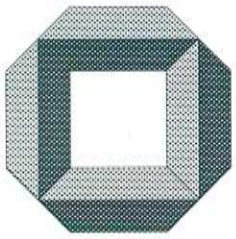


Institut für Logik, Komplexität
und Deduktionssysteme
Universität Karlsruhe

Am Fasanengarten 5
D–76128 Karlsruhe

Lokalisieren von Gesichtern mit Hilfe von neuronalen Netzen

Diplomarbeit von

H. Martin Hunke

angefertigt am

Computer Science Department
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

hunke@cs.cmu.edu

14. Juli 1994

Betreuer: Prof. Dr. Alex Waibel
Stefan Manke

Ich erkläre, daß ich die vorliegende Arbeit selbständig verfaßt und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Pittsburgh, U.S.A., den 15. Juli 1994

Inhaltsverzeichnis

Zusammenfassung

Zwischenmenschliche Kommunikation verwendet eine Vielzahl teilweise redundanter Zusatzinformationen zur gesprochenen Sprache, die die Kommunikationssicherheit wesentlich erhöhen. Bisherige Entwicklungen zur Auswertung visueller Informationen, z.B. Mimik, Augen- und Lippenbewegungen, erfordern ein stabiles Kamerabild des Sprechers und sind daher in ihrer praktischen Anwendbarkeit begrenzt. In dieser Diplomarbeit wurde ein System entwickelt, das mit Hilfe kuenstlicher neuronaler Netze eine Umgebung selbstaendig nach Gesichtern durchsucht, eines der Gesichter auswaehlt und durch Nachfuehren der Kamera und Objektivbrennweite ein stabiles Kamerabild in Realzeit auch von sich bewegenden Personen erzeugt.

Kapitel 1

Einführung

Eine Vielzahl der dem Menschen zur Verfügung stehenden Kommunikationsmittel sind für den Computer heute nicht oder nur eingeschränkt verwendbar. Künstliche Sprachsynthese, Sprach-, Gestik-, Mimik- und Handschrifterkennung sind den menschlichen Fähigkeiten weit unterlegen. Eine Weiterentwicklung auf diesem Gebiet würde die Kommunikation zwischen Computer und Mensch wesentlich effizienter gestalten und die Kommunikationssicherheit erhöhen.

Viele aktuelle Forschungsarbeiten beschäftigen sich mit der Erkennung visueller Informationen, um eine dem Menschen gerechtere Form der Kommunikation mit dem Computer zu ermöglichen. Sprecherunabhängige Spracherkennung ist ein wesentlicher Schritt in diese Richtung. Hohe Erkennungsraten sind allerdings nur bei qualititiv hochwertigen Sprachsignalen mit hohem Signalrauschabstand erzielbar. Hintergrundgeräusche führen zur Zeit zu deutlichen Leistungseinbußen. Die gleichzeitige Auswertung von Lippenbewegungen zusätzlich zur auditiven Spracherkennung erhöht die Erkennungsrate gerade in diesen Situationen, da ähnliche Laute häufig mit sehr unterschiedlichen Lippenbewegungen verknüpft sind („m“ und „n“) [2] und [5].

Das Verfolgen von Augenbewegungen ermöglicht die Berechnung eines Blickpunktes und bietet eine komfortablere Alternative zu heute üblichen Dateneingabemethoden [1]. In Kombination mit der Spracherkennung sind z.B. Textverarbeitungssysteme denkbar, bei denen die Manipulation des Textes durch Aussprache von Schlüsselwörtern bei gleichzeitigem Auswählen eines Objektes durch Augenbewegungen wesentlich vereinfacht wird. Ein Beispiel wäre das Kommando „lösche dieses Wort“ bei gleichzeitigem Betrachten eines Wortes.

Die Erkennung der Identität eines Sprechers anhand einer Abbildung des Gesichtes bietet zusätzliche Information [3], die z.B. zum automatischen Laden einer individuellen Arbeitsumgebung genutzt werden kann. Einige weitere Forschungsansätze auf diesem Gebiet finden sich in [20].

Bisherige Entwicklungen, die visuelle Informationen des Gesichtes des Kommunikationspartners verwerten, sind auf ein stabiles Kamerabild angewiesen, das das Gesicht in vorgegebener Größe, Ausrichtung und Position enthalten muß. Menschen sind jedoch gerade beim Sprechen in ständiger Bewegung, so daß die Anwendbarkeit bestehender Systeme daher stark begrenzt ist. Ein System zum Lokalisieren von Gesichtern in Kamerabildern kann den das Gesicht enthaltenden Bildausschnitt skalieren und nachfolgenden Systemen zur Verfügung stellen. Lippenlesesysteme können dann ein stabiles Bild des Sprechergesichtes in konstanter Position und Größe verwenden, obwohl sich der/die Sprecher(in) frei bewegen kann.

1.1 Motivation und Aufgabenstellung

Zielsetzung dieser Diplomarbeit ist die Entwicklung eines Systems zum Lokalisieren von Gesichtern in Kamerabildern zur visuellen Mensch–Maschine Kommunikation. Um freie Sprecherbewegungen zu ermöglichen, wurde die Fähigkeit des Sytems zur automatischen Kameranachführung und Einstellung der Objektivbrennweite entsprechend der Entfernung des Gesichtes von der Kamera mit in die Aufgabenstellung einbezogen. Das im Rahmen dieser Arbeit entwickelte System ist in der Lage, einen Raum auf die Präsenz von Personen zu untersuchen, eine dieser Personen nach noch näher spezifizierten Gesichtspunkten auszuwählen und die Kameraposition und Einstellung der Objektivbrennweite der ermittelten Position anzupassen und entsprechend der Personenbewegung laufend anzugleichen. Der das Gesicht enthaltende Bildausschnitt wird automatisch skaliert und Systemen zur Auswertung dieser Information zur Verfügung gestellt.

Die verwendete handelsübliche Videokamera CCD–TR 101 von Sony wurde auf zwei Schrittmotoren montiert, die eine horizontale und vertikale Drehbewegung ermöglichen und von einer seriellen Schnittstelle angesteuert wurden. Da die Kamera eine Einstellung der Objektivbrennweite per Fernbedienung zuließ, wurde die Kontrolle dieser Funktion nach einer Manipulation der Fernbedienung, die den Anschluß derselben an eine serielle Schnittstelle des Computers ermöglichte, dem Rechner zugänglich. Die Kamerabilder werden von einem Framegrabber in den Rechner eingelesen, der die analogen Videosignale in digitale RGB–Bitmuster umwandelt. Die anschließenden Berechnungen werden vollständig auf einer HP 9000/735 durchgeführt.
Zur zufriedenstellenden Lösung der Aufgabe muß das System in der Lage sein, in Echtzeit

- beliebige Gesichter in beliebiger Umgebung zu finden.
- ein davon automatisch ausgewähltes Gesicht durch Kameranachführung im Bild zu behalten.
- die Objektivbrennweite der Kamera einzustellen, um eine gleichbleibende Auflösung des lokalisierten Gesichtes auch bei Entfernungsänderungen zu garantieren.
- die Position und Größe des Gesichtes innerhalb des Kamerabildes zu bestimmen, so daß nachfolgenden Verarbeitungen durch Ausschneiden und Skalieren des Gesichtes ein stabiles Bild in konstanter Größe und Position zur Verfügung gestellt wird.
- Merkmale eines Gesichtes während der Lokalisierung zu lernen, um in aufeinander folgenden Bildern jeweils das gleiche Gesicht lokalisieren zu können.
- eine Anpassung auf unterschiedliche Beleuchtungssituationen vorzunehmen.

Die Auswahl eines Gesichtes könnte in Zukunft von einem Mikrofon–Array unterstützt werden, so daß Gesichter bevorzugt in Richtungen gesucht werden, aus denen ein Sprachsignal geortet wurde. Umgekehrt kann die Information über Gesichtspositionen dazu verwertet werden, in bestimmte Richtungen hineinzuhören und dadurch Hintergrundgeräusche zu unterdrücken[6].

1.2 Ansatz und Kapitelüberblick

Eine Auflistung einiger verwandter Forschungsarbeiten und Lösungsansätze findet sich im Kapitel 2.

