläutert und ein Einblick in das menschliche Gesichtswahrnehmungssystem gegeben.

Kapitel 2 befaßt sich mit der internen Repräsentation von Bildern in der Form etikettierter Graphen als Ausgangspunkt für ein effizientes Erkennsystem. Es wird gezeigt, wie Ergebnisse neuronaler Untersuchungen des Gehirns in das Verfahren Eingang finden.

Die vollständige Beschreibung und Begründung des Verfahrens zur Gesichtserkennung mit Hilfe elastischer Graphen folgt in Kapitel 3.

Im 4. Kapitel werden die Erkennleistungen des Erkennsystems auf der ORL-Gesichtsdatenbank unter verschiedenen Versuchsbedingungen diskutiert, Kapitel 5 beschreibt die Resultate auf der UMIST-Datenbank.

Eine Zusammenfassung der Ergebnisse, ein Vergleich mit anderen Erkennsystemen und eine Erläuterung möglicher Verbesserungen dieses Systems folgt im abschließenden 6. Kapitel.

2 Repräsentation von Bildern

Dieses Kapitel beschäftigt sich mit der Verwaltung der in einem Bild enthaltenen Information. Es wird versucht, die ersten Schritte der Datenverarbeitung im Gehirn zu imitieren. Die in späteren Kapiteln eingeschlagene Vorgehensweise löst sich zwar von der Analogie zum menschlichen Sehsystem, doch erscheint es sinnvoll, sich zumindest ansatzweise der gleichen Eingangsdaten zu bedienen.

2.1 Die Datenstruktur

Grauwertbilder werden in der Bildverarbeitung typischerweise als n-dimensionale Vektoren gespeichert, wobei n die Anzahl der Bildpunkte ist und die Koeffizienten die Grauwerte beschreiben. Zur Bilderkennung würde es prinzipiell genügen, den ein unbekanntes Bild beschreibenden Vektor durch einen geeigneten Algorithmus mit einer Gallerie von Vektoren bekannter Identität zu vergleichen. Werden zwei Vektoren von diesem Algorithmus als hinreichend ähnlich bewertet, so ist das Bild "erkannt" worden.

Nun verhält es sich so, daß die gewählte Datenstruktur für die Konstruktion eines Vergleichs-Algorithmus von entscheidender Bedeutung ist. Eine vektorielle Bildrepräsentation, die z.B. für Darstellung am Bildschirm geeignet ist, enthält keine (direkte) Information über die räumliche Beziehung zweier Bildpunkte. Zudem ist der Informationsgehalt eines einzelnen Pixels (ein oder zwei Byte) gering, so daß sie zu Gruppen zusammengefaßt werden müssen, um beim Vergleich mit einem anderen Bild Mehrdeutigkeiten zu vermeiden.

2.2 Etikettierte Graphen

In dieser Arbeit wird das von Wiskott et.al. [Wis95] vorgeschlagene Konzept etikettierter Graphen verwendet. Diese bestehen aus einem (im Sinne der Graphentheorie) nichtorientierten Graphen, dessen Knoten relevanten Punkten in einem Bild entsprechen: Eine feste Anzahl von Punkten wird an ausgewählten Pixeln plaziert und bildet einen "Graphen", dessen Kanten dem Pixelabstand zweier Knoten entsprechen. Lokale Bildinformationen in der Umgebung jedes Knoten werden diesem zugeordnet und führen zum Konzept des etikettierten Graphen.

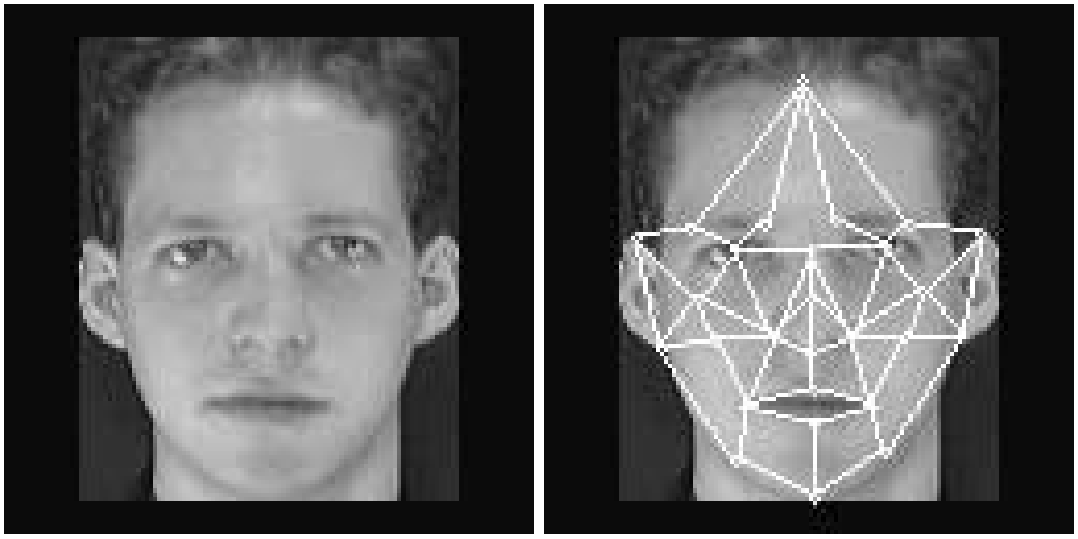

Abb. 2.1: Etikettierter Graph zur Beschreibung eines Bildes. An jedem Knoten ist lokale Bildinformatione kodiert.

Die Ähnlichkeit zweier Graphen hängt dann von der Ähnlichkeit und Position (Verzerrung der Kanten) ihrer Knoten ab.

2.3 Bildtransformation mit Gabor Wavelets

Die lokalen Bildinformationen, mit denen jeder Knoten etikettiert wird, ergeben sich aus einer Wavelettransformation des Bildes. Die komplexen Koeffizienten der Transformation werden in Form eines Jet genannten Vektors angeordnet. Dies soll im folgenden näher erläutert werden.

2.3.1 Wavelettransformationen

Ist $\psi(x)$ eine quadratintegrable Funktion im $\mathbb{R}^n$ und $\mathbb{G}$ eine Gruppe von Abbildungen von $\mathbb{R}^n$ auf sich selbst, so beschreibt die Menge aller Funktionen $\{\psi(g(\vec{x})) \mid g \in \mathbb{G}\}$ eine Wavelettransformation (vergleiche z.B. Würtz [Wür94]):

$$(W(f(\vec{x}))(g(\vec{x})) :=< \psi(g(\vec{x})) \mid f(\vec{x}) >= \int \overline{\psi}(g(\vec{x}))f(\vec{x})d^n\vec{x}. \qquad (2.1)$$

Dabei wird $\psi(\vec{x})$ als Mutterwavelet bezeichnet. Die Forderung nach endlicher Norm (Quadratintegrabilität) ist für eine allgemeine Wavelettransformation nicht nötig, aber für die Bildverarbeitung nützlich: Dadurch ergibt sich sofort, daß die Funktionen im Phasenraum eng lokalisiert sind. Eine Einführung in Wavelettransformationen findet man bei [Chu92].

Handelt es sich bei $\mathbb{G}$ um die Gruppe aller Translationen

$$\vec{x} \longrightarrow \vec{x'} = \vec{x} - \vec{a}, \tag{2.2}$$

so entspricht die Transformation einer Faltung:

$$\begin{aligned} W(f(\vec{x}))(g(Translation\ um\ Vektor\ \vec{a})) &= < \psi(\vec{x} - \vec{a}) \mid f(\vec{x}) > & (2.3)\\ &= \int \overline{\psi}(\vec{x} - \vec{a}) f(\vec{x}) d^n\vec{x} & (2.4)\\ &= (\psi * f)(\vec{a}). & (2.5) \end{aligned}$$

In (2.5) wurde $\psi(\vec{x}) = \overline{\psi}(-\vec{x})$ benutzt, was für später in dieser Arbeit verwendete ψ gilt. Der Vorteil dieser Darstellung liegt darin, daß gemäß dem Faltungstheorem eine Faltung

$$(\psi * f)(\vec{a}) = F^{-1}(F(\psi)F(f))(\vec{a}) \tag{2.6}$$

über Fouriertransformationen F berechnet werden kann. Mit Hilfe der diskreten Fast-Fourier-Transformation (DFFT) kann so die Wavelettransformation rasch numerisch ausgewertet werden. In dieser Arbeit wurde die von Press [Pre92] beschriebene Routine implementiert.

2.3.2 Gabortransformationen

Betrachtet werden Funktionen aus $\mathfrak{L}^2(\mathbb{R}^2)$ im Phasenraum, wobei die Transformation zwischen Orts- und Frequenzraum durch die Fouriertransformation gegeben ist. Gabor [Gab46] zeigte für den eindimensionalen Fall, daß keine Funktion in beiden Räumen zugleich beliebig scharf lokalisiert sein kann. Vielmehr gilt für das Phasenraumvolumen die Unschärferelation (Würtz [Wür94])

$$V(f) \geq 2^{-d}. \tag{2.7}$$

Funktionen mit minimalem Phasenraumvolumen heißen Gaborfunktionen. Als Gabortransformationen werden in dieser Arbeit Wavelettransformationen bezeichnet, deren Mutterwavelet eine Gaborfunktion ist. Im folgenden wird die von Würtz [Wür94] beschriebene zulässige Gaborfunktion verwendet. Sie hat die Form einer ebenen Welle unter einer einhüllenden Gaußglocke:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right)\left[exp(i\vec{k}\vec{x}) - exp(-\sigma^2/2)\right]. \tag{2.8}$$

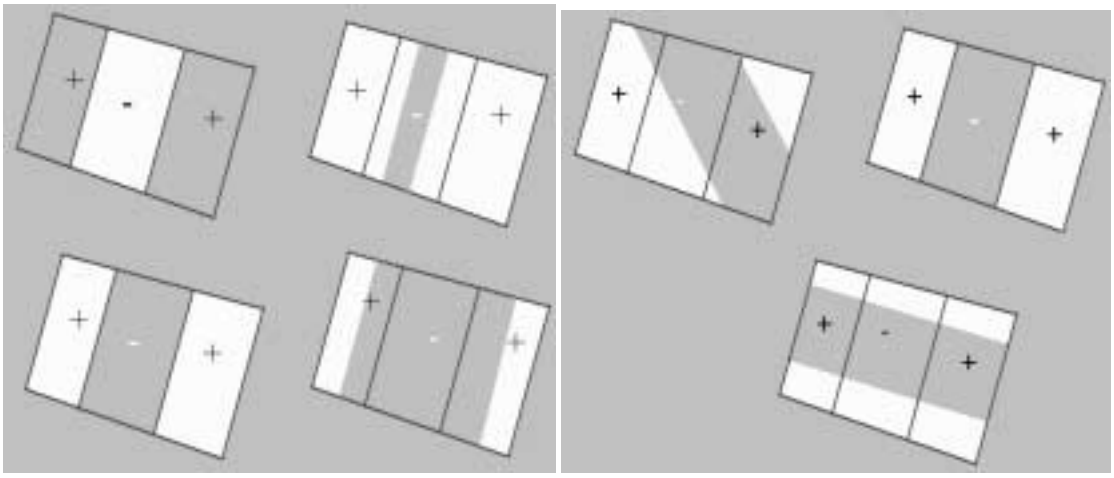

Abb. 2.2: Das "receptive field" einfacher Zellen: Die hellen Bereiche der Zelle werden angeregt. Je nach Vorzeichen betroffenen Gebiets wirkt der Stimulus exhibitorisch (+) oder inhibitorisch (-) auf die Aktivität der Zelle. Das linke Teilbild zeigt die Frequenzabhängigkeit der Aktivität (max. links unten), das rechte die Richtungsabhängigkeit (max. rechts oben).

Dieses spezielle Mutterwavelet wird mit neurologischen Erkenntnissen motiviert: Im primären visuellen Cortex von Säugetieren findet sich eine Vielzahl von Nervenzellen mit unterschiedlichen Eigenschaften. Die einfachsten wurden von Hubel und Wiesel [HuW62] prosaisch "simple cells" (einfache Zellen) genannt. Das Antwortverhalten dieser Zellen ist folgendermaßen charakterisiert:

- Linearität: Die Aktivität der Zelle ist in etwa linear zur äußeren Anregung.
- Lokalisierung im Ortsraum: Jede Zelle ist mit einem korrespondierenden Bereich ("receptive field") auf der Netzhaut verbunden. Reize außerhalb dieses Bereichs beeinflussen die Zelle nicht.
- Lokalisierung im Frequenzraum: Unter Frequenz ist hier die räumliche Variation der Lichtintensität zu verstehen. Werden einfache Zellen mit einer Intensitätsverteilung in Form einer ebenen Welle angeregt, so zeigen sie maximale Aktivität bei einer bestimmten Wellenrichtung und Frequenz. Mit zunehmender Abweichung von diesen Vorgaben fällt die Aktivität ab. Zudem verhalten sich die Zellen indifferent gegenüber einer räumlich konstanten Beleuchtung.

Einzelne Neuronen haben stets eine positive Aktivität – auch eine inaktive Zelle produziert ein Rauschen. Faßt man vier Nervenzellen zusammen, so kann man ihre Aktivitäten mit einer komplexwertigen Funktion beschreiben. Die positiven und negativen Werte des Real- und Imaginärteils werden dabei von je zwei

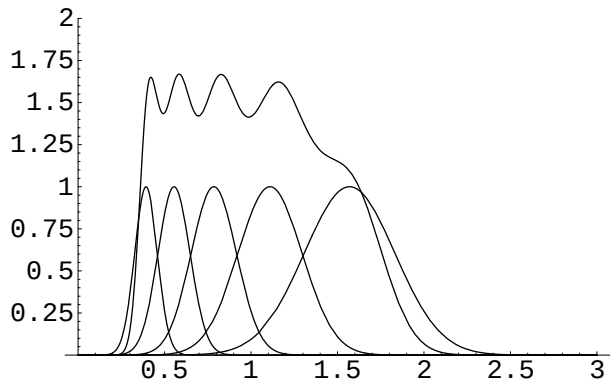

Abb. 2.3: Schnitt durch fünf Gaborkerne im Frequenzraum mit $\sigma = 6$ und Zentralfrequenzen von $k_0 = 0.39$, $k_1 = 0.56$, $k_2 = 0.79$, $k_3 = 1.11$ und $k_4 = 1.57$. Die obere Kurve ergibt sich aus der quadrierten Summe der Kerne. Verwendet man unendlich viele Frequenzen, so ergibt dies eine Konstante ([Wür94]).

Zellen mit stimulierender bzw. hemmender Wirkung auf nachfolgende Neuronen repräsentiert.

Die Gaborfunktion (2.8) beschreibt die Antwort einer solchen Gruppe. Der von Würtz eingeführte zweite Term macht den Gaborkern zulässig, d.h.

$$\int \psi(\vec{x}) d^2x = 0. \tag{2.9}$$

Dies bedeutet, daß die Filterfunktion nicht auf räumlich konstante Signale reagiert, bzw. daß die Fouriertransformierte

$$\left(F\psi_{\vec{k}}\right)(\vec{w}) = exp\left(-\frac{\sigma^2(\vec{w}-\vec{k})^2}{2\vec{k}^2}\right) - exp\left(-\frac{\sigma^2(\vec{w}+\vec{k})^2}{2\vec{k}^2}\right) \tag{2.10}$$

für die Frequenz $\vec{w} = 0$ gleich Null ist.

Von Field [Fie87] wurde gezeigt, daß die Amplitude von fouriertransformierten "natürlichen" Bildern in etwa wie $\frac{1}{|w|}$ abfällt. Der Normierungsfaktor von $\psi(\vec{x})$ wurde so gewählt, daß diese Abhängigkeit gerade kompensiert wird. Dadurch erhält man bei der Wavelttransformation für verschiedene Werte von $\vec{k}$ vergleichbare Antworten. Die Breite der Gaborkerne im Fourierraum wird damit zu $\frac{k}{\sigma}$ (vgl. auch Abbildung 2.3).

2.4 Jets

Die durch die Gabortransformation gewonnenen Koeffizienten werden zu sogenannten Jets zusammengefaßt. Biologisches Vorbild sind die Hypersäulen (Hypercolumns).

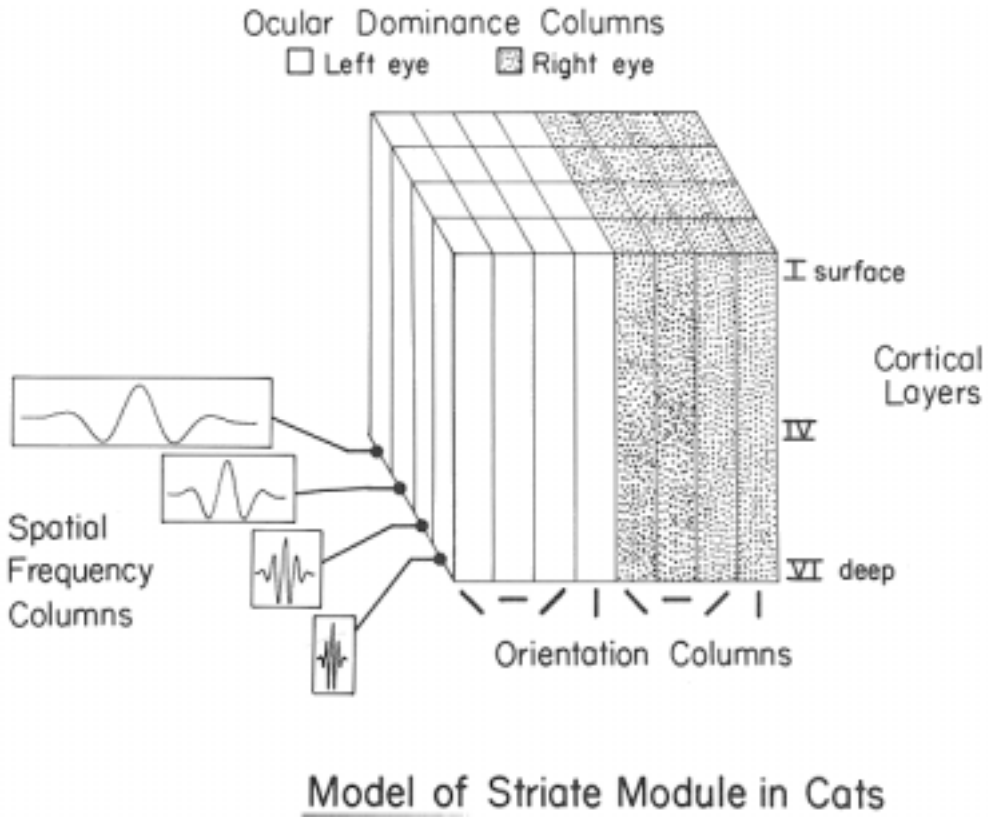

Abb. 2.4: Die Neuronen im V1 sprechen auf feste Orientierungen und Frequenzen an. Alle zu einem bestimmten Punkt auf der Netzhaut gehörenden Neuronen werden zu Hypersäulen zusammengefaßt. (Aus: http://mulan.eng.hmc.edu/~rwang)

2.4.1 Hypersäulen

Die Sehnervenfasern führen vom Auge über den Corpus geniculatum zum primären visuellen Kortex (area striata, Sehrinde, Area 17, V1-Areal). Die Verarbeitung erfolgt in sogenannten corticalen Säulen (Hypercolumns, Hypersäulen) von Neuronen. Je die Hälfte der Neuronen einer Hypersäule empfängt Reize vom linken bzw. rechten Auge. Die Neuronen jeder Hälfte sind in Säulen angeordnet, die jeweils auf den Stimulus einer bestimmten Stelle der Netzhaut reagieren. Innerhalb einer Säule sind Nervenzellen, die auf bestimmte Frequenzen und Orientierungen (orientation columns) reagieren (siehe Abbildung 2.4). Beschränkt man sich auf ein einzelnes Auge (Kameralinse, 2D Photo), so kann diese Struktur als Jet dargestellt werden.

2.4.2 Jets

Ein Jet ist ein n-dimensionaler komplexwertiger Vektor, der einen kleinen Bereich von Grauwerten um einen gegebenen Bildpunkt beschreibt. Die einzelnen Koeffizienten werden dabei an einem gegebenen Bildpunkt $\vec{x}$ aus einer Wavelettransformation des Bildes $I(\vec{x'})$

$$I_j(\vec{x}) = \int I(\vec{x'})\psi_j(\vec{x} - \vec{x'})d^2\vec{x'} \tag{2.11}$$

mit einer Familie von Gaborfunktionen (2.8) mit unterschiedlichen Wellenvektoren $\vec{k}_j$ gewonnen. Der Wellenvektor legt dabei fest, welche Orientierung und Frequenz des Signals am stärksten in den Koeffizienten eingeht. Dies ist am anschaulichsten zu verstehen, wenn man die Wavelettransformation als Faltung betrachtet:

$$I_j(\vec{x}) = F^{-1}(F(\psi)F(I(\vec{x'})))(\vec{x}) \tag{2.12}$$

Im Frequenzraum ist jeder Gaborkern stark um seine Zentralfrequenz $\vec{k}$ lokalisiert. Die Koeffizienten $\hat{I}$ der Fouriertransformation des Bildes $I(\vec{x'})$ werden so angeordnet, daß der Punkt $\hat{I}_{\vec{k}=0}$ (Mittelwert des Grauwertbildes) in die Bildmitte von $\hat{I}$ zu liegen kommt. Für jeden Wert $\vec{k}_j$ wird nun $\hat{I}$ mit dem entsprechenden Gaborkern $\hat{\psi}_k = F(\psi_k)$ punktweise multipliziert und das Resultat zurücktransformiert. Ordnet man nun diese wavelettransformierten Bilder stapelartig übereinander an, so entsprechen die Werte der übereinander liegenden Pixel gerade den Koeffizienten des an dieser Position lokalisierten Jets. Üblicherweise werden die komplexen Koeffizienten als Amplitude und Phase gespeichert. Der Betrag eines Jets wird auf eins normiert, womit sie robust gegen variierenden Kontrast werden. Abbildung (2.5) zeigt das Resultat.

2.4.3 Parameter

Die Breite eines einzelnen Gaborkerns wird durch $\frac{k}{\sigma}$ bestimmt. Je schärfer die Kerne im Fourierraum lokalisiert sind, desto mehr Orientierungen müssen gesampelt werden, um den Fourierraum sinnvoll abzudecken. Gewählt wurde wie bei Wiskott [Wis95]

$$\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu cos\varphi_\mu \\ k_\nu sin\varphi_\mu \end{pmatrix}, \qquad k_\nu = 2^{-\frac{\nu+2}{2}}\pi, \qquad \varphi_\mu = \mu\frac{\pi}{8}$$

mit 5 Frequenzen ($\nu = 0, ..., 4$) und 8 Orientierungen ($\mu = 0, ..., 7$), wobei $j = \mu + 8\nu$ ist (Abbildung 2.6). Dies ergibt an jedem Bildpunkt einen 40-dimensionalen

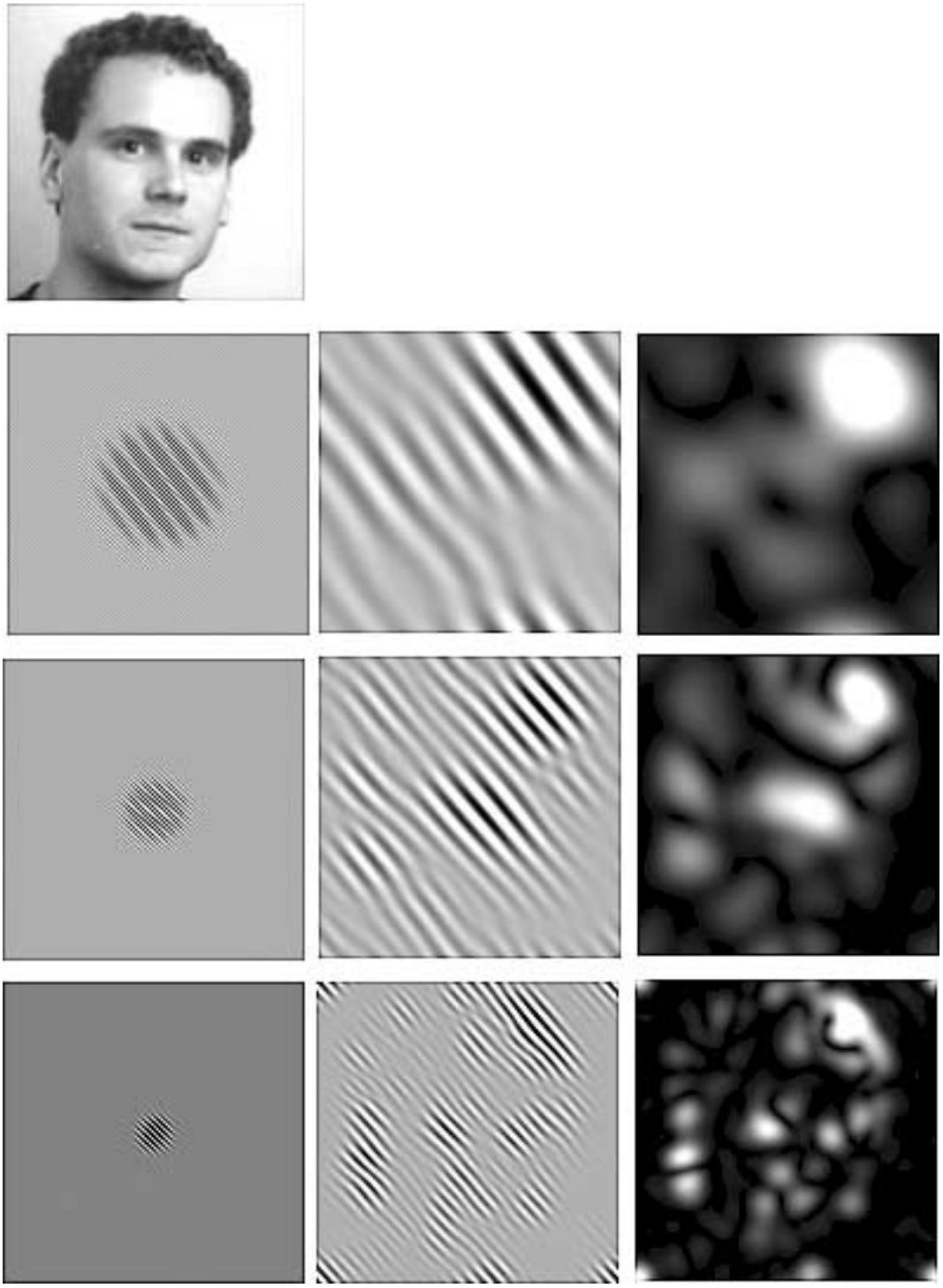

Abb. 2.5: Veranschaulichung der Gabortransformation: Die linke Spalte zeigt den Realteil von drei Gaborfunktionen gleicher Orientierung und unterschiedlicher Frequenz. Wird das obenstehende Gesicht mit ihnen gemäß (2.11) transformiert, so resultiert ein komplexwertiges Bild. Die mittlere Spalte zeigt dessen Realteil (der Imaginärteil hat identische Struktur), in der rechten Spalte ist die Amplitude dargestellt. Die Werte des Realteils oszillieren näherungsweise im Takt und in Richtung von $\vec{k}$, wobei sich die Amplitude nur langsam verändert.