Ein wesentlicher Aspekt der hier entwickelten Kameranachführung ist die Echtzeitfähigkeit. Um eine hohe Verarbeitungsgeschwindigkeit zu erreichen, werden die Kamerabilder in einer

situationsgebundenen, geringen Auflösung betrachtet. Im Kapitel 3 werden Eigenschaften von Gesichtern diskutiert, die vom Rechner bei der Auswertung der Kamerabilder verwendet werden können. Im wesentlichen werden die Merkmale Farbe und Bewegung verwendet. Ein Gesichtsfarbenklassifikator (FCC, von **F**ace **C**olor **C**lassifier) teilt dabei normierte Farben in Hintergrund- und Gesichtsfarben ein und paßt sich automatisch unterschiedlichen Beleuchtungsverhältnissen und Gesichtsfarben an. Durch die Abstraktion von den von der Kamera gelieferten Farbwerten auf ein Maß für Gesichtsfarbe ist es möglich, die Abhängigkeit gegenüber Hardware, Beleuchtungssituationen und Gesichtsfarben weitgehend auf den FCC zu reduzieren und eine bzgl. dieser Faktoren invariante Bilddarstellung zu entwickeln. Folglich ist es nicht mehr erforderlich, bereits trainierte, neuronale Netze nach Austausch von Hardware–Komponenten oder Veränderung der Beleuchtungssituation neu zu trainieren. Theoretische Grundlagen der Chromatik, die Konstruktion des darauf aufbauenden FCC's, die Untersuchung der Objektbewegung und die Betrachtung zusammenhängender Objekte bilden die Themen dieses Kapitels.

Das Kapitel 4 zeigt die Entwicklung einer Kameranachführung unter Verwendung der in Kapitel 3 besprochenen Merkmale. Grundlage der Gesichtslokalisierung bildet das Auffinden zusammenhängender Bereiche, die sich gegenüber dem Hintergrund bewegen und Gesichtsfarben enthalten. Da die träge, physikalische Einstellung der Kameraposition und Objektivbrennweite der Echtzeitforderung nicht genügt, wird ein virtuelles Kamerakonzept eingeführt. Dabei wird ein Teilausschnitt des Kamerabildes als virtuelle Kamera definiert, deren Größe und Position verzögerungsfrei verändert werden können.

Um zusätzlich zu den bereits verwendeten Merkmalen eine Formerkennung durchzuführen, werden künstliche neuronale Netze in das Gesamtsystem eingegliedert. Die konnektionistische Methode, mit den Farb-, Bewegungs- und Forminformationen die Position und Größe eines Gesichtes zu ermitteln, wird im Kapitel 5 vorgestellt. Mehrschichtige Perzeptronen (MLP) werden mit der Backpropagation–Methode trainiert und liefern bei Vorgabe des Kamerabildes Position und Größe enthaltener Gesichter. Es werden mehrere Netztopologien und Vorverarbeitungsmethoden vorgestellt. Durch Verwendung von FCC's können die Netze auf Farbinformationen zurückgreifen, ohne daß eine Veränderung der verwendeten Hardware, der Beleuchtungssituation oder der zu erkennenden Gesichtsfarben ein erneutes, zeitaufwendiges Training der Netze erfordern würde.

Eine Methode zur Generierung einer beliebigen Anzahl von realistischen Bildern und Filmsequenzen mit bewegten Gesichtern wird im Kapitel 6 entwickelt, die für die Bildung der Trainingsmengen für die neuronalen Netze verwendet wird.

Die Leistungsfähigkeit der in dieser Arbeit entwickelten Methoden und Systeme wird im Kapitel 7 untersucht. Anhand aufgezeichneter Filmsequenzen mit unterschiedlichen Personen und in verschiedenen Umgebungen werden Zuverlässigkeit und Genauigkeit der Lokalisierungen und Kameranachführung ermittelt. Dazu wird die Position von Gesichtern in jedem Bild der Sequenzen manuell eingegeben und mit den vom System generierten Ausgaben verglichen.

Die Kapitel 8 und 9 geben eine Übersicht über die entwickelten Methoden, die erzielten Ergebnisse und ihre möglichen Implikationen auf zukünftige Entwicklungen.

1.3 Danksagung

Mein Dank gilt meinem Betreuer Prof. Alexander Waibel für die Möglichkeit, diese Diplomarbeit an der Carnegie Mellon University durchführen zu können, die starke Unterstützung bei der Anschaffung der benötigten Hardware sowie die mir gewährten Freiräume bei der Entwicklung der Arbeit.

Für die zusätzliche Betreuung meiner Diplomarbeit von der Universität Karlsruhe aus möchte ich herzlich Stefan Manke danken.

Für die zahlreichen interessanten Diskussionen möchte ich Michael Finke danken, dessen konstruktive Stellungnahmen und Hilfsbereitschaft eine wertvolle Bereicherung darstellten.

Die Mitarbeiter des Robotikinstituts unterstützten mich mit ihrer Erfahrung bei der Auswahl der benötigten Hardware. Mein Dank gilt insbesondere Dean Pomerleau, mit ich mehrfach interessante Gespräche führen konnte.

Die Hilfsbereitschaft und Freundlichkeit der Mitarbeiter der Spracherkennungsgruppe an der Carnegie Mellon University sind mir eine große Hilfe gewesen. Insbesondere möchte ich mich bei Ricky Houghton, Arthur E. McNair und Torsten Zeppenfeld bedanken, die stets hilfreiche Ansprechpartner waren.

Mein Dank gilt auch den fast 100 Studenten und Studentinnen, die sich für Kameraaufnahmen zur Verfügung gestellt und damit wesentlich zum Gelingen dieser Arbeit beigetragen haben.

Kapitel 2

Verwandte Arbeiten

Eine Vielzahl von Arbeiten beschäftigt sich mit dem Erkennen von Merkmalen in Gesichtern. Während in frühen Arbeiten Abstandsmessungen von markanten Punkten in Gesichtern zur Erkennung der Identität verwendet wurden [16], werden in jüngerer Zeit vorwiegend neuronale Netze für diese Aufgabe verwendet. Zweischichtige, lineare Netzwerke bildeten die ersten konnektionistischen Modelle, die autoassoziativ Gesichter speichern, erkennen und restaurieren [12] konnten. Ein anderer autoassoziativer Ansatz verwendet Back–Propagation, um Bilder von Gesichtern zu komprimieren. Das Training des Netzes mit identischen Bildern als Eingabe und gewünschter Ausgabe führt zu einer komprimierten Darstellung der Bilder in der Zwischenschicht [3]. Diese kann von weiteren Netzwerken zur Erkennung des Geschlechts [8], der Identität [7] und der Mimik [3] verwendet werden. Diese Methoden operieren nur auf Bildern mit einem Gesicht in vorgegebener Position und Größe.

Lokalisierungen von Merkmalen in Gesichtern sind den in dieser Diplomarbeit betrachteten Problemen verwandter. Der Vergleich mit Durchschnittsmustern durch Template Matching ermöglicht die Lokalisierung von Augen in Gesichtern [11]. In derselben Publikation wurde dieser Ansatz mit mehrschichtigen Perzeptronen gelöst. Dazu wird ein Fenster zeilenweise über das Bild bewegt und der Fensterinhalt in ein dreischichtiges Netzwerk eingespeist. Die Ausgabe des Netzwerkes gibt an, ob im Fensterinhalt Augen detektiert wurden. Eine andere Arbeit verwendet Template Matching zum Verfolgen von Gesichtern und Substrukturen in Gesichtern, wie Augen und Mund, wobei die zu verfolgenden Bereiche interaktiv bestimmt werden [15]. Die Methode des Template Matching eignet sich zum schnellen Wiederfinden bekannter Muster.

Eine Arbeit zur Lokalisierung beliebiger Gesichter in Zeitungsbildern bildet ein errechnetes Kantenbild auf ein Gesichtsmodell ab [9]. Die Berechnung einer Hauptachsentransformation auf dem Raum aller Bilder, die ein zentriertes Gesicht enthalten, führt zu sogenannten Eigenfaces, die zur Erkennung von Gesichtern benutzt werden [18]. Dieser Ansatz ermöglicht auch eine Aussage über die Existenz eines Gesichtes in dem Bild. Durch Verschieben eines betrachteten Bildausschnittes über das Gesamtbild kann ein Gesicht auch lokalisiert werden. Eine Verknüpfung mit Bewegungsinformation erlaubt das Verfolgen von Gesichtern in nahezu Echtzeit bei einfachen Hintergründen.

Eine Kombination eines der oben beschriebenen, vergleichsweise langsamen Verfahren zur Lokalisierung beliebiger Gesichter könnte in Verbindung mit einem Template Matcher zum schnellen Verfolgen des nunmehr bekannten Gesichtes einen Ansatzpunkt zur Lösung der hier gestellten Aufgabe bieten. Das Verfolgen eines Gesichtes kann nach einer erfolgreichen ersten Lokalisierung als Wiederfinden desselben Objektes beschrieben werden. Bei diesen Lokalisierungen zum Wiederfinden kann der Suchraum auf eine Maximalumgebung um die letzte Position des Gesichtes eingeschränkt werden, die entsprechend der Objektgeschwindigkeit festgelegt werden kann.

Weiterhin wird der Objektraum der zu erkennenden Gesichter dahingehend eingeschränkt, daß das wiedergefundene Gesicht eine hinreichende Ähnlichkeit zu dem Gesicht aus dem letzten Bild aufweisen muß. Da das Verfolgen auf diese Weise als eingeschränkte Suche beschrieben werden kann, bildet eine Zweiteilung des Problems und eine Behandlung der Teilprobleme mit unterschiedlichen Methoden einen unnatürlichen Schnitt.

Die in dieser Arbeit vorgestellte Lösung führt die Suche auf die Erkennung von Merkmalen zurück, die in jedem Gesicht enthalten sind. Beim Wiederfinden des Gesichtes wird derselbe Suchalgorithmus mit eingeschränktem Suchraum und Anpassung der betrachteten Merkmale auf das zu wiederzufindende Gesicht verwendet.

Kapitel 3

Merkmale zur Gesichtslokalisierung

Welche Eigenschaften haben Gesichter, die sie von anderen Objekten unterscheiden und sind diese auch bei einer geringen Auflösung des Kamerabildes zuverlässig erkennbar? Bei Bildern mit einer hohen Auflösung führt der Mensch die Lokalisierung eines Gesichtes auf die Erkennung von Substrukturen zurück und sucht nach einem zusammenhängenden, ovalen Objekt mit Augen, Nase, Mund und Ohren. Bei geringen Auflösungen sind diese Strukturen nicht immer sichtbar, und die Verwendung der im folgenden untersuchten Merkmale geben hier zuverlässigere Informationen. In den nächsten Abschnitten werden die Merkmale

1. Farbe
2. Bewegung
3. zusammenhängendes Objekt

einzeln betrachtet und anschließend die Verknüpfung dieser Eigenschaften zur Gesichtslokalisierung und Kameranachführung erklärt.

3.1 Farbe

Die Verwendung von Farbe als Merkmal wirft folgende Schwierigkeiten auf:

- Die Umsetzung der Lichtwerte in ein analoges Videosignal in der Kamera sowie die Digitisierung im Framegrabber sind nicht linear und nicht normiert. Unterschiedliche Hardware kann zu deutlich unterschiedlichen Farbwerten führen. Es wird daher im folgenden ein Verfahren entwickelt, daß eine Farbanpassung in kurzer Zeit auf unterschiedliche Hardware ermöglicht.
- Die Gesichtsfarben unterschiedlicher Personen variieren stark; selbst bei einer einzelnen Person treten Unterschiede durch Erröten, Sonnenbräunung, usw. auf. Reflexionen der Umgebung auf dem Gesicht können zu Farbverschiebungen führen. Im folgenden wird daher eine Methode entwickelt, die eine automatische Anpassung des Systems auf alle Gesichtsfarben und sich ändernde Beleuchtungssituationen erlaubt.
- Die Verwendung von Farbwerten als Eingabe für neuronale Netze dehnt die Abhängigkeit von der Hardware und der Beleuchtungssituation auf die gelernten Gewichte des Netzes aus. Eine im folgenden entwickelte Abstraktion von den von der Kamera gelieferten Farbwerten ermöglicht es, diese Abhängigkeit zu umgehen.

In diesem Kapitel werden Methoden zur Normierung von Kamerabildern und Farbverteilungen entwickelt. RGB–Werte der Kamera werden dazu in chromatische Farben umgerechnet, die keine Helligkeitsanteile mehr enthalten. Die unterschiedlichen Farbverteilungen von Gesichtern und Hintergründen werden dazu genutzt, Klassifikatoren für Gesichtsfarben (FCC, von **F**ace **C**olor **C**lassifier) zu entwickeln, die Farben dem Gesichts- oder dem Hintergrundbereich zuordnen. Die Bildung eines generellen FCC's (GFCC) erlaubt die Lokalisierung von beliebigen, aber unbekannten Gesichtern, indem alle Hautfarben berücksichtigt werden. Nach erfolgreicher Lokalisierung kann der GFCC entsprechend dem gefundenen Gesicht auf einen individuellen FCC (IFCC) deduziert werden, so daß nur noch die tatsächlich in dem Gesicht auftretenden Farben als Gesichtsfarben klassifiziert werden. Das gleiche Verfahren erlaubt auch eine automatische Anpassung an sich verändernde Beleuchtungsverhältnisse.

3.1.1 Chromatikdiagramm

Das von der Kamera gelieferte Bild wird im Computer digitisiert und in einer Pixelmatrix abgespeichert. Jedes Pixel enthält 3 diskrete Werte für die absoluten Farbanteile an den Grundfarben Rot, Grün und Blau im Wertebereich $[0, \frac{1}{255}, \frac{2}{255}, \ldots 1]$. Pixel werden im folgenden in der Schreibweise $Q = (R, G, B)$ angegeben. Die Pixel $Q_1 = (0.1, 0.5, 0.1)$ und $Q_2 = (0.2, 1.0, 0.2)$ besitzen die gleichen relativen Farbanteile, unterscheiden sich jedoch in ihrer Helligkeit. Eine Umrechnung in chromatische Farben enspricht einer Normierung der Farbanteile bzgl. der absoluten Summe der RGB–Werte und führt theoretisch zu einer gegenüber Helligkeit invarianten Darstellung [21]. Damit läßt sich das Merkmal „Gesichtsfarbe" unabhängig von der Helligkeit spezifizieren. Durch die Nichtlinearität des CCD–Chips in der Kamera und der nicht genormten Digitisierung des analogen Videosignals im Framegrabber bleibt eine geringfügige Abhängigkeit der normierten Farben von der Farbhelligkeit erhalten. Die Farbnormierung bildet die RGB–Werte auf das Chromatikdiagramm gemäß $f : \mathbb{R}^3 \mapsto \mathbb{R}^2$ mit

$$q = (r, g) = f(R, G, B)$$

und

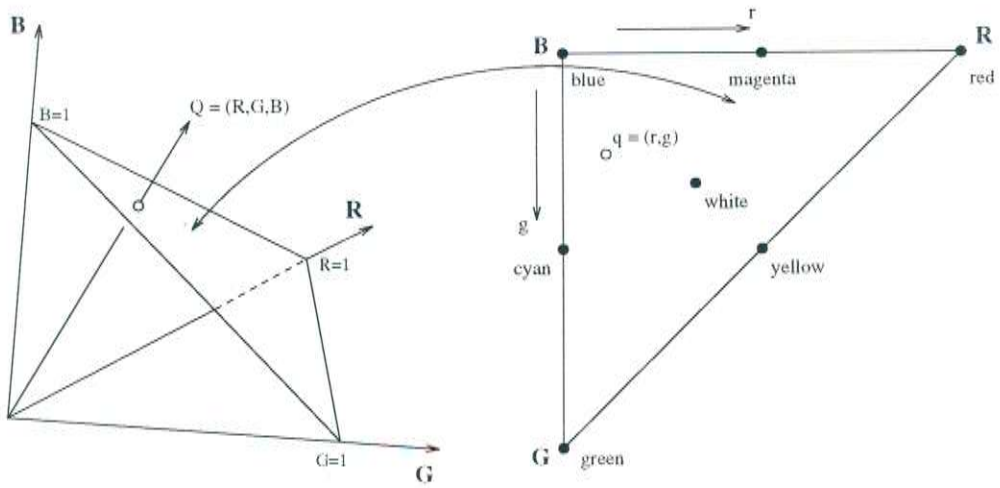
$$r = \frac{R}{R + G + B}$$
$$g = \frac{G}{R + G + B}$$

ab. Der normierte Blauwert b wird zu $r + g + b = 1$ ergänzt und ist damit redundant. Alle Farben normierter Helligkeit $q = (r, g)$ lassen sich in ein 2–dimensionales Chromatikdiagramm aufzeichnen (Bild 3.1 (a)). Im Bild (b) ist ein Beispiel der im nächsten Kapitel eingeführten Farbverteilungen abgedruckt, um den Zusammenhang mit dem Chromatikdiagramm zu verdeutlichen.