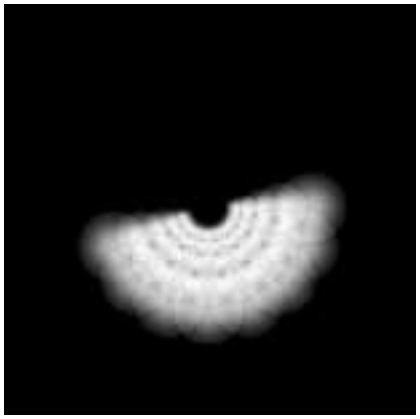

Abb. 2.6: Superposition aller 40 verwendeten Gaborkerne im Fourierraum. Da das komplexe Spektrum reeller Grauwertbilder bereits durch die Werte eines Halbraums vollständig beschrieben ist, genügt es die untere Hälfte abzutasten.

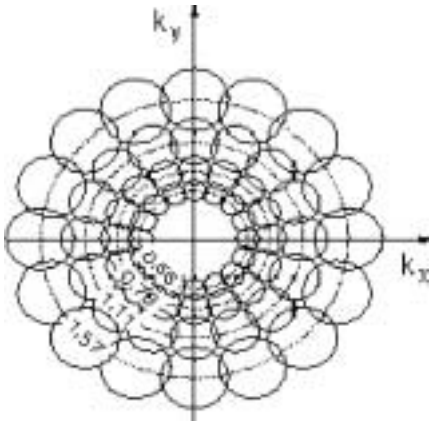

Abb. 2.7: Die Gaborkerne mit Konturlinien bei 37% ihres Maximums. Es werden 8 Richtungen zwischen 0° und 157,5° verwendet. Kerne mit der kürzesten Zentralfrequenz von 0,39 sind nicht dargestellt (Aus: www.neuroinformatik.ruhr-uni-bochum.de).

Jet mit Koeffizienten $J_j = a_j \exp(i\phi_j)$ wobei $(j = 1, ..., 40)$. Der Wert von σ wird durch die Zahl der Orientierungen bestimmt: Für $\sigma = 6 \approx 2\pi$ nimmt die Breite der Kerne mit wachsendem k geeignet zu, so daß der Fourierraum vollständig abgedeckt wird. Die Grauwertbilder werden stets auf einem Feld von 128x128 Pixeln transformiert, so daß ein FFT-Algorithmus eingesetzt werden kann. Aufgrund der Normierung ist $\sum_j a_j = 1$.

2.4.4 Rekonstruktion

Bei der Gabortransformation wird die Datenmenge vergrößert: statt NxN Pixeln, deren Grauwert durch je ein Byte beschrieben ist, liegen nun an jedem Punkt 40 komplexe Koeffizienten (mit je 2 Byte für Real- und Imaginärteil) vor, was die Datenmenge um den Faktor 160 vergrößert. Gleichzeitig verringert sich die in den Jetbildern enthaltene Informationsmenge, da die Rekonstruktion das Originalbild nur approximiert (siehe Abbildung (2.8)). Die Rekonstruktion erfolgt gemäß Würtz [Wür94] mit

Abb. 2.8: Links das Originalbild, rechts die Rekonstruktion des Bildes aus der Gabortransformation. Der "Heiligenschein" ist für die Rekonstruktionen typisch.

$$I(\vec{x}) = F^{-1}\sum_{\vec{k}} F(W(\vec{x},\vec{k})F(\vec{\psi}). \tag{2.13}$$

Für die in dieser Arbeit gewählten Parameter ist die Gesamtrekonstruktion dem Originalbild sehr ähnlich, wenn man von der Verschiebung des absoluten Grauwerts absieht (der Mittelwert wurde durch die Wahl der Gaborwavelets gezielt entfernt).

Aus einem Jetbild werden nur die Jets einiger weniger Punkte zu einem Graph zusammengefaßt, was die Datenmenge wieder erheblich reduziert. Demnach ist auch der Speicherbedarf für eine Gallerie von Graphen vergleichsweise gering. Allerdings ist eine Rekonstruktion des Bildes aus dem Graph nicht sinnvoll möglich.

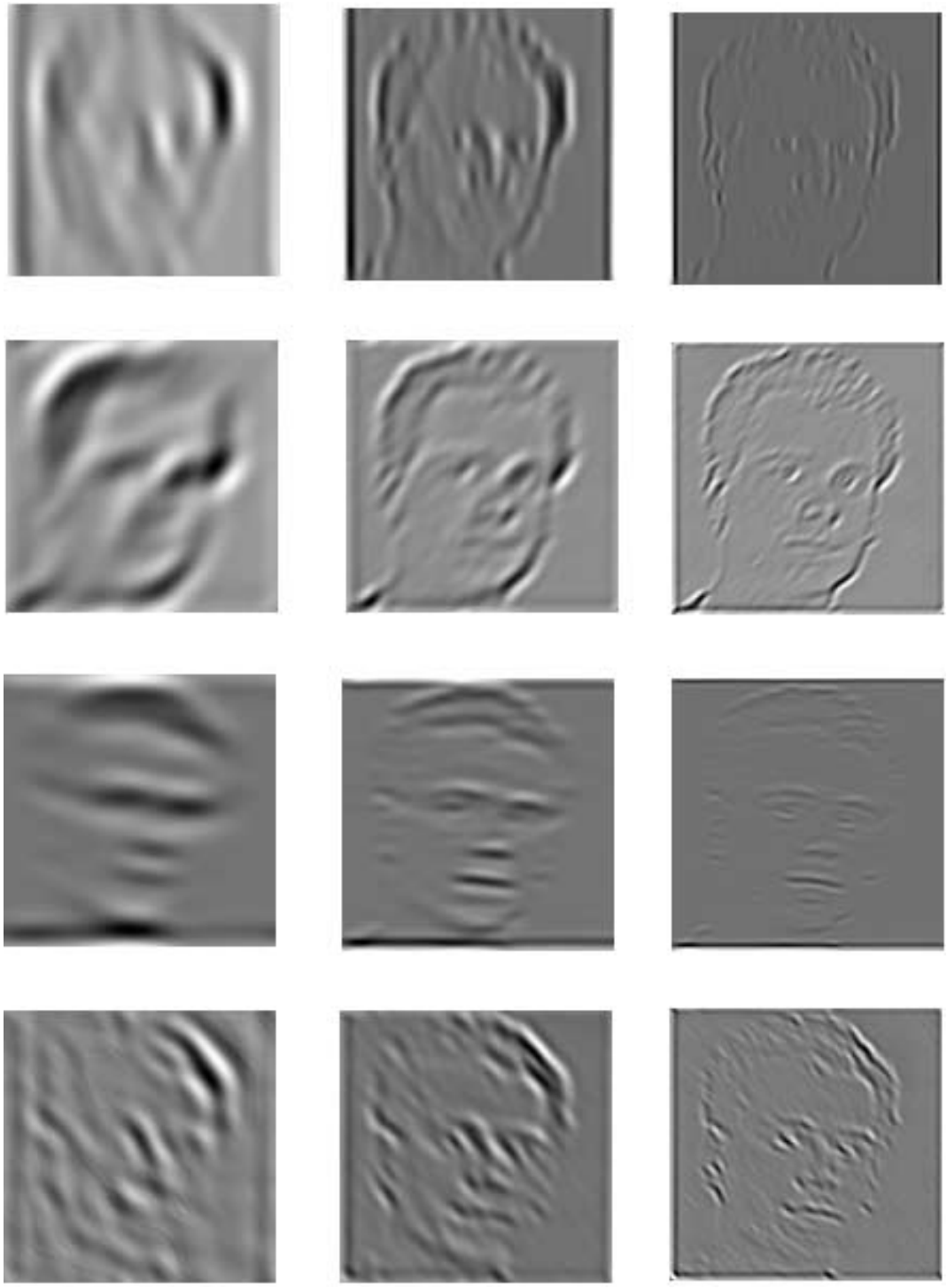

Abb. 2.9: Rekonstruktion mit $\sigma = 2$ für verschiedene Orientierungen (Zeilen) und Frequenzen (Spalten). Verwendet wurden vier Orientierungen (von oben: $0°$, $-45°$, $-90°$ und $-135°$) und drei Frequenzen (von links: $k_0 = 0.4$, $k_1 = 0.77$, $k_2 = 1.5$).

3 Gesichtserkennung mit elastischen Graphen

Im vorherigen Kapitel wurde beschrieben, wie ein Bild durch eine Gabor-Transformationen dargestellt werden kann, und wie die Informationsmenge durch Beschränkung auf wenige Punkte des Bildes - die Knoten eines Graphen - reduziert werden kann. Hier soll gezeigt werden, wie dieses Konzept zur Gesichtserkennung eingesetzt werden kann.

3.1 Das Übereinstimmungs-Problem

Zwei Bilder ein und desselben Gegenstands sind im allgemeinen sehr verschieden: Die Perspektive ist eine andere, gewisse Teile verdecken andere Teile, die Bildpunkte sind als Ganzes oder auch relativ zueinander verschoben. Um die Aufnahmen miteinander vergleichen zu können ist es wichtig zu wissen, welche Punkte der beiden Aufnahmen denselben Punkt des realen Gegenstands darstellen, d.h. welche Bildpunktpaare der Aufnahmen *korrespondieren*. Das Übereinstimmungs-Problem (Correspondence Problem) ist das zentrale Problem der automatischen Bilderkennung. Mit einem punktweisen Vergleich der Grauwerte einzelner Pixel (skalare Größe) kann es nicht bewältigt werden, da es im Vergleichsbild zu viele ähnliche Punkte gäbe, die eine 1 zu 1 Zuordnung verhindern. Daher müssen größere Informationseinheiten miteinander verglichen werden. In dieser Arbeit sind dies die im vorherigen Kapitel eingeführten Jets.

3.2 Repräsentation eines Gesichts

Die zur Gesichtsrepräsentation benutzten Graphen sind objektangepaßt. Die Knoten sind an Punkten angebracht, von denen angenommen werden kann, daß sie gesichtsspezifische Informationen tragen (z.B. Pupille, Nasenspitze, Kinn, nicht aber Haar). Zudem wird darauf geachtet, daß diese Punkte gut von einem Menschen identifiziert werden können. Dies ist wichtig, damit bei der Erstellung der Graphen für jedes Gesicht gleich vorgegangen werden kann. Zusätzliche Knotenpunkte können vom System automatisch hinzugefügt werden, beispielsweise in der

Mitte der Strecke zwischen Ohrläppchen und Pupille. Die Kanten des Gesichtsgraphen verbinden nun nicht etwa alle Knoten untereinander, sondern werden nach zwei Kriterien ausgewählt. Zum einen werden benachbarte Knoten verbunden, zum anderen wird versucht, durch die Anzahl der Kanten den Graph mit speziellen Deformationseigenschaften auszustatten: Knoten an wenig veränderlichen Gesichtsteilen sollen vergleichsweise starr, Knoten an durch Mimik stark veränderlichen Gesichtspunkten beweglich sein.

3.3 Vergleich von Graphen

Um die "Ähnlichkeit" $S_G(G_1, G_2)$ zweier Graphen G_1, G_2 zu bestimmen, kann man die "Ähnlichkeit" $S_J(J_1, J_2)$ der einander entsprechenden Jets J_1, J_2 über alle Knoten aufsummieren. Da im allgemeinen nicht jeder Graph alle Knoten enthält (z.B. Haar verdeckt ein Ohr), wird diese Summe noch mit der Anzahl N der in beiden Graphen vorhandenen Knoten normiert. Da in dieser Arbeit die Gesichtsgraphen aus vornormierten Gesichtern gewonnen werden, sind in jedem Falle genügend gepaarte Knoten vorhanden. Ist die Zahl der in beiden Graphen vorhandenen Knoten zu klein, könnte andernfalls das Ergebnis des Vergleichs stark vom Zufall beeinträchtigt werden.

Neben dem Vergleich der Knoten können auch Verzerrungen der Kanten K_1, K_2 mit einem geeigneten Algorithmus $S_D(K_l^1, K_l^2)$ berücksichtigt werden. Sind die Vorschriften zur Charakterisierung der Jet-Ähnlichkeit $S_J(J_1, J_2)$ und der Verzerrung $S_D(K_1, K_2)$ gegeben, so ist die Ähnlichkeit zweier Graphen S_G dann

$$S_G(G_1, G_2) = \frac{1}{N}\sum_n S_J(J_n^1, J_n^2) - \lambda \sum_l S_D(K_l^1, K_l^2), \tag{3.1}$$

wobei λ die Gewichtung von Knotenähnlichkeit und Verzerrungsfreiheit bestimmt (im folgenden ist, sofern nicht anders angegeben, λ=3). Dabei läuft n über die Zahl der Knoten und l über die Zahl der Kanten.