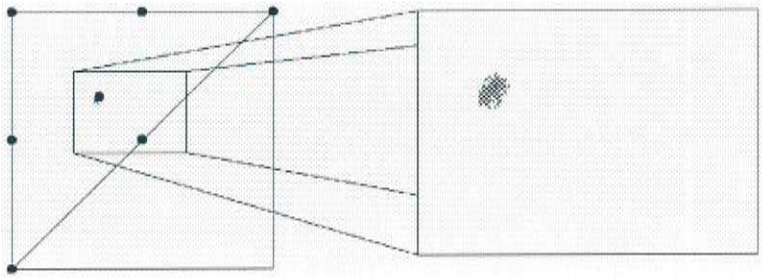
3.1.2 Farbverteilungen

Um Gesichtsfarben von Hintergrundfarben abgrenzen zu können, werden im folgenden Farbverteilungen von Bildern untersucht. Dazu werden die Häufigkeiten $N_{r,g}$ des Auftretens jeder der möglichen $\frac{256^2}{2}$ Farben $q = (r, g)$ des Chromatikdiagramms ermittelt. Die normierten Häufigkeiten $\overline{N}_{r,g}$ werden als Farbverteilung definiert:

$$\overline{N}_{r,g} = \frac{N_{r,g}}{\max_{i+j \leq 255} N_{i,j}}$$



(a)



(b)

Abbildung 3.1: Chromatikdiagramm
(a) Raum aller RGB–Werte und Chromatikdiagramm, (b) Beispiel einer Farbverteilung

In den abgebildeten Farbverteilungen sind die auftretenden Farben zur besseren Darstellung im Bereich Grau bis Schwarz repräsentiert. Farben mit $\overline{N}_{r,g} < \frac{1}{256}$, also Farben, die 256 mal seltener als die häufigste Farbe oder noch seltener auftreten, werden Weiß dargestellt.

In den Bildern 3.2 (a) und (b) sind Original und zugehörige Farbverteilung eines Kamerabildes dargestellt. Die deutlich sichtbare diagonale Linie in der Farbverteilung deutet auf einen Anteil Farben im Originalbild hin, der keinen Blauanteil enthält. Ursache ist eine von der Kamera hervorgerufene Farbverfälschung, die hier durch eine geringere Sensibilität für die Farbe Blau auffällt. Der rechte Teil der Farbverteilung zeigt eine Ausschnittsvergrößerung.

Die Bilder 3.2 (c) und (d) zeigen die gleiche Szene, unterdrücken aber in (c) alle Bereiche, die Farben enthalten, die außerhalb des markierten Bereichs der Farbverteilung (d) liegen. Die dadurch isolierte Fläche besitzt einen scharf begrenzten, intensiven Blauton, der für das Blue–Screen–Verfahren verwendet wird (s. Kapitel 6.3.2). Die deutliche Farbverschiebung der Kamera wird beim Vergleich mit dem Chromatikdiagramm in Bild 3.1 (a) sichtbar, in der dieser Blauton nahe der Farbe Weiß liegt. Zum Vergleich wird in Bild 3.2 (e) ein anderer Farbbereich (f) ausgewählt.

3.1.3 Farbverschiebungen

Die Beleuchtungsverhältnisse spielen eine entscheidende Rolle bei der Farbgebung eines Objektes. Dasselbe Objekt reflektiert bei einer Innenaufnahme mit künstlicher Beleuchtung signifikant andere Farben als bei einer Außenaufnahme. Das Sonnenlicht besitzt einen weitaus höheren

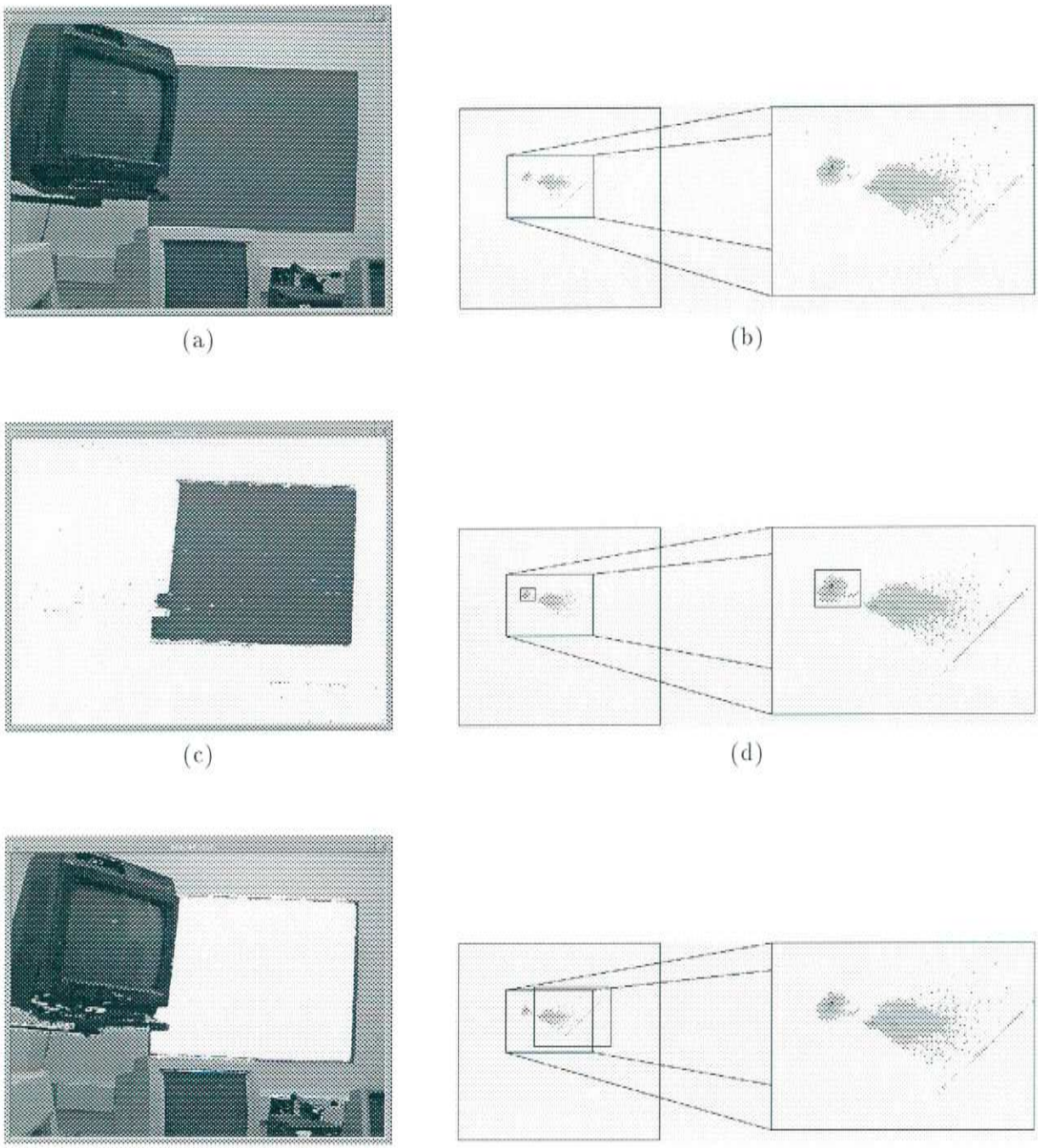


Abbildung 3.2: Kamerabild mit zugehöriger Farbverteilung
(a) und (b) Original und Farbverteilung, (c) und (d) Auswahl der blauen Farben, (e) und (f) Auswahl der nicht–blauen Farben

Blauanteil als das Licht von Neonröhren. Bei Außenaufnahmen ist der Blauanteil innerhalb von Schattenregionen noch intensiver [13]. Die Farbverteilungen sind daher bei Außenaufnahmen mehr zur Farbe Blau und bei Innenaufnahmen zur Farbe Gelb verschoben. Moderne Videokameras besitzen zum Ausgleich einen Weißabgleich, der eine meist automatische Anpassung an die Beleuchtungsverhältnisse oder eine manuelle Einstellung auf Innen- oder Außenaufnah-

men ermöglicht. Trotz Verwendung des Weißabgleichs wurden bei Innen- und Außenaufnahmen deutlich unterschiedliche Farbwerte desselben Objektes gemessen.