Da in der zweiten Summe nicht mit der Anzahl der Kanten am Knoten normiert wird, sind stark vernetzte Knoten unbeweglicher als andere, bzw. ihre Ähnlichkeit stärker von der Geometrie abhängig. Der Entwurf eines Programms, das einen Graphen unbekannter Identität mit einer Gallerie von bekannten Graphen vergleicht und die beste Übereinstimmung ausgibt, ist somit unkompliziert. Es stellt sich aber die Frage, wie aus einem Bild ein geeigneter, das Gesicht repräsentierender Graph gewonnen werden kann.

3.4 Erstellung der Graphen

3.4.1 Manuelle Erstellung

Die einfachste Methode, aus einer Aufnahme einen Gesichtsgraphen zu extrahieren, besteht darin, dies manuell vorzunehmen. Dabei werden die festgelegten Punkte vom Benutzer identifiziert und dem Programm eingegeben. Anschließend werden aus der Gabortransformierten des Bildes die dazugehörigen Jets gewonnen. Die Kanten entsprechen dann einfach den Abständen der Knoten. Die Notwendigkeit einer äußeren Eingabe macht diesen Weg jedoch zeitaufwendig und für die Praxis ungeeignet.

3.4.2 Automatische Erstellung

Hat man die Aufnahme eines Gesichts und einen Gesichtsgraphen, der manuell aus einer zweiten Aufnahme extrahiert wurde, so kann man folgendermaßen vorgehen:

Die unbekannte Aufnahme wird gabortransformiert und der Graph bekannter Identität wird auf diesem aus Jets bestehenden Datenfeld plaziert. Die unter den Knoten des Graphen liegenden Jets können als ein zweiter etikettierter Graph mit der selben geometrischen Struktur interpretiert werden. Die Ähnlichkeit der beiden Graphen wird nun über (3.1) bestimmt, und der Prozess beginnt mit einer verschobenen Position des bekannten Graphen erneut. Dies wird solange wiederholt, bis entweder die Ähnlichkeit einen festgelegten Schwellwert überschreitet, oder bis alle möglichen Positionierungen des Graphen ausgeschöpft sind.

Prinzipiell kann man auf diese Weise auch ein Bild konsekutiv mit einer Vielzahl von etikettierten Graphen aus einer Gallerie vergleichen. Jeder dieser Vergleichsschritte prüft aber nur, ob ein spezielles Gesicht auf der unbekannten Aufnahme zu sehen ist. Dies bedeutet bei großen Gallerien einen erheblichen Rechenaufwand, zudem wird ein unbekanntes Gesicht nur schlecht klassifiziert.

Daher ist es wünschenswert die beiden Schritte "Auffinden eines Gesichts" und "Erkennen eines bekannten Gesichts" voneinander zu trennen. Zur Extraktion eines Gesichts benötigt man folglich einen Graph, der nicht ein einzelnes Gesicht repräsentiert, sondern allen menschlichen Gesichtern ähnlich ist. Dies wird mit einem Face Bunch Graph (siehe Wiskott [Wis95]) erreicht.

3.4.3 Der Face Bunch Graph (FBG)

Der FBG wird erzeugt, indem eine bestimmte Anzahl Z von manuell erzeugten Gesichtsgraphen (hier: Z=70) zusammengefaßt wird. Die Kanten des FBG werden aus dem Mittel aller dieser Graphen berechnet, die Knoten werden mit allen Knoten der Graphen etikettiert. Am Knoten "rechte Pupille" im FBG sind somit die Pupillen-Jets von Z Personen zusammengefaßt.

Bei der Plazierung des FBG auf einem gabortransformierten Bild wird für jede Position des Graphen an jedem Knoten der ähnlichste Jet ausgewählt. Dadurch ergibt sich eine große Generalisierungsfähigkeit: Durch Kombination einzelner Jets verschiedener Personen können neue Gesichtsgraphen gebildet werden (Z^N Möglichkeiten, N=Anzahl der Knoten, $Z > 1$). Daher sollten die benutzten Modellgesichter so unterschiedlich wie möglich sein. Sind im FBG keine Gesichter mit Bärten, Brillen oder starker Mimik vertreten, so können nur ungenaue Graphen solcher Gesichter extrahiert werden. Aus allen Graphen des FBG wird ein angepaßtes Gesicht zusammengesetzt und zum Vergleich mit dem unbekannten Bild benutzt. Die unter der optimalen Plazierung liegenden Jets werden danach zu einem Gesichtsgraphen zusammengefaßt. Dieser nunmehr automatisch erstellte Graph kann mit einer Gallerie verglichen oder ihr hinzugefügt werden.

3.4.4 Erstellen einer Gallerie

Die Gallerie besteht aus der Gesamtheit aller etikettierter Graphen der Bilder von Personen, deren Identität dem System bekannt ist. Gewöhnlich ist ihre Anzahl größer als die Zahl der Graphen im FBG, wobei Überschneidungen möglich sind. Mit Hilfe des FBG können aus einer Gesichtsdatenbank problemlos Graphen gewonnen und zu einer Gallerie hinzugefügt werden. Es ist lediglich darauf zu achten, daß die Normierung (Größe, Perspektive) der Eingabebilder dem Format der im FBG enthaltenen Graphen entspricht.

3.5 Ausführliche Darstellung der elastischen Graphenanpassung

Für die Charakterisierung der Ähnlichkeit eines etikettierten Bildgraphen B mit dem FBG wird folgende Funktion verwendet:

$$S_G(B, FBG) = \frac{1}{N}\sum_n \underset{m}{max}\big(S_J(J_n^B, J_n^{FBG(m)})\big) - \lambda \sum_l \frac{(\Delta\vec{x}_l^B - \Delta\vec{x}_l^{FBG})^2}{(\Delta\vec{x}_l^{FBG})^2}. \quad (3.2)$$

Dabei wird für jeden Konten n der FBG-Jet m verwendet, bei dem S_J maximal wird. Der Index n läuft über die Anzahl N der Knoten und Index l über die Anzahl Kanten.

Als Vorschrift für die Ähnlichkeit zweier Jets wird die Amplitudenähnlichkeit verwendet:

$$S_J(J, J') = \sum_{j=1} a_j a'_j. \quad (3.3)$$

Der Index j läuft über die Anzahl der Gaborkerne (Koeffizientenzahl des Jets).

Im ersten Term wird für jeden Knoten der jeweils ähnlichste Jet des FBG aufsummiert. Die zweite Summe beschreibt die Verzerrung des extrahierten Graphen gegenüber dem Durchschnittsgesicht des FBG. Dabei ist $\Delta\vec{x}$ ein Verktor, der den Verlauf einer Kante beschreibt. Das quadratische Kraftgesetz wurde der Anschaulichkeit halber verwendet – prinzipiell sind eine Vielzahl anderer Bindungskräfte denkbar.

Die Extraktion eines geeigneten Bildgraphen erfolgt in mehreren Schritten:

1. Finden des Gesichts: Für jeden Knoten des FBG werden die zugehörigen Jets zu einem durchschnittlichen Jet gemittelt. Dieser durchschnittliche etikettierte Graph wird in der linken oberen Ecke des Bildes positioniert und mit einer Schrittweite von vier Pixeln über das Bild geschoben. Dabei wird keine Deformation des Graphen zugelassen ($\lambda \to \infty$). Um die Postion mit größter "Ähnlichkeit" herum wird dieser Prozeß mit einer Schrittweite von einem Pixel wiederholt. Die daraus resultierende Position dient als Startposition für den nächsten Schritt.

2. Verwenden aller Jets der FBG: Um die Startpostion herum ($\pm$ 9 Pixel) wird der FBG auf jeder möglichen Stelle plaziert. Dabei werden alle Jets eines Knotens für den Vergleich zugelassen.

3. Verzerren des Graphen: In der dritten Stufe wird der Wert von λ endlich und Graphendeformation möglich. Jeder einzelne Knoten kann sich nun von seiner Startposition wegbewegen. Diese Position wird "angenommen" wenn der FBG als Ganzes dem Bildgraphen ähnlicher wird. Dies hängt davon ab, ob der Ähnlichkeitsgewinn durch bessere Jet-Ähnlichkeit den Ähnlichkeitsverlust durch eine stärkere Verzerrung des Graphen übertrifft. Dieser Schritt wird mehrmals wiederholt, da dann der Graph als Ganzes ohne starke Verzerrung über das Bild wandern kann – die "Relaxation" eines Knoten kann die Bewegung eines anderen Knoten ermöglichen. An der auf diese Art ermittelten Position wird der Bildgraph extrahiert.

3.6 Verwendung von Phaseninformation

Die Ähnlichkeitsfunktion (3.3) berücksichtigt nur die Ähnlichkeit der Amplituden der Jets. Dies hat praktische Gründe, da das Verhalten der Amplitude wesentlich stabiler ist: Aus Abbildung (2.5) ist ersichtlich, daß die Amplitude örtlich nur langsam variiert und somit Jets von benachbarten Punkten vergleichbare Amplituden haben.

Die Phase dagegen rotiert in Abhängigkeit von der Zentralfrequenz $\vec{k}$ rasch mit einer Periode von wenigen Pixeln, d.h. ähnliche Phasenwerte deuten nicht

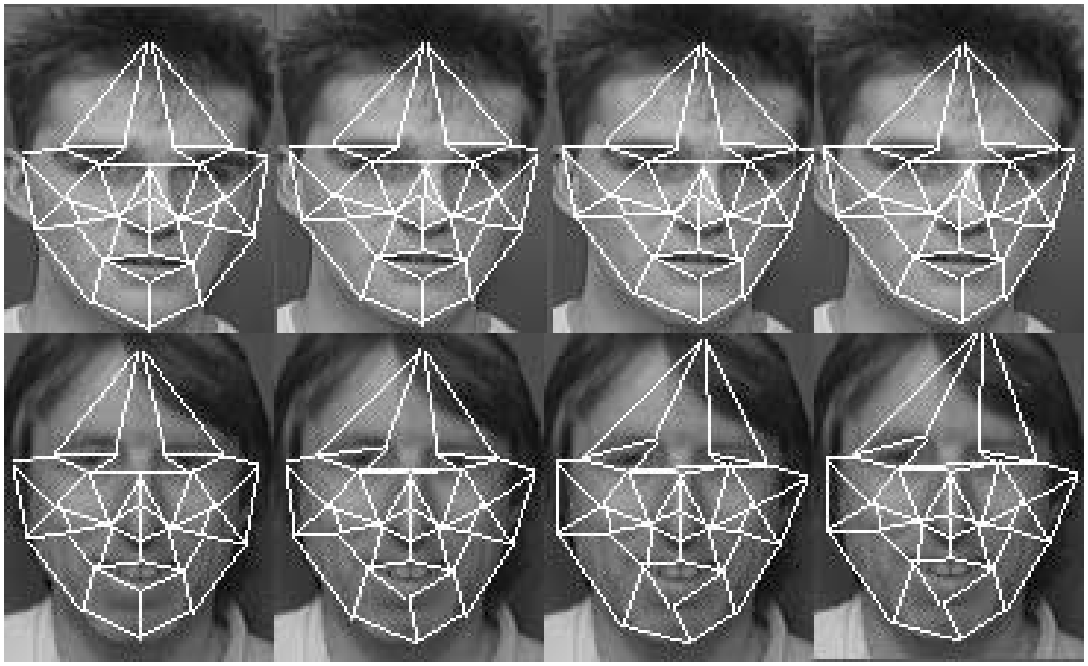

Abb. 3.1: Anpassung der Graphen: Der unter (3.5) beschriebene Schritt 3 wird mehrmals ausgeführt, wobei sich der etikettierte Graph den Ähnlichkeitsmerkmalen des Eingangsbildes anpaßt. Von dem durchschnittlichen Graph (links) ausgehend bewegen sich die Knoten zu den lokalen Ähnlichkeitsmaxima (Jetähnlichkeit, (3.3)). Die dabei entstehenden Verzerrungen des Graphen wirken dem "Auseinanderfließen" entgegen. In der oberen Reihe erkennt man, wie die für das linke Ohr zuständigen Knoten schrittweise auf den korrekten Positionen plaziert werden. Bei der unteren Bildfolge ist auffällig, wie der Scheitel zwar anfangs richtig zu liegen kommt, daß System dies aber aufgrund des ungewöhnlichen Haaransatzes nicht erkennt und vergeblich nach einer horizontalen Kante sucht. Die Graphenähnlichkeiten (FBG zu extrahiertem Graphen) gemäß (3.2) sind für die obere Reihe: 0.905016, 0.990401, 0.993716 und 0.993855, für die untere Reihe 0.863719, 0.90716, 0.91319 und 0.915754. Je nach Elastizität wird nach einigen Schritten (hier vier mit $\lambda = 0.1$) ein stabiler Zustand erreicht.

auf ähnliche Grauwertumgebungen der Jets hin. Es liegt daher nahe, für das Erkennsystem lediglich die Amplitudeninformation zu verwenden.

Es zeigt sich aber, daß gerade die Phaseninformation für das menschliche Wahrnehmungssystem bedeutsam ist. Dies erkennt man anhand von Abbildung (3.2): Werden zwei Aufnahmen fouriertransformiert und die Phaseninformation vertauscht, so zeigt die Rücktransformation das Gesicht derjenigen Person, deren Phase verwendet wurde.

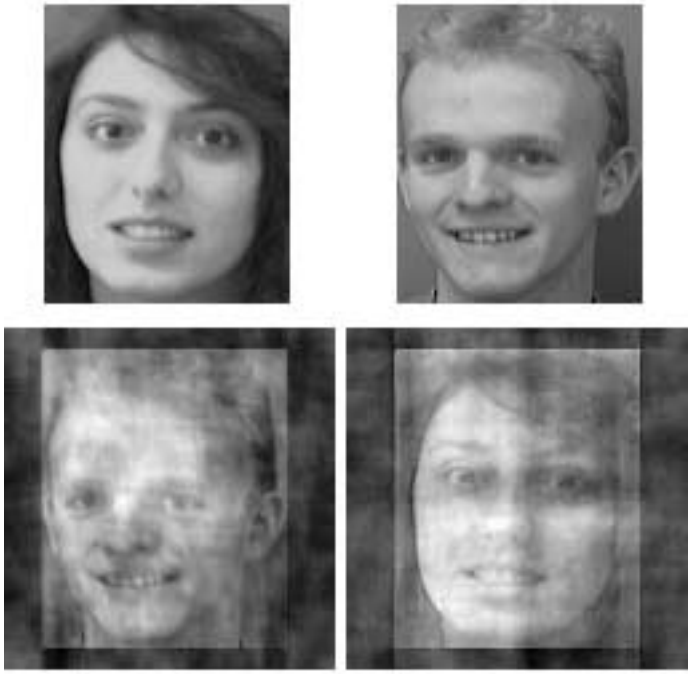

Abb. 3.2: Die beiden oberen Gesichter wurden fouriertransformiert und ihre Phaseninformation vertauscht. Die Rücktransformation zeigt, daß die Phase die für die menschliche Gesichtsidentifikation relevanten Daten enthält.

Ein weiterer Grund der die Verwendung von Phaseninformation nahelegen würde ist, daß bei der Amplitudenbildung das Vorzeichen verloren geht. Dadurch wird die Polarität einer Kante unklar, ein Wechsel von hell nach dunkel und von dunkel nach hell ergibt identische Amplituden, so daß das System beispielsweise zwischen Unter- und Oberlippe nur schwer unterscheiden kann.

Um die Phasen zweier Jets sinnvoll vergleichen zu können, muß der Teil des Phasenunterschieds, der nur aufgrund der verschiedenen Positionen zustande kommt, kompensiert werden. Fleet [Fle92] hat gezeigt, daß die Phase im allgemeinen das Verhalten einer ebenen Welle hat, deren Frequenz als die der Zentralfrequenz des Gaborwavelets approximiert werden kann (siehe Abbildung 3.3). Die Ausnahme sind Positionen, an denen die Amplitude sehr klein wird, wodurch das Phasenverhalten instabil wird. Daher wurden in dieser Arbeit für Jetkoeffizienten mit kleiner Amplitude sowohl Phase als auch Amplitude gleich Null gesetzt.

Mit dieser Näherung ergibt sich die Phasendifferenz, die auf die Verschiebung der Jets zurückzuführen ist, zu

$$\Delta\phi = \vec{d}\vec{k}_j, \tag{3.4}$$

und die Ähnlichkeitsfuntion kann als

$$S_J(J, J') = \sum_{j=1} a_j a'_j cos(\phi_j - \phi'_j - \vec{d}\vec{k}_j) \tag{3.5}$$

definiert werden. Dabei ist $\vec{k_j}$ die Zentralfrequenz und $\vec{d}$ der Abstand der Positionen, an denen der jeweilige Jet gewonnen wurde. Wiskott [Wis95] schätzt diesen Abstand wiederum aus der Phasendifferenz ab, was für nah benachbarte Jets möglich ist. Für unsere Zwecke ist es jedoch ausreichend, jedem Graphen zusätzlich zu den relativen Jetpositionen noch einen absoluten Orientierungswert mitzugeben. Dieser legt fest, an welcher Pixelposition der Graph aus dem Bild gewonnen wurde.

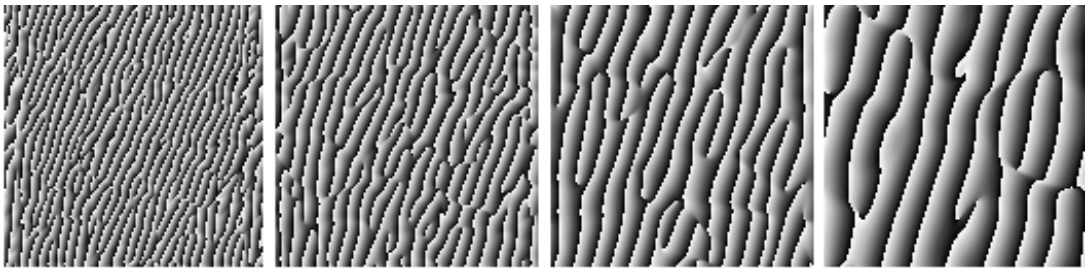

Abb. 3.3: Die vier Abbildungen zeigen die Phase eines gabortransformierten Gesichts für eine Richtung ($\phi = -22.5°$)und vier verschiedene Frequenzen (von links $k_0 = \frac{\pi}{2}$ bis $k_4 = \frac{\pi}{2\frac{5}{2}}$).

3.7 Erkennung

Der Bildgraph kann nun auf einfache Weise mit allen Graphen einer Gallerie G verglichen werden.

Wie schon angesprochen, ist nicht notwendigerweise für jeden Knoten des FBG ein entsprechender Punkt im unbekannten Bild vorhanden. Beispielsweise kann ein Ohr durch Haar oder Profilstellung des Gesichts verdeckt sein. In diesem Falle kommt der entsprechende Knoten des FBG gezwungenermaßen falsch zu liegen. Es wäre wünschenswert, eine effiziente Möglichkeit zur Erkennung solcher Fehlplazierungen in das System zu integrieren. Daher wurde versucht, diese Knoten anhand ihrer mutmaßlich schlechten Jet-Ähnlichkeit zu erkennen und ab einem Schwellwert auszuschließen. Dies führte aber nicht zum gewünschten Resultat.

Arbeiten zur "Pose Detection" zeigen diese Schwierigkeit deutlich: Man benötigt Kenntnis der Pose, um zu wissen, welche Knoten fehlen, andererseits soll gerade aus den fehlenden Knoten auf die Pose geschlossen werden. In dieser Arbeit werden lediglich Knoten ausgeschlossen, die außerhalb der 92x112 Pixel großen

Aufnahmen zu liegen kommen. Das System schlägt dann den ähnlichsten Graphen als Identität des Bildgraphen vor.

4 Erkennleistung auf der ORL-Datenbank

4.1 Beschreibung der ORL-Datenbank

Das Erkennsystem wurde anhand der "Database of Faces"(früher in Besitz der Olivetti Research Laboratories) der AT&T Laboratories Cambridge erstellt und erprobt.

Die Datenbank umfaßt 40 Personen, die mit je 10 Aufnahmen vertreten sind. Bei allen Bildern ist der Hintergrund dunkel und die Blickrichtung in etwa frontal in die Kamera gerichtet. Bei einigen Personen wurden die Aufnahmen zu verschiedenen Zeitpunkten gemacht, wobei Beleuchtung und Gesichtsausdruck variierten, zudem wurden Brillenträger sowohl mit als auch ohne Brille photographiert. Die 92x112 Pixel großen Bilder verfügen über 256 Graustufen. Abbildung (4.1) zeigt einen Ausschnitt.

Zur Konstruktion des FBG wurden von 35 Personen für je zwei Bilder manuell etikettierte Graphen gemäß Abbildung 2.1 erstellt. Mit Hilfe des FBG konnten nun automatisch Graphen für alle 400 Gesichter erstellt werden. Dabei ist zu beachten, daß sich automatisch erstellte Graphen von manuellen Graphen derselben Aufnahme unterscheiden, da dem FBG zwar alle "korrekten" Knoten zur Verfügung stehen, den Kanten aber die durchschnittliche Geometrie aller Graphen im FBG zugrunde liegt.

4.2 Erkennleistung bei verschiedenen Gallerien

In Abbildung (4.2) wird die Erkennleistung des in Kapitel 3 vorgestellten Systems in Abhängigkeit von der Zahl der Modellgesichter in der Gallerie dargestellt. Erkennleistung bedeutet hier wie im folgenden wieviele der vom System als "ähnlichster Graph" ermittelten Graphen tatsächlich korrekt zugeordnet wurden. Die Zahl der Modellgesichter gibt jeweils an, wie groß die Gallerie für den Graphenvergleich ist. Die kleinste Gallerie umfaßt 40 Gesichter (eines von jeder Person), die größte 360 (alle anderen Aufnahmen der zu identifizierenden Person sind in der Gallerie enthalten). Intern werden mehrere Aufnahmen derselben Person als verschieden betrachtet, d.h. diese Aufnahmen konkurrieren auch untereinander

Abb. 4.1: Die ersten 10 Personen aus der ORL Datenbank mit ihren verschiedenen Ansichten. Man beachte die stark wechselnden Gesichtsausdrücke z.B. der 7. Person von oben.

um die Zuordnung als "ähnlichster Graph". Lediglich in der Aufbereitung der Ergebnisse für den Benutzer werden alle Zuordnungen zu einer der korrekten Aufnahmen zusammengefaßt.

Bei den Berechnungen mußte das System versuchen, alle 400 Gesichter zu erkennen. Die erste Aufnahme (1. Spalte) der ersten Person (1. Zeile) wird als erste dem System zur Erkennung übergeben. Danach folgen alle weiteren Aufnahmen der 1. Spalte. In dieser Weise werden auch alle weiteren Spalten abgearbeitet.

Die Gallerie setzt sich dabei aus Graphen der jeweils nächsten Spalten (siehe Abbildung 4.1) zusammen, so daß in keinem Falle ein aus dem zu erkennenden Bild gewonnener Graph in der Gallerie enthalten ist.

Bei einem Modellgesicht pro Person (Gallerie umfaßt eine, d.h. die nächste Spalte) werden 291 der 400 Gesichter korrekt erkannt, in weiteren 56 Fällen wird die richtige Person auf Rang 2 bis 5 (von 40 möglichen) gesetzt. Umfaßt die Gallerie volle neun Modellgesichter (alle restlichen Spalten) so werden 393 Gesichter erkannt. Weitere vier richtige Zuordnungen landen auf den Rängen zwei bis fünf. Bei drei Aufnahmen liegt die richtige Zuordnung schlechter als Rang 5.

Zur Erkennung einer Aufnahme werden je nach Größe der Gallerie 3,0 bis 4,7 Sekunden benötigt. Alle Rechnungen wurden auf einer Compac ES/40 durchgeführt.

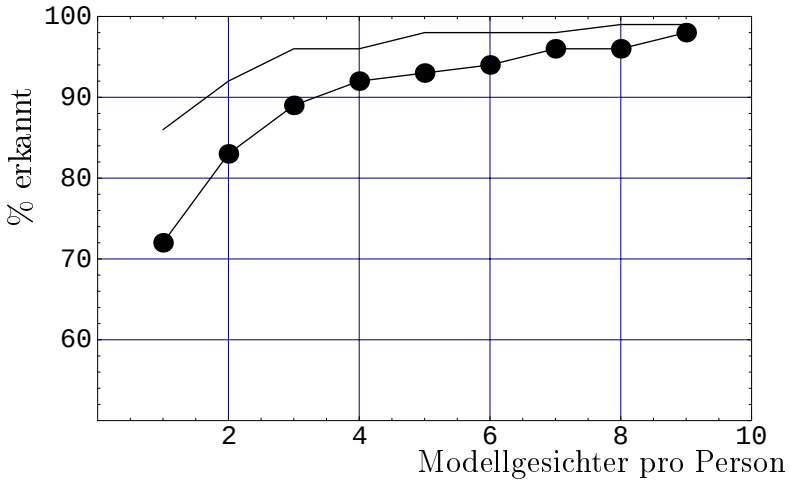

Abb. 4.2: Erkennleistung in Prozent in Abhängigkeit von der Zahl der Modellgesicher pro Person in der Gallerie. Die durchgezogene Linie zeigt richtige Zuordnung auf Rang 1 bis 5, die gepunktete Kurve zeigt korrekte Zuordnung auf Rang 1.

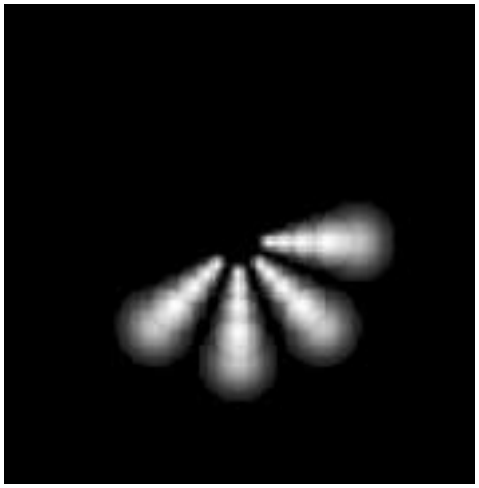

Abb. 4.3: Die Superposition von 20 Gaborkernen im Frequenzraum. Die dazwischen liegenden Orientierungen werden nicht erfaßt.