Unter verschiedenen Beleuchtungsverhältnissen wurde dazu dasselbe Gesicht mehrfach aufgenommen und jeweils die häufigste Farbe $\overline{p} = (\overline{r}, \overline{g})$ in der Farbverteilung gemessen. In Tabelle 3.1 sind Durchschnitt und Varianz der gemessenen Werte angegeben:

Einstellung der Kamera	Beleuchtungsverhältnisse	Mittelwert	Varianz	
Automatik	Innenaufnahmen	(113,90)	(3.7,2.7)	a_1
Automatik	Außenaufnahmen	(103,78)	(2.9,0.4)	a_2
Automatik	Innen- und Außenaufnahmen	(108,84)	(5.2,5.8)	
Innenaufnahmen	Innenaufnahmen	(107,97)	(1.1,0.7)	m_1
Innenaufnahmen	Außenaufnahmen	(80,80)	(3.1,3.1)	f
Außenaufnahmen	Innenaufnahmen	(160,88)	(8.9,4.0)	f
Außenaufnahmen	Außenaufnahmen	(112,77)	(0.1,0.5)	m_2
entprechend Beleuchtung	Innen und Außenaufnahmen	(109,89)	(2.6,9.9)	

Tabelle 3.1: Farbverschiebungen unter verschiedenen Beleuchtungssituationen

Die gemessenen Werte sind hardwarespezifisch (s. Anhang) und nur tendenziell auf andere Kameras übertragbar. Interessant sind hier nur die Verschiebungen der Mittelwerte und die Varianz bei verschiedenen Einstellungen der Kamera.

Deutlich sind die extremen Verschiebungen bei falscher Einstellung der Kamera (mit f markiert). Die Unterschiede der Mittelwerte zwischen Innen- und Außenaufnahmen sind bei manueller (m_1 und m_2) Einstellung auf die Beleuchtungsverhältnisse wesentlich geringer als bei automatischer (a_1 und a_2) Anpassung. Es liegt daher nahe, die Kamera entsprechend den Beleuchtungsverhältnissen manuell einzustellen und auf die Automatik zu verzichten. Keine der Anpassungen erlaubt es, gleiche Farben bei Innen- und Außenaufnahmen zu erzielen.

3.1.4 Gesichtsfarbenklassifikation (FCC)

Nur wenige aller möglichen Farben treten tatsächlich in Gesichtern auf. Die Farbe Blau kann in Hintergründen oder in Augen auftreten, nicht jedoch als Hautfarbe. Die Konstruktion eines FCC's wird an Bild 3.3 demonstriert.

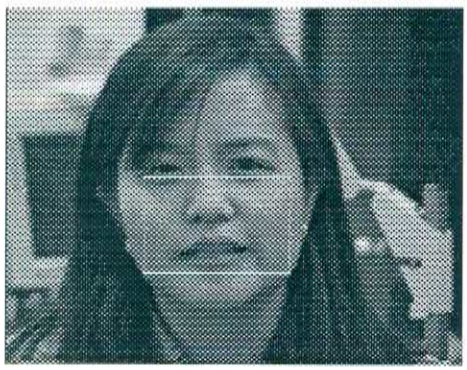
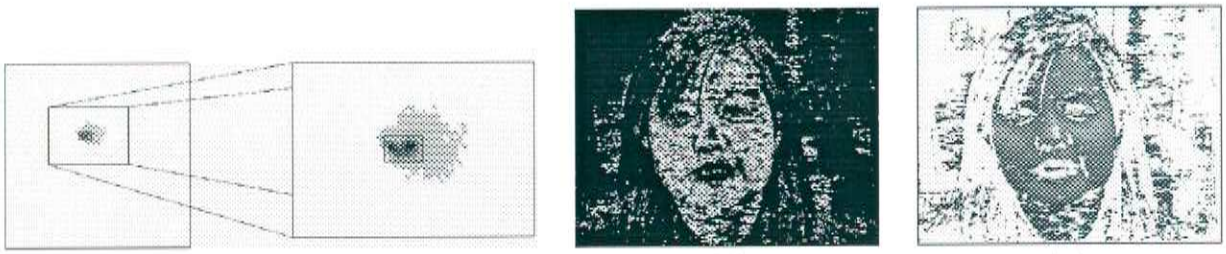


Abbildung 3.3: Gesicht

Abbildung 3.4: Bereichsauswahl

Von dem eingezeichneten Gesichtsausschnitt (Bild 3.4) wir die Farbverteilung berechnet, die als ein auf dieses individuelle Gesicht angepaßter IFCC (von **I**ndividual **F**ace **C**olor **C**lassifier)

betrachtet werden kann, der die im Beispielausschnitt häufiger auftretenden Farben als Hautfarben und alle anderen als Nicht–Hautfarben einteilt. Bild 3.5 zeigt den Farbenklassifikator und seine Anwendung. In (a) sind die Farbverteilung des Bildausschnittes und eine Vergrößerung des wesentlichen Teils der Verteilung dargestellt. Die Anwendung des IFCC's auf das gesamte Gesicht wird in (b) demonstriert. Jedes Pixel $q = (r, g)$ wird mit dem Grauwert $\overline{N}_{r,g}$ der Farbverteilung eingezeichnet. Pixel, deren chromatische Farbwerte in der Farbverteilung häufig auftreten und daher in (a) dunkel eingezeichnet sind, werden in (b) hell dargestellt und umgekehrt. Die Hautfarben konzentrieren sich auf einen eng umgrenzten Bereich des Chromatikdiagramms, der in (a) durch ein Rechteck markiert wurde. In (c) wurden alle Pixel, deren chromatische Farbwerte außerhalb des markierten Bereichs liegen, weiß dargestellt, alle anderen Pixel in ihrem ursprünglichen Grauwert. Deutlich sind die Bereiche mit Hautfarbe und die Ausblendung der Augen und des Mundes erkennbar, da diese Bereiche keine Hautfarbe enthalten. Man beachte, daß der Mund ausgeblendet wird, obwohl er in dem Bildausschnitt, der zur Bildung des IFCC's verwendet wurde, enthalten war. Die Einteilung als Nicht–Hautfarbe ist durch die geringe relative Häufigkeit des Auftretens $\overline{N}_{r,g}$ mit $(r, g) \in$ Mundfarbe möglich.



(a) (b) (c)

Abbildung 3.5: IFCC
(a) IFCC, (b) Anwendung des IFCC's, (c) Ausblendung der Nicht–Hautfarben

Einige der in diesem Vorgehen implizit enthaltenen Voraussetzungen sind nicht immer erfüllt. Die angegebenen, folgenden Abschnitte zeigen Lösungen zu den sich daraus ergebenden Problemen.

1. **Schätzung der Farbverteilung mit wenigen Pixeln**
   Der IFCC wurde mit einem Bildausschnitt bestimmt, von dem bekannt war, daß er hauptsächlich Hautfarben enthält. Wenn dieser Ausschnitt nur in einer geringen Auflösung zur Verfügung steht, reicht die Pixelzahl nicht aus, um von der Farbverteilung des Ausschnitts auf eine tatsächliche Farbverteilung des Gesichtes zu schließen.

2. **Genereller FCC**
   Die Farbverteilung eines zu lokalisierenden Gesichtes ist im Allgemeinen nicht a priori gegeben. Es muß daher ein genereller FCC (GFCC) entwickelt werden, der alle Hautfarben erkennt und sich nach Lokalisierung eines Gesichtes auf die gefundene Hautfarbe anpaßt.