4.3 Erkennleistung bei halber Datenmenge

Die in den Graphen verwendeten Jets umfassen 40 Koeffizienten. Diese resultieren aus den Antworten der Gabortransformation mit 40 Werten für k_j (fünf Frequenzen mal acht Orientierungen, siehe (2.4.3)). Verwendet man nur die Hälfte der Orientierungen, so läßt sich die Datenmenge pro Graph halbieren. Der Fourierraum wird (sofern die Breite $\sigma = 6$ in den Gaborkernen beibehalten wird) nur unvollständig abgedeckt, siehe Abbildung (4.3).

Abbildung (4.4) zeigt, daß die Erkennleistung dennoch nur mäßig abnimmt. So findet man 275 Treffer bei einem Modellgesicht, weitere 66 korrekte Zuordnungen wenigstens auf Rang 2 bis 5. Da eine Gallerie mit zwei 20-koeffizientigen Modellgesichtern pro Person die gleiche Datenmenge wie eine Gallerie mit einem 40-koeffizientigen Modellgesicht enthält, schneidet das System mit den "kurzen" Jets sogar besser ab: Die Zwei-Modellbilder-Gallerie mit kurzen Jets erkennt 81% gegenüber den 72% der Ein-Modellbild-Gallerie mit langen Jets. Auch beim Vergleich der Vier-, Sechs- und Acht-Bildgallerien mit den Zwei-, Drei- und Vier-Bildergallerien schneidet das System mit kurzen Jets besser ab.

Der Grund dafür ist, daß die Ähnlichkeit zweier Bilder in hohem Maße von der Ansicht abhängt und das System mit kurzen Jets doppelt so viele Posen zur Verfügung hat.

Da gewisse Richtungen im Frequenzraum bei der Gabortransformation mit nur 20 Koeffizenten gar nicht erfaßt werden, könnte man diesen Bildern noch ein Bild überlagern, das nur aus Frequenzen in diesen Richtungen besteht, und das System würde keinen Unterschied bemerken. Die Robustheit der Erkennleistung gegenüber Reduzierung der Zahl der Gaborkerne zeigt, daß die in der Datenbank enthaltenen Gesichter keine prädominanten Frequenzen in den nicht abgedeckten

Bereichen aufweisen.

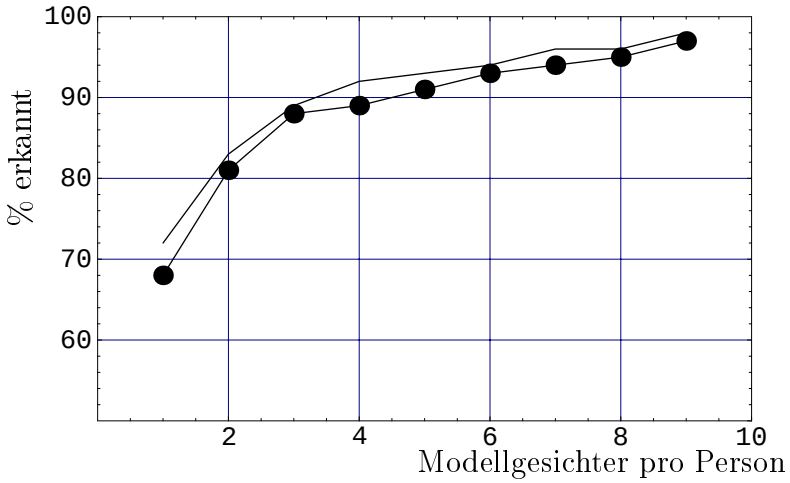

Abb. 4.4: Erkennleistung in Prozent in Abhängigkeit von der Zahl der Modellgesicher pro Person in der Gallerie bei Halbierung der Zahl der Richtungen der Gaborkerne (Abbildung 4.3). Die gepunktete Kurve zeigt korrekte Zuordnung auf Rang 1 unter Verwendung von nur 20 Koeffizienten pro Jet. Die durchgezogene Kurve zeigt zum Vergleich das Ergebnis für 40 Koeffizienten (volle Zahl der Richtungen).

4.4 Spiegelung der Bilder

Da Gesichter annäherend symmetrisch sind, sollte das System ein direkt in die Kamera blickendes, an der vertikalen Achse gespiegeltes Gesicht ebenso leicht erkennen wie das Original. Die meisten Aufnahmen der ORL-Datenbank sind nicht völlig frontal, so daß aufgrund der Posenempfindlichkeit des Systems eine Erkennung schwierig wird. Bei Verwendung einer großen Zahl von Modellgesichtern sollten aber genügend Abbildungen mit Links- oder Rechtsprofil zur Verfügung stehen, um eine ähnlich gute Erkennleistung zu ermöglichen. Abbildung (4.5) zeigt jedoch eine deutliche Abnahme der Erkennleistung. Möglicherweise ist in der Gallerie eine Blickrichtung bevorzugt, d.h. für die einzelnen Personen überwiegt meist eine Profilstellung. Wahrscheinlicher ist jedoch der Einfluß von Asymmetrien, beispielsweise der Frisur oder Beleuchtung.

Insgesamt wurden unter Verwendung von einem Modellgesicht pro Person nur 52% der Gesichter erkannt (gegenüber 72% für ungespiegelte). Auch für volle neun Modellgesichter lag die Erkennleistung mit 82% (gegenüber 98%) deutlich niedriger. Es liegt nahe, eine Kombination beider Verfahren zu erproben: Das Originalbild und seine Spiegelung werden mit der Gallerie verglichen und das jeweils bessere Ergebnis verwendet. Es zeigte sich jedoch, daß zwar bislang nicht

erkannte Gesichter identifiziert wurden, gleichzeitig aber neue Fehlerkennungen auftraten, so daß das Ergebnis im Endeffekt schlechter ausfiel.

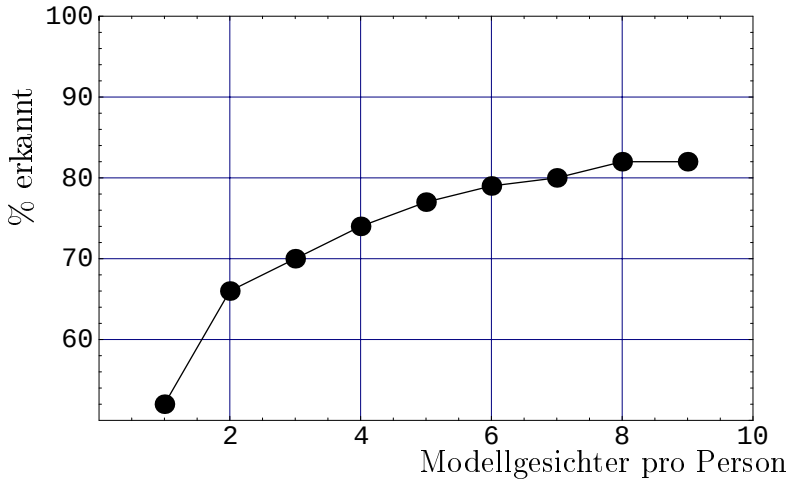

Abb. 4.5: Werden die Gesichter gespiegelt und mit ungespiegelten Gesichtern verglichen, so nimmt die Erkennleistung stark ab. Der Kurvenverlauf entspricht der normalen Erkennleistung (Abbildung 4.2), die absoluten Werte liegen allerdings damit verglichen um 15-20 Prozentpunkte niedriger.

4.5 Verdecken des oberen Bildteils

Bei der Konstruktion der Graphen wurde nur ein einziger Knoten oberhalb der Augenbrauen angebracht (Scheitel). Auf diese Weise sollte das System robust gegen veränderliche Frisuren sein.

Um dies zu überprüfen, wurden dem System teilverdeckte Aufnahmen zur Erkennung vorgelegt: Das obere Viertel der 112 Pixel hohen Bilder wurden verdeckt, vergleiche auch Abbildung (4.6). Das Ergebnis zeigt, daß die Erkennleistung nur wenig zurückging. Für ein Modellgesicht erniedrigte sich die Erkennleistung um 7% auf 65%, für drei und mehr Modellgesichter ist sie praktisch identisch mit der bei Erkennung unverdeckter Gesichter. Abbildung (4.7) zeigt den bekannten Anstieg der Erkennleistung mit der Zahl der verwendeten Modellgesichter.

4.6 Verwendung eines Identitäts-Bündel-Graphen(IBG)

Zur Extraktion eines Gesichtsgraphen wird ein FBG benutzt, also ein Algorithmus verwendet, der aus einer festen Menge bekannter Graphen ein Gesicht "zu-

Abb. 4.6: Das obere Viertel des Gesichts wurde schwarz überschrieben. Da sich nur ein Knoten oberhalb der Augenbrauen befindet, wird die Leistung des Systems nur wenig beeinträchtigt. Dies sollte es robust gegen veränderliche Frisuren oder Hüte machen.

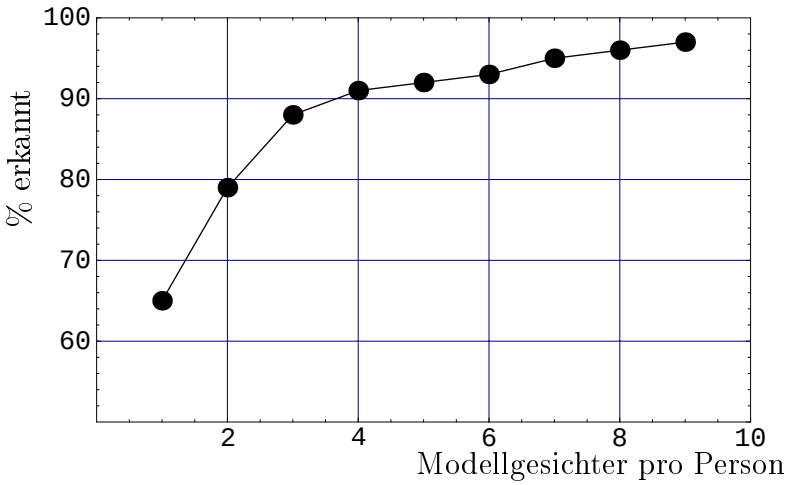

Abb. 4.7: Das Verdecken des oberen Bildbereichs (25%) führt nur zu einer geringen Abnahme der Erkennleistung. Dies zeigt, daß die Graphen in guter Näherung von Frisur oder Hüten nicht beeinträchtigt werden.

sammenpuzzelt", welches dem unbekannten Bild möglichst ähnlich sieht. Danach wird an dieser Stelle der Gesichtsgraph extrahiert und dieser wiederum mit einer Gallerie verglichen.

Da in der Gallerie mehrere Graphen derselben Person zur Verfügung stehen (sofern die Zahl der Modellgesichter größer als eins ist), liegt es nahe, das Prinzip des FBG nicht nur zur Graphenerstellung, sondern auch zum Graphenvergleich zu nutzen. Dazu wird zunächst der gewöhnliche Vergleichsalgorithmus benutzt. Die fünf besten Graphen der Gallerie werden dann in Identitäts-Bündel-Graphen(IBGs) verwandelt: Jeder der fünf Graphen wird mit den anderen Graphen derselben Person analog zum FBG zusammengefügt. Die Geometrie ist die des extrahierten Graphen, an den Knoten sitzen nun sämtliche Jets der dieser

Person zugeordneten Graphen.

Für den Vergleich werden nun alle fünf IBGs mit dem extrahierten Graphen verglichen, wobei an den Knoten jeweils der beste Jet des IBGs verwendet wird. Beispielsweise kann der Jet am linken Auge der einen Aufnahme und der Jet am rechten Auge einer anderen Aufnahme derselben Person verwendet werden. Die Identität des ähnlichsten IBG wird dann als Erkennung vorgeschlagen.

Die Erkennleistung wurde durch dieses Verfahren nur minimal verbessert und zwar im Mittel nur um 0,5%. Eine genauere Analyse zeigt, daß einer Anzahl von korrekten Verbesserungen eine ähnlich große Zahl von neuen Fehlerkennungen gegenübersteht. Das bedeutet, daß sich durch Kombination der Jets verschiedener Ansichten derselben Person zwar Graphen generieren lassen, die den Graphen unbekannter Aufnahmen dieser Person sehr ähnlich sind, daß dabei aber auch Kombinationsgraphen anderer Personen dem unbekannten Graph ähnlich werden können.

4.7 Variation der Elastizität

Sämtliche vorherige Erkennleistungen bezogen sich auf relativ starre Graphen (λ=3). Es war das Ziel dieser Arbeit ein robustes System zu entwickeln, das gegen mäßige Änderungen der verwendeten Parameter möglichst unempfindlich ist. Andernfalls hätte die Gefahr einer speziell für die untersuchten Datenbanken optimierten Parameterwahl bestanden.

In den Abschnitten 4.3 und 4.5 wurden Veränderungen der Anzahl der Waveletkoeffizienten und des verwertbaren Bildbereichs untersucht. Auch bei Verwendung eines deutlich kleineren Wertes, λ=1, ist die Erkennleistung unverändert hoch. Für ein Modellgesicht liegt die Erkennrate bei 73%, auch bei größeren Gallerien unterscheiden sich die Erkennleistungen für die beiden Werte von λ um weniger als einen Prozentpunkt (siehe auch Abbildung (4.8)).