3. **Berücksichtigung des Hintergrundes**
   Wenn die Farbverteilungen des Hintergrundes und des Gesichtes nicht hinreichend disjunkt sind, werden Bereiche des Hintergrundes als potentielle Gesichtsbereiche fehlklassifiziert. Bei der Bildung eines IFCC müssen die Hintergrundfarben daher mitberücksichtigt werden.

4. **Deduktion des generellen Klassifikators**
   Der zur Bildung des IFCC verwendete Bildausschnitt enthielt keine Bereiche des Hinter-

grundes oder der Haare. Wenn diese Voraussetzungen nicht gegeben sind, kann eine Konstruktion des IFCC durch Deduktion des GFCC entsprechend des vorgegebenen Auschnittes eine Verwendung der Nicht–Gesichtsfarben vermeiden, indem nur Farben als Gesichtsfarben zugelassen werden, die bereits im GFCC enthalten sind.

Schätzung der Farbverteilung mit wenigen Pixeln

In unserem vorigen Beispiel stand eine hohe Anzahl von repräsentativen Pixeln des Gesichtsausschnittes (Bild 3.4) zur Verfügung, die eine gute Schätzung der Farbverteilung des gesamten Gesichtes zuließen. In Bild 3.6 ist der gleiche Ausschnitt in einer geringeren Auflösung gezeigt. Die Berechnung und Anwendung eines IFCC's aufgrund dieser verringerten Pixelzahl ist in Bild 3.7 dargestellt.

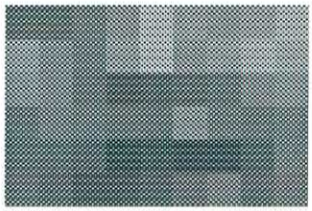
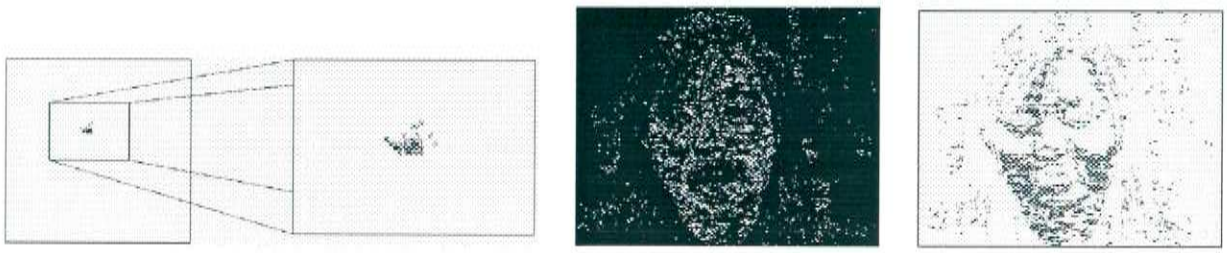


Abbildung 3.6: Vergröberter Bildausschnitt



(a) (b) (c)

Abbildung 3.7: IFCC mit wenigen Pixeln
(a) IFCC, (b) Anwendung des IFCC, (c) Ausblendung der Nicht–Hautfarben

Die geringere Helligkeit des Gesichtes nach Anwendung dieses Klassifikators in (b) gegenüber dem Klassifikator aus Bild 3.5 ist deutlich. Dies liegt daran, daß die relative Häufigkeit der Farben innerhalb des vergröberten Ausschnittes sich nicht mehr auf das gesamte Gesicht verallgemeinern lassen. In der Farbverteilung (a) ist ein Bereich markiert, der die häufigsten Farben des Ausschnittes umfaßt. In (c) sind alle anderen Farben ausgeblendet. Die nicht ausgeblendeten Farben sind in dem Bereich des Gesichtes konzentriert, der dem verwendeten Gesichtsausschnitt entspricht. Der gebildete IFCC ist offenbar nur lokal gültig.

Es liegt nahe, die Farbe jedes Pixels als Repräsentant für im Chromatikdiagramm benachbarte Farben zu betrachten. Als benachbart gelten dabei Farben, deren Abstand im Varianzbereich der Farbverschiebungen liegt (s. Abschnitt 3.1.3). Durch die Digitisierung der analogen Farbsignale innerhalb des Framegrabbers werden ähnliche Farben bereits in Gruppen gleicher digitaler Werte zusammengefaßt. Bild 3.8 zeigt einen Fehler, der durch diese Digitisierung auftritt. Obwohl die Häufung der Punkte nahe der Grenze zwischen beiden Klassen liegt, wird durch die Digitisierung Klasse **a** als wesentlich häufiger vertreten gewertet als Klasse **b**. Diese Verzerrung der Häufigkeiten kann durch einen Tiefpaßfilter verringert werden, der als lokaler Operator auf

die Farbverteilung entsprechend Bild 3.9 angewendet wird. Dieses Verfahren der digitalen Bildverarbeitung ist u.a. in [14] erläutert. Bild 3.10 zeigt in (a) die gefilterte Farbverteilung und in (b) und (c) die wesentlich verbesserten Ergebnisse.

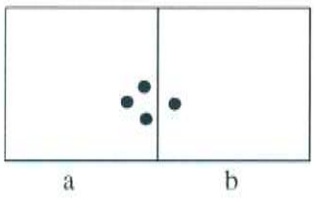


Abbildung 3.8: Digitisierungsfehler

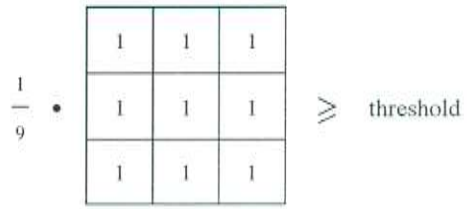
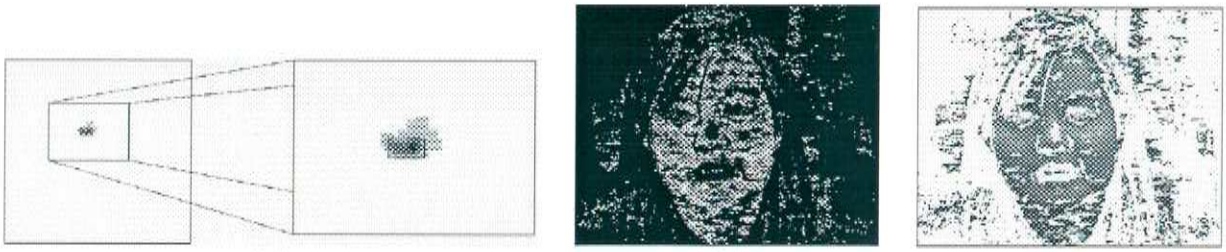


Abbildung 3.9: Tiefpaßfilter



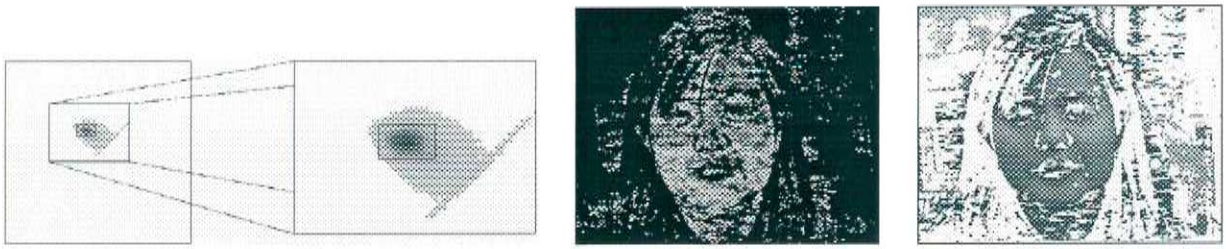
(a) (b) (c)

Abbildung 3.10: IFCC nach Tiefpaßfilterung
(a) IFCC nach Tiefpaßfilterung, (b) Anwendung des IFCC, (c) Ausblendung der Nicht–Hautfarben

Genereller FCC

Das Verfahren, einen IFCC ausgehend von einem lokalen Bildausschnitt zu bilden, der für das gesamte Gesicht gute Ergebnisse liefert, wurde im vorigen Abschnitt entwickelt. Bei vielen Problemen, z.B. dem Lokalisieren eines unbekannten Gesichtes, steht kein Bildausschnitt mit repräsentativen Farben zur Verfügung. Ein a priori Klassifikator, der alle Hautfarben erkennt, kann diese Aufgabe lösen.

Die normierten Werte aller Hautfarben liegen nahe beieinander. Bild 3.11 zeigt in (a) eine Farbverteilung, die aus Aufnahmen von 30 Gesichtern mit unterschiedlichen Hautfarben entstanden ist. Die Übergänge zwischen asiatischer, schwarzer und weißer Hautfarbe sind fließend; es entstehen keine getrennten Häufungen. Gesichter unterschiedlicher Hautfarbe reflektieren ähnliche Farben, aber in unterschiedlichen Stärken. Die Stärke der Farbreflektion wird als Helligkeit von dem FCC herausgefiltert. Der Bereich der häufigsten Gesichtsfarben (dunkel in Bild 3.11 (a)) ist eng umgrenzt. Diese einmalig für die verwendete Hardware und Beleuchtungssituation zu bestimmende Farbverteilung bildet einen generellen FCC (GFCC), der alle Hautfarben a priori erkennt und ist der einzige Teil der Kameranachführung, der bei Veränderung der Hardware angepaßt werden muß.



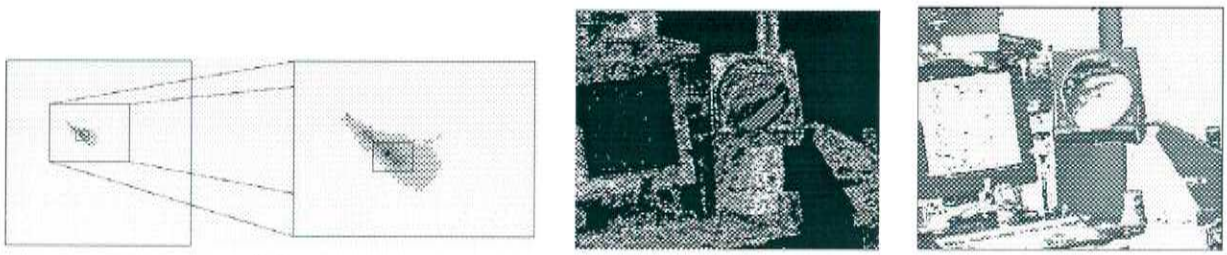
(a) (b) (c)

Abbildung 3.11: GFCC
(a) GFCC, (b) Anwendung des GFCC, (c) Ausblendung der Nicht–Hautfarben

Die Anwendung (b) des GFCC's vergrößert jedoch die Wahrscheinlichkeit, Bereiche des Hintergrundes fälschlich als Bereiche des Gesichtes zu klassifizieren. In (c) ist dies deutlich erkennbar. Dieser Klassifikator wird deshalb nur anfangs bei der Gesichtssuche verwendet. Nach erfolgreicher Lokalisierung des zuerst unbekannten Gesichtes kann die Information über die tatsächlich im Gesicht auftretenden Farben zur Bildung eines IFCC's verwendet werden. Dieses Verfahren wird im Abschnitt "Deduktion des generellen Klassifikators" beschrieben.

Berücksichtigung des Hintergrundes

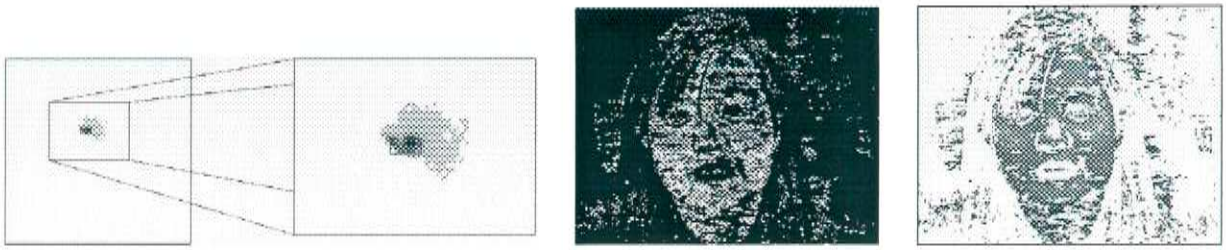
Die bisher gebildeten Farbenklassifikatoren wurden unabhängig vom Hintergrund gebildet. Enthält der Hintergrund Farben, die von einem FCC als Gesichtsfarben gedeutet werden, kommt es zu Fehlklassifikationen. Der Hintergrund des Bildes 3.4 ist in Bild 3.12 mit zugehöriger Farbverteilung gezeigt.



(a) (b) (c)

Abbildung 3.12: Hintergrund
(a) Farbverteilung des Hintergrundes, (b) Anwendung des FCC, (c) Ausblendung der Nicht–FCC–Farben

Der markierte Bereich der Farbverteilung (a) liegt in einem Bereich des Chromatikdiagramms, der auch Gesichtsfarben der bisher konstruierten FCC's enthält. Die Farbverteilungen von Hintergründen und Gesichtern können nicht als disjunkt betrachtet werden. Dies war auch schon früher an den Ergebnissen der Anwendung eines Klassifikators erkennbar. Man betrachte z.B. die hellen Bereiche des Hintergrundes in Bild 3.10 (b). Die konjunktive Verknüpfung der Farbverteilungen von Hintergrund und Gesicht läßt den FCC nur diejenigen Farben des Gesichtes als Gesichtsfarben klassifizieren, die nicht zu häufig auftretenden Hintergrundfarben gehören (Bild 3.13). Die Einschränkung der Farbverteilung (a) gegenüber der ursprünglichen Verteilung in Bild 3.5 (a) ist deutlich erkennbar.



(a) (b) (c)

Abbildung 3.13: Berücksichtigung des Hintergrundes
(a) FCCmit Subtraktion der Hintergrund–Farbverteilung, (b) Anwendung des FCC, (c) Ausblendung der Nicht–Hautfarben

Eine Ausklammerung aller Farben, die im Hintergrund vorkommen, würde eine zu starke Einschränkung des Klassifikators bedeuten. Wird eine Farbe, die selten als Hintergrundfarbe vorkommt, als Gesichtsfarbe klassifiziert, wird ein Fehler produziert, der weit geringer sein kann, als die Ausklammerung derselben Farbe, wenn sie sehr häufig im zu lokalisierenden Gesicht auftritt. Anstelle einer konjunktiven Verknüpfung wird daher eine Subtraktion der Hintergrund–Farbverteilung von der Farbverteilung des Gesichtes durchgeführt.

Deduktion des generellen Klassifikators

Durch sich verändernde Beleuchtungsverhältnisse, z.B. durch Bewegung des Gesichtes, verändert sich auch die Farbverteilung desselben. Eine fortlaufende Anpassung des IFCC's an diese Umgebungsverhältnisse ist deshalb erstrebenswert. Bei der von der Kamera gelieferten, fortlaufenden Bildsequenz kann diese Anpassung auf einfache Weise erreicht werden. Die Anwendung des IFCC teilt ein Bild in Gesichts- und Hintergrundbereiche ein. Die als Gesichtsfarben klassifizierten Pixel werden zur Bildung eines neuen IFCC's verwendet, der nur noch die Farben als Gesichtsfarben klassifizieren soll, die tatsächlich im lokalisierten Gesicht auftreten. Die Kombination des alten und neuen Klassifikators erlaubt daher eine graduelle Anpassung auf sich verändernde Lichtverhältnisse. Im nächsten Bild werden die mit dem kombinierten Klassifikator als Gesichtsbereiche klassifizierten Pixel zur Bildung des nächsten IFCC verwendet, usw.

Da jedoch nicht garantiert werden kann, daß der gefundene Gesichtsbereich keine Teile des Hintergrundes oder z.B. der Frisur enthält, kann eine Anpassung des Farbklassifikators auf diese unerwünschten Farben im nächsten Bild zu einem größeren Anteil dieser Farben führen. Es ist daher sinnvoll, bei der Anpassung des Klassifikators durch einen Bildausschnitt nur diejenigen Farben zu berücksichtigen, die als Gesichtsfarben möglich sind. Diese Information ist bereits in dem GFCC enthalten, der nur Farben berücksichtigt, die einer Hautfarbe zugeordnet werden können. Bei der Konstruktion eines neuen IFCC's durch einen Bildausschnitt werden durch Schnittmengenbildung mit dem GFCC alle Farben ausgeklammert, die nicht als Gesichtsfarben in Frage kommen.