Die mit verschiedenem λ extrahierten Graphen unterscheiden sich dagegen stark (siehe auch Abbildung (3.1)) in ihrer Geometrie, so daß die Gallerie für jeden Wert einzeln angelegt werden muß.

Die Ergebnisse unter Verwendung eines IBG zeigen eine leichte Verbesserung der Erkennleistung. Dies legt den Schluß nahe, daß auch ein simples, nicht "gesichtsangepaßtes" Gitter von Punkten mit starren Verbindungen ähnlich gute Ergebnisse wie die hier verwandten Graphen erbringen würde.

4.8 Zur Verwendung der Phaseninformation

Um das Potential des Phasenvergleichs abschätzen zu können, wurden zunächst die aus der Gabortransformation eines Gesichts entstehenden Jets mit einem Jet verglichen, der der linken Pupille der gleichen Aufnahme entnommen worden war

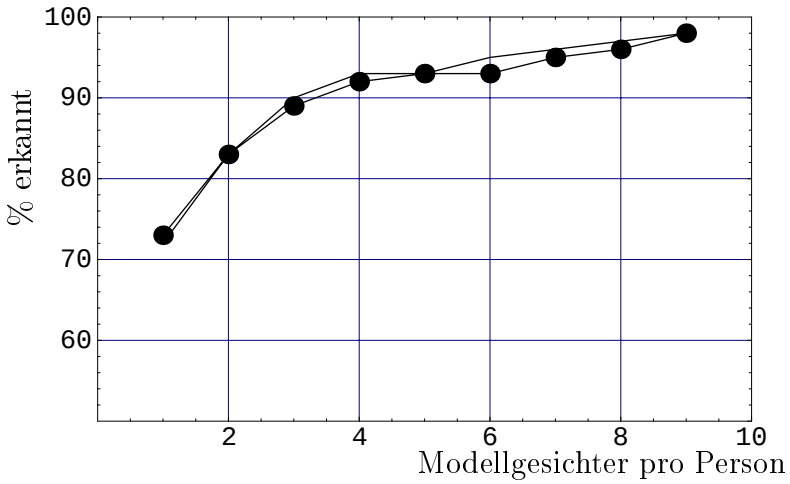

Abb. 4.8: Mit λ=1 erzielte Erkennleistung. Obwohl eine "weiche" Graphenstruktur eine wesentlich stärkere Verzerrung der aus einem Bild extrahierten Graphen erlaubt, ist die Erkennleistung davon unbeinträchtigt. Die durchgezogene Linie zeigt die Erkennleistung unter Verwendung von Identitäts-Bündel-Graphen(IBG)

(Abbildung (4.9)).

Beim Amplitudenvergleich werden die Umgebungen um beide Augen als ähnlich eingestuft, beim Vergleich mit Phase wird das linke Auge sehr scharf eingegrenzt. Dieses Eingrenzen kann in Stufen erfolgen, wenn man zunächst nur die Phasenübereinstimmung einer langen Wellenlänge (entspricht $k_4=\frac{\pi}{8}$, kurz im Fourierraum) verwendet und dann konsekutiv zu kürzeren Wellenlängen übergeht.

Dabei ist allerdings zu bedenken, daß hierbei ein identischer Jet gematcht wurde. Für ein Erkennsystem muß aber ein Jet, der aus einer anderen Aufnahme derselben Person gewonnen wurde, für den Vergleich verwendet werden. Es gilt daher abzuschätzen, ob beispielsweilse ein Augen-Jet ein Auge in einer unbekannten Aufnahme mit Hilfe der Phaseninformation präziser lokalisieren kann.

Dazu wurde folgende Rechnung durchgeführt: Von den 35 Personen des FBG sind je zwei präzise, da von Hand erstellte, Graphen vorhanden. Eines der Bilder aus der 10. Spalte der Gesichtsdatenbank wird gabortransformiert, so daß man ein Feld von Jets erhält. Dann werden sämtliche Jets der Graphen von Bildern der 1. Spalte mit diesem Jetbild verglichen. Die besten Übereinstimmungen werden mit drei Verfahren gesucht: Für den reinen Amplitudenvergleich, den Amplituden- plus Phasenvergleich und für eine Mischform, bei der um das Maximum des Amplitudenvergleichs 10 Pixel weit die beste Amplituden- und Phasenübereinstimmung gefunden wird. Die dabei ermittelten Positionen können nun mit den echten Positionen verglichen werden, da diese von Hand ermittelt wurden. Die jeweils beste Vergleichsmethode für diesen Jet wird gespeichert, und nachdem alle 35x20 Jets der 1. Spalte abgearbeitet wurden, zeigt sich beispiels-

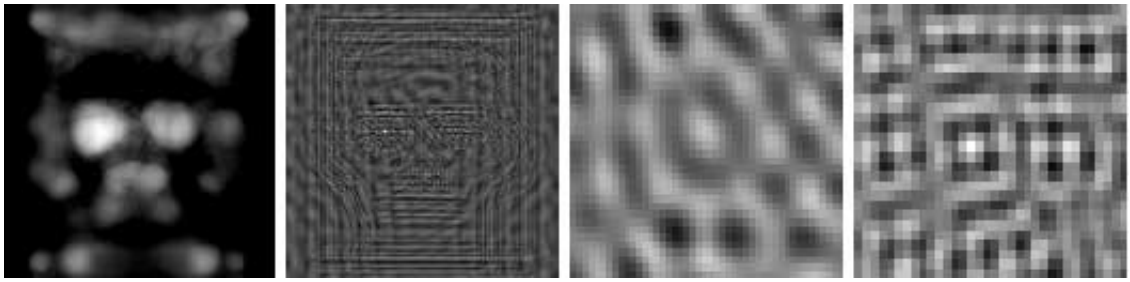

Abb. 4.9: Vergleich der Ähnlichkeit des Jets am linken Auge mit allen anderen Jets derselben Aufnahme. Links die Ähnlichkeit gemäß (3.3, Ampitudenvergleich), dann gemäß (3.5, Phasenvergleich). Die beiden rechten Aufnahmen zeigen vergrößerte Ausschnitte des Phasenvergleichs um die linke Pupille, wobei jeweils alle Richtungen aber nur eine Frequenz (k_4 und k_2) verwendet wurden. Dadurch kann eine "ähnliche" Umgebung in mehreren Stufen eingegrenzt werden (siehe unten).

weise für Person 1, daß für 325 Jets der Amplituden- plus Phasenvergleich am erfolgreichsten war, während der reine Amplitudenvergleich und die Mischform nur 233 bzw. 142 Erfolge hatten. Wiederholt man dies für alle anderen Bilder der 10. Spalte, so bestätigt sich, daß der Amplituden- plus Phasenvergleich etwa doppelt so viele Erfolge wie der Amplitudenvergleich hat, während die Mischform am schlechtesten abschneidet.

Somit ist gezeigt, daß die Phaseninformation zur Lokalisierung einer Gesichtspartie eingesetzt werden kann. Dabei soll aber nicht verschwiegen werden, daß in vielen Fällen auch völlig falsche Positionen ermittelt wurden. Werden mit dieser Vorgehensweise Graphen aus einer Ansicht extrahiert, so eignen sich diese Graphen zwar für eine mittelmäßige Erkennleistung, ihre Geometrie ist aber völlig verzerrt.

Werden aber ganze Graphen sinnvoller Struktur miteinander gemäß der Phasenähnlichkeitsfunktion (3.5) verglichen, so zeigt sich wie von Wiskott [Wis95] beobachtet, daß der reine Amplitudenvergleich die bessere Erkennleistung erbringt. Bereits Würtz [Wür94] kritisierte, daß aufgrund der starken Frequenzunterschiede in den Komponenten eines Jets die Phaseninformation nicht überzeugend verwertet werden kann. Auch Variationen von (3.5), wie Quadrierung des Cosinus-Terms oder unterschiedliche Gewichtung der verschiedenen Frequenzen, führten nicht zu einer Verbesserung der Erkennleistung.

5 Erkennleistung auf der UMIST-Datenbank

Das Erkennsystem wurde vollständig am Beispiel der ORL-Datenbank entwickelt und erprobt. Das Herzstück, der Face Bunch Graph (FBG) besteht aus manuell definierten Graphen von Gesichtern dieser Datenbank. Es stellt sich daher die Frage, ob die Erkennleistung des Systems stark von dieser spezifischen Datenbank abhängt, beziehungsweise, ob das System auf einer anderen Datenbank ähnlich gute Resultate erbringt. In diesem Kapitel wird die Erkennleistung auf der UMIST Datenbank beschrieben.

5.1 Beschreibung der UMIST-Datenbank

Die UMIST (University of Science and Technology in Manchester) Face Database besteht aus 564 Aufnahmen von 20 Personen. Die Zahl der Aufnahmen pro Person schwankt zwischen 18 und 47. Das erste Bild zeigt die Person jeweils im extremen Rechtsprofil, welches dann graduell in eine Frontalansicht übergeht (letzte Aufnahme, siehe Abbildung(5.1)). Es handelt sich wiederum um Grauwertbilder (256 Stufen) mit einer Auflösung von 112x92 Pixeln.

Da der FBG nur aus Frontalansichten an der ORL-Datenbank erstellt wurde, stellte sich die Frage, wie das Erkennsystem nun mit einer unbekannten Datenbank zurechtkommen würde, in der ein erheblicher Teil der Gesichter in Profilansicht vorliegt.

Dies bedeutet, daß die Graphen der Gallerie zwar vom System automatisch generiert werden konnten, diese aber bei den Profilansichten "falsch" zu liegen kommen: Knoten, die im FBG Augen, Nase, Mund etc. markieren, liegen nun an völlig anderen Stellen. Diese sind jedoch nicht zufällig: Es sind die besten Übereinstimmungen, die vom Vergleichsalgorithmus gefunden wurden.

5.2 Erkennleistung bei verschiedenen Gallerien

Das System kann ein Gesicht nur dann erkennen, wenn die Pose des Vergleichsbilds der Gallerie hinreichend ähnlich ist. Wenn die Gallerie mehrere Modellge-

Abb. 5.1: Alle Ansichten der ersten Person der UMIST Datenbank. Man beachte, daß von den einzelnen Personen unterschiedlich viele Ansichten zur Verfügung standen.

sichter pro Person enthält, werden diese deshalb gleichmäßig über die verschiedenen Ansichten verteilt (d.h., als Modellgesichter werden möglichst untereinander äquidistante Aufnahmen ausgewählt, um möglichst viele unterschiedliche Posen zu berücksichtigen). Wie bei der ORL-Datenbank ist auch hier niemals das zu erkennende Bild in der Gallerie enthalten.

Die Auswertung zeigt, daß die Erkennleistung für ein Modellgesicht sehr schlecht ist (38%). Dies liegt daran, daß hier nur eine Ansicht im "Halbprofil" zum Vergleich zur Verfügung steht und somit Frontal- oder Profilaufnahmen nicht erkannt werden können. Die Erkennleistung nimmt aber bei Erhöhung der Zahl der Modellgesichter rasch zu und sättigt bei etwa 95%.

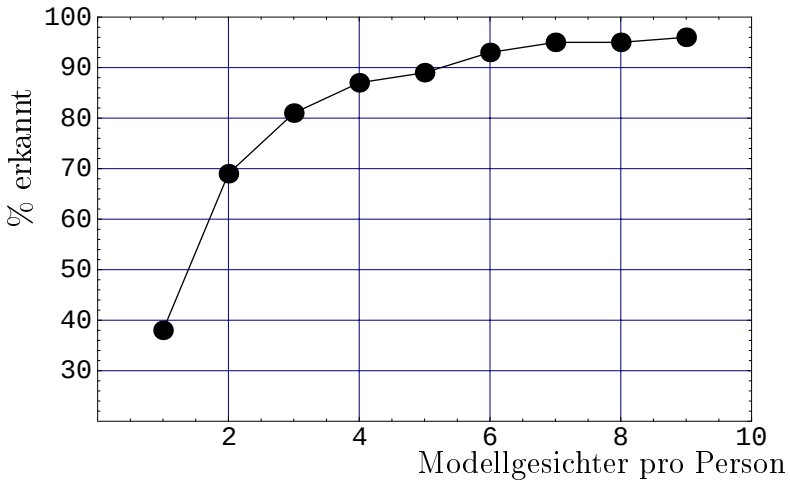

Abb. 5.2: Erkennleistung in Prozent in Abhängigkeit von der Zahl der Modellgesichter pro Person in der Gallerie.

5.3 Verwendung eines Identitäts-Bündel-Graphen(IBG)

Wie bei der ORL-Datenbank kann auch für die UMIST-Datenbank die Verwendung von IBGs erprobt werden. Hierbei ist zu beachten, daß die Datenbank Aufnahmen aus sehr verschiedenen Blickwinkeln beinhaltet. Würde man alle Graphen aus der verwendeten Gallerie zu IBGs zusammenfassen, so würden sehr verschiedene Ansichten kombiniert, was der Erkennung abträglich wäre. Daher werden in diesem Falle nur benachbarte Graphen zusammengefaßt (in der Datenbank wechselt der Blickwinkel kontinuierlich, siehe Abbildung (5.1).

Wie bei der ORL-Gallerie zeigte sich auch hier nur eine unwesentliche Verbesserung (weniger als ein Prozentpunkt) der Erkennleistung.

6 Ergebnisse

In diesem abschließenden Kapitel werden die Ergebnisse besprochen und mit den Resultaten anderer Arbeiten verglichen.

6.1 Erkennleistungen anderer Systeme

Der objektive Vergleich der verschiedenen publizierten Ansätze zur Gesichtserkennung ist schwierig, da jeweils die genauen Versuchsbedingungen verschieden sind. Zielsetzung, Computerleistung und insbesonders die verwendete Datenbank und die Zahl der Trainingsgesichter (bei neuronalen Netzen) bzw. der Anzahl von Modellgesichtern in der Gallerie machen einen allgemeinen Vergleich fast unmöglich. Daher sollen hier nur Erkennleistungen angeführt werden, die mit der bekannten ORL-Datenbank erzielt wurden.

1994 berichten Pentland, Moghaddam und Starner [Pen94] eine Erkennleistung von 90% mit einem Eigenface-Ansatz[1]. Im selben Jahr erzielt Samaria [Sam94] mit Hidden Markov Modellen (HMM)[2] unter Verwendung von fünf Trainingsgesichtern eine Erkennleistung von 95%.

Ebenfalls mit fünf Trainingsgesichtern erkennt 1997 ein neuronales Netz von Lawrence, Giles, Tsoi und Bach [Law97] 96,2% der ORL Gesichter. Zhang, Yan und Lades [Zha97] beschreiben 1997 sowohl einen Ansatz mit Eigenfaces wie auch mit elastischen Graphen, wobei jeweils Erkennleistungen von 80% erreicht werden.

Ben-Arie und Nandy [Ben98] entwickeln 1998 ein Erkennsystem, daß mit fünf Trainingsgesichtern 92,5%, mit acht volle 100% erreicht. Ihr System basiert auf einer "Volumetric Frequency Representation (VFR)", welche im Grunde aus einem Satz von zweidimensionalen diskreten Fouriertransformationen verschiedener Modellansichten besteht. Auf einem Pentium mit 200MHz werden pro Erkennung zwischen 320 und 512 Sekunden benötigt.

[1] Der bedeutsame Eigenface-Ansatz wurde von Turk und Pentland [Tur91] entwickelt. Mittels einer Hauptachsentransformation (Principal Components Analysis, PCA) werden zunächst Trainingsgesichter in Eigenfaces genannte Eigenvektoren transformiert. Diese spannen den "face space" auf. Ein zu erkennendes Gesicht wird in den face space projiziert und die euklidische Distanz zu den Projektionen der Trainingsgesichter berechnet.

[2] Hidden Markov Modelle beschreiben die statistischen Eigenschaften eines Signals.

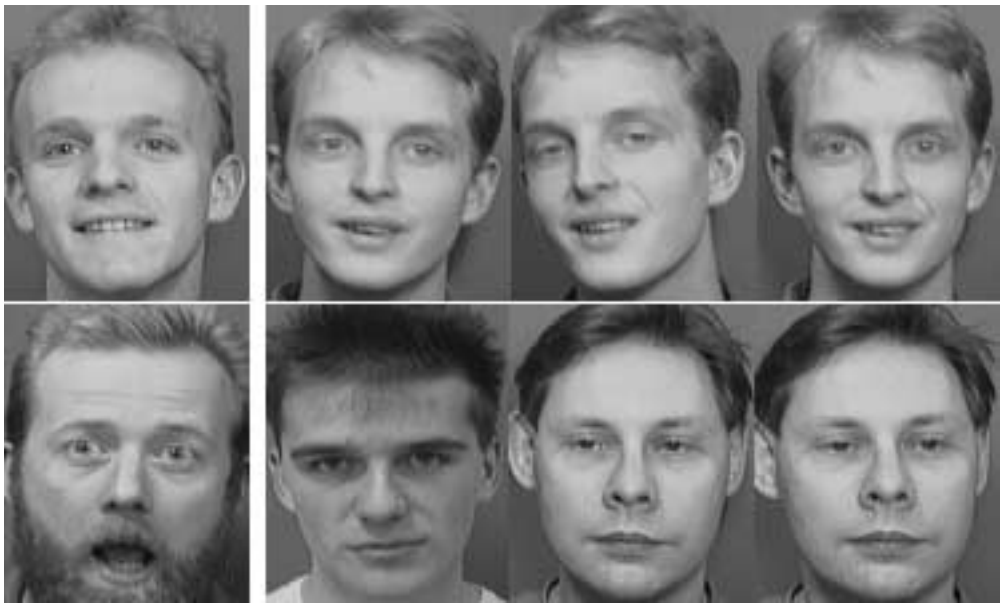

Abb. 6.1: Fehlerkennungen des Systems. Die linke Aufnahme beider Spalten sollte erkannt werden. Die jeweils drei rechts abgebildeten Gesichter wurden als ähnlichste Aufnahmen bewertet. Die Gallerie enthielt neun Modellgesichter pro Person.

Kürzlich beschrieben de Vel und Aeberhard [Vel00] ihren linienbasierten Ansatz, mit welchem sie unter Verwendung von fünf ausgewählten Trainingsgesichtern eine Erkennleistung von 99,8% erzielten. Wurden die Trainingsgesichter zufällig gewählt, so wurden noch 80,8% erkannt. Standen dem System sogar fünf Ansichten der zu erkennenden Person zur Verfügung, so wurden 100% bzw. 98% der Gesichter erkannt. Die Rechenzeit pro Erkennung liegt zwischen 0,79 und 5,0 Sekunden. Für ihr System gilt die Einschränkung, daß die Kontur des Gesichts dem System bekannt und nicht (teil-)verdeckt sein darf.

Das in dieser Arbeit beschriebene System erkennt bei Verwendung von fünf zufällig gewählten Trainingsgesichtern 94% der Bilder. Stehen dem System sogar alle fünf übrigen Ansichten einer zu erkennenden Person zur Verfügung, so werden volle 100% richtig zugeordnet. Mit einer Rechenzeit von wenigen Sekunden pro Erkennung ist es auch für realistische Anwendungen geeignet, zumal auch die Aufnahme neuer Gesichter in die Gallerie automatisch erfolgt, wobei pro Ansicht etwa 2 Sekunden benötigt werden. Als besonderer Vorzug sticht die demonstrierte Robustheit gegen teilverdeckte Bilder hervor.

6.2 Die Verwechslungen

Es wurde bereits angesprochen, daß aus den Fehlleistungen des Systems wichtige Rückschlüsse gezogen werden können. Wenn das System Gesichter verwechselt, die auch einem Menschen als ähnlich erscheinen, so darf vermutet werden, daß beide Methoden die Eingangsdaten ähnlich bewerten.

Bei neun Modellgesichtern werden 392 der 400 Gesichter richtig erkannt, in weiteren vier Fällen befindet sich die korrekte Zuordnung auf Rang 2 bis 5. Die vier anderen Fehlerkennungen sind aber besonders gravierend, zwei von ihnen sind in der Abbildung (6.1) dargestellt. Besonders die Zuordnung des bärtigen Gesichts in der zweiten Reihe enttäuscht: Obgleich die Datenbank mehrere bärtige Personen enthält, wird keines von ihnen vom System vorgeschlagen. Allerdings ist das Gesicht insoweit ausgezeichnet, daß es in der Datenbank das einzige mit weit geöffnetem Mund ist.

Abbildung (6.2) zeigt dagegen je drei Aufnahmen zweier Personen aus der UMIST-Datenbank. Während sie einem Menschen als ähnlich erscheinen, verwechselt das System sie nie.

Betrachtet man die nach unseren Maßstäben absurden Fehlerkennungen in Abbildung (6.1), so wirkt die generell zuverlässige Funktion des Systems verblüffend. Aber gerade dies läßt Chancen erkennen, die Leistungsfähigkeit weiter zu steigern, indem die offensichtlich nicht verwerteten Informationen in ein späteres (Kombinations-) System integriert werden.

Abb. 6.2: Zwei Personen der UMIST Datenbank. Obgleich sich die Person in der oberen und in der unteren Zeile nach menschlichen Maßstäben ähneln, verwechselt das System die beiden Personen nie.

Gesichtserkennung mit elastischen Graphen erweist sich als ein konzeptionell unkompliziertes Verfahren, welches sich effizient implementieren läßt und zuverlässig gute Erkennleistungen erbringt. Es ist aber klar, daß das Verfahren nur den Ausgangspunkt für ein umfassenderes Erkennungssystem bilden kann.

So ist die Behandlung und interne Repräsentation verschiedener Ansichten derselben Person im normalen wie im um Identitätsbündelgraphen (IBGs) erweiterten System nur unzufriedenstellend gelöst. Auch wird ein leistungsfähigeres System die Phaseninformation ausnutzen und zwischen verschiedenen Blickwinkeln unterscheiden müssen.

Danksagung

Zuerst möchte ich mich bei Prof. Dr. Günter Wunner für die hervorragende Betreuung bedanken. Er ermöglichte diese für ein Institut für Theoretische Physik ungewöhnliche Diplomarbeit und unterstützte mich engagiert, ließ aber auch genug Freiraum für selbstständiges Arbeiten.

Dr. habil. Stefan Luding danke ich für sein Interesse und die Übernahme des Mitberichts.

Für die bereitwillige Hilfe bei Rechnerproblemen bin ich Dirk Engel und Dr. Georg Wöste zu Dank verpflichtet.

Ein Dankeschön geht auch an die Kaffeerunde für die offenen Diskussionen auch der abwegigsten Themen. Zudem wurde ich auch beim Umgang mit LaTeX und Mathematica tatkräftig unterstützt.

Anne danke ich für das Korrekturlesen dieser Arbeit und für weiteres, für das hier kein Raum ist.

Mein besonderer Dank gilt meinen Eltern. Ich hätte es nicht besser treffen können.

Abbildungsverzeichnis

Literaturverzeichnis

[Ben98] Jezekiel Ben-Arie und Dibyendu Nandy; A Volumetric/Iconic Frequency Domain Representation for Objects with Applications for Pose Invariant Face Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 20, No 5, May 1998

[Chu92] Charles K. Chui, An Introduction to Wavelets, Academic Press 1992

[Dia77] R. Diamond und S. Carey. Developmental Changes in the Representation of Faces. Journal of Experimental Child Psychology. 231-22, 1977

[Fie87] D. Field, Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A, 4(12):2379-2394, 1987

[Fle92] D. J. Fleet, Measurement of Image Velocity. Kluwer Academic Publishers, Dordrecht, Netherlands, 1992

[Gab46] Dennis Gabor, Theory of Communication, Journal IEE 93:429-457, 1946

[Ger97] Susanne Gerl, 3D-Gesichtserkennung mit selbstorganisierendem mehrkanaligem Matching-Verfahren, VDI Reihe 10 Nr. 488, Düsseldorf: VDI Verlag 1997

[Gra98] Daniel B. Graham and Nigel M. Allison, in Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and System Sciences, Vol. 163. H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelmann-Soulie and T.S. Huang (eds), pp 446-456, 1998

[Hay82] D.C. Hay und A.W.Young. The Human Face. Normality and Pathology in Cognitive Functions, Academic Press, London, 1982

[HuW62] D.H. Hubel und T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiology (lond.) 160:106-154, 1962

[Jae93] Bernd Jähne, Digitale Bildverarbeitung, Springer Verlag 1993

[Law97] A. Lawrence, C.L. Giles, A.C. Tsoi, A.D. Bach; Face Recognition: A Convolutional Neural Network Approach, IEEE Trans. Neural Networks, vol 8, no 1, pp 98-113, 1997

[Ley79] R. G. Ley und M. P. Bryden. Hemispheric Differences in Processing Emotions and Faces. Brain and Language. 7:127-138, 1979

[Mor88,99] Hans P. Moravec, Mind children : the future of robot and human intelligence - Cambridge, Mass. : Harvard Univ. Pr., 1988.

Computer übernehmen die Macht : vom Siegeszug der künstlichen Intelligenz, Hamburg : Hoffmann und Campe, 1999

FAZ 26. Juli 2000

[Pen94] A.P. Pentland, B. Moghaddam, T. Starner, View-based and Modular Eigenfaces for face Recognition, Proc. 1994 IEEE Conf. Computer Vision and Pattern Recognition (CVPR'94), 1994

[Pre92] W.H. Press, Numerical Recipes in C, Cambridge Univ. 1992

[Sam94] F. Samaria, Face Recognition Using Hidden Markov Models, PhD thesis, University of Cambridge, Cambridge, U.K., 1994

[Tur91] M. Turk, A.P. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience, 1991, 3(1), 71-86

[Vel00] Olivier de Vel und Stefan Aeberhard, Line Based Face Recognition under Varying Pose, IEEE Trans. on Pattern Analysis and Machine Intelligenc, 2000

[Wis95] Laurenz Wiskott, Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis. PhD thesis Ruhr-Universität Bochum, 1995

[Wür94] Rolf P. Würtz, Multilayer dynamic link networks for establishing image point correspondences and visual object recognition, PhD thesis Ruhr-Universität Bochum, 1994

[Zha97] J. Zhang, Y. Yang, M. Lades; Face Recognition: Eigenface, Elastic Matching and Neural Nets, Proc. IEEE vol 85, pp 1423-1435, 1997