3.2 Bewegung

Da statische Gesichter, z.B. in an der Wand hängenden Bildern, nicht berücksichtigt werden sollen, bildet Bewegung ein zur Lokalisierung eines Gesichtes notwendiges Kriterium. Ein schnelles Verfahren, um bei statischem Hintergrund Bewegung von Objekten zu ermitteln, ist die Differenzbildung von zeitlich aufeinander folgenden Grauwertbildern. Dazu wird der Abstand $\|q_1 - q_2\|$ der Grauwerte von korrespondierenden Pixeln berechnet.

Dieses Verfahren ist wesentlich schneller als die Berechnung des optischen Flusses [19], unterliegt aber folgenden Einschränkungen:

- Der Hintergrund in aufeinander folgenden Bildern muß konstant sein. Dazu darf sich weder die Kameraposition noch die Einstellung der Objektivbrennweite verändern. Bei der Lokalisierung eines Gesichtes ist diese Einschränkung unwesentlich, da die Kamera erst bei der Nachführung bewegt wird.
- Die Bewegung wird nur an den Bereichen eines Objektes erkannt, an denen sich die Helligkeit ändert. Gleichmäßige Objekte weisen nur im Umriß eine Differenz zum Hintergrund auf. Durch die geringen Auflösungen der Bilder sind die Gesichtsbereiche jedoch genügend strukturiert.
- Wenn sich ein Objekt von Position A nach Position B bewegt, ergibt das Differenzbild an beiden Positionen hohe Werte. Diese Methode läßt keinen Aufschluß über Bewegungsrichtung zu. Daher wird nur erkannt, daß ein Objekt in A *oder* B ist. Diese Information wird durch Betrachtung von Gesichtsfarbe in beiden Positionen geliefert (s. Kapitel 3.4).

Die Bewegung eines Gesichtes wird bei ersten Suche als notwendiges Kriterium verwendet. Während der anschließenden Kameranachführung liefert eine Bewegung des Gesichtes zusätzliche Information, indem sie eine Trennung des Gesichtes vom Hintergrund bei stillstehender Kamera erlaubt, wird aber nicht mehr zwingend vorausgesetzt, um auch nicht bewegte Gesichter im nächsten Bild wiederfinden zu können.

3.3 Zusammenhängende Objekte

Die bisher untersuchten Merkmale ermöglichen es, von jedem Pixel einzeln festzustellen, ob es zu einem Bereich gehört, der sich bewegt und Gesichtsfarbe enthält. Um die zu einem gemeinsamen Objekt gehörenden Pixel zusammenfassen zu können, werden die Pixel auf Nachbarschaftsbeziehungen untersucht. Im fogenden bezeichnen $x(p)$ und $y(p)$ die Koordinaten eines Pixels p.

Def. 1: Zwei Pixel p_1 und p_2 heißen *benachbart*, wenn ihre Koordinaten der Bedingung $\|x(p_1) - x(p_2)\| \leq 1 \wedge \|y(p_1) - y(p_2)\| \leq 1$ genügen.

Def. 2: Eine Menge $\mathcal{M}$ heißt *zusammenhängend*, wenn

1. $\mathcal{M} = \{\mathrm{p}\}$ oder
2. $\mathcal{M} = \mathcal{N} \cup p$, so daß $\mathcal{N}$ eine zusammenhängende Menge ist und $\exists q \in \mathcal{N} : p$ ist benachbart zu q.

Def. 3: Ein Pixel p hat das Merkmal

- *Gesichtsfarbe bzgl. eines FCC's*, wenn $\overline{N}_{r,g} \geq \alpha$ gilt, mit $p = (r, g)$ und N der Farbverteilung des FCC's.
- *Bewegung*, wenn für das korrespondierende Pixel q aus dem zeitlich vorhergehenden Bild $\|p - q\| \geq \beta$ gilt.

Def. 4: Eine Menge $\mathcal{M}$ *besitzt das Merkmal* $\mathcal{C}$, wenn alle Pixel $p \in \mathcal{M}$ dieses Merkmal besitzen.

Def 5.: Ein *Objekt* $\mathcal{O}$ mit dem Merkmal $\mathcal{C}$ ist eine zusammenhängende Menge von Pixeln mit dem Merkmal $\mathcal{C}$, so daß gilt:
$\forall p \notin \mathcal{O} : \mathcal{O} \cup p$ ist nicht zusammenhängend oder hat nicht das Merkmal $\mathcal{C}$.

Def. 6: Die *Größe* eine Objektes $\mathcal{O}$ berechnet sich als
$\max_{p_1,p_2 \in \mathcal{O}}(\|x(p_1) - x(p_2)\| \cdot \|y(p_1) - y(p_2)\|)$.

Die Schwellwerte α und β sind empirisch ermittelt und von der verwendeten Hardware abhängig. Die Suche nach einem Gesicht kann mit diesen Definitionen einfach formuliert werden:

Suche das größte Objekt mit den Merkmalen Bewegung und Hautfarbe.

Die Auswahl des größten Objektes zielt auf die Auswahl des Gesichtes, das der Kamera am nächsten steht. Die Definition eines Objektes ist sehr instabil gegenüber kleinen Veränderungen am Ursprungsbild und beinhaltet keinerlei Formerkennung. Eine Verbesserung kann durch Tiefpaßfilterung erreicht werden. Ein anderer Ansatz zur Objekterkennung wird im Kapitel über künstliche neuronale Netze besprochen.

3.4 Verknüpfung der Merkmale und Objekterkennung

Mit den Defnitionen aus Abschnitt 3.3 läßt sich die Gesichtssuche einfach beschreiben. In den Bildern 3.14 (a) und (b) sind zwei zeitlich aufeinander folgende Aufnahmen abgebildet, deren Differenz (c) Aufschluß über die zeitliche Veränderung gibt. In (d) sind alle Pixel mit dem Merkmal Bewegung hell dargestellt. Eine Vertauschung der Ursprungsbilder würde zu dem gleichen Ergebnis führen, so daß die aktuelle Position des sich bewegenden Objektes alleine aus einem Differenzbild nicht ermittelt werden kann. Sie läßt sich aber mit der Verknüpfung des Merkmals Farbe errechnen, da das Bild nur an der aktuellen Position des Gesichtes auch Gesichtsfarbe aufweist. Die Anwendung des GFCC wird in (e) illustriert. Die Verwendung des generellen Klassifikators und die Auswahl eines Hintergrundes, der viele Gesichtsfarben enthält, führt zu vielen Bereichen, die fehlerhaft als Gesichtsfarbe eingeteilt werden. Tiefpaßfilterung mit dem lokalen Operator aus Bild 3.9 ergibt die Menge aller Pixel mit dem Merkmal Gesichtsfarbe (f). Die Tiefpaßfilterung führt dazu, daß einzelne Pixel mit Gesichtsfarbe nicht berücksichtigt werden. Umgekehrt werden zwei Objekte, die nur durch ein Pixel getrennt sind, als zusammengehörig erkannt. Von der Menge aller Pixel, die beiden Merkmalen genügen (g), wird das größte Objekt ausgewählt (h). In (i) wurden diese Pixel durch ihren ursprünglichen Grauwert ersetzt.

Die Pixelmenge des größten Objektes wird dazu verwendet, aus dem generellen Klassifikator einen auf das individuelle Gesicht angepaßten IFCC zu bilden, so daß im nächsten Bild eine bessere Lokalisierung möglich ist, da die tatsächlich im Gesicht auftretenden Farben nun bekannt sind.

Ein wesentlicher Nachteil in der ausschließlichen Verwendung von Farbe und Bewegung ist die Vernachlässigung der Form. Hals, Arme und Hände werden als mögliche Gesichter interpretiert. Im Allgemeinen ist die Größe eines Gesichtes entsprechend obiger Definition größer als die anderer Objekte, aber eine Verbindung mehrerer Objekte, z.B. durch Führen einer Hand an das Gesicht, führt zur Interpretation des Gesamtumrisses als Gesicht. Eine Formerkennung ist unter anderem durch die Verwendung von neuronalen Netzen möglich und wird an entsprechender Stelle behandelt.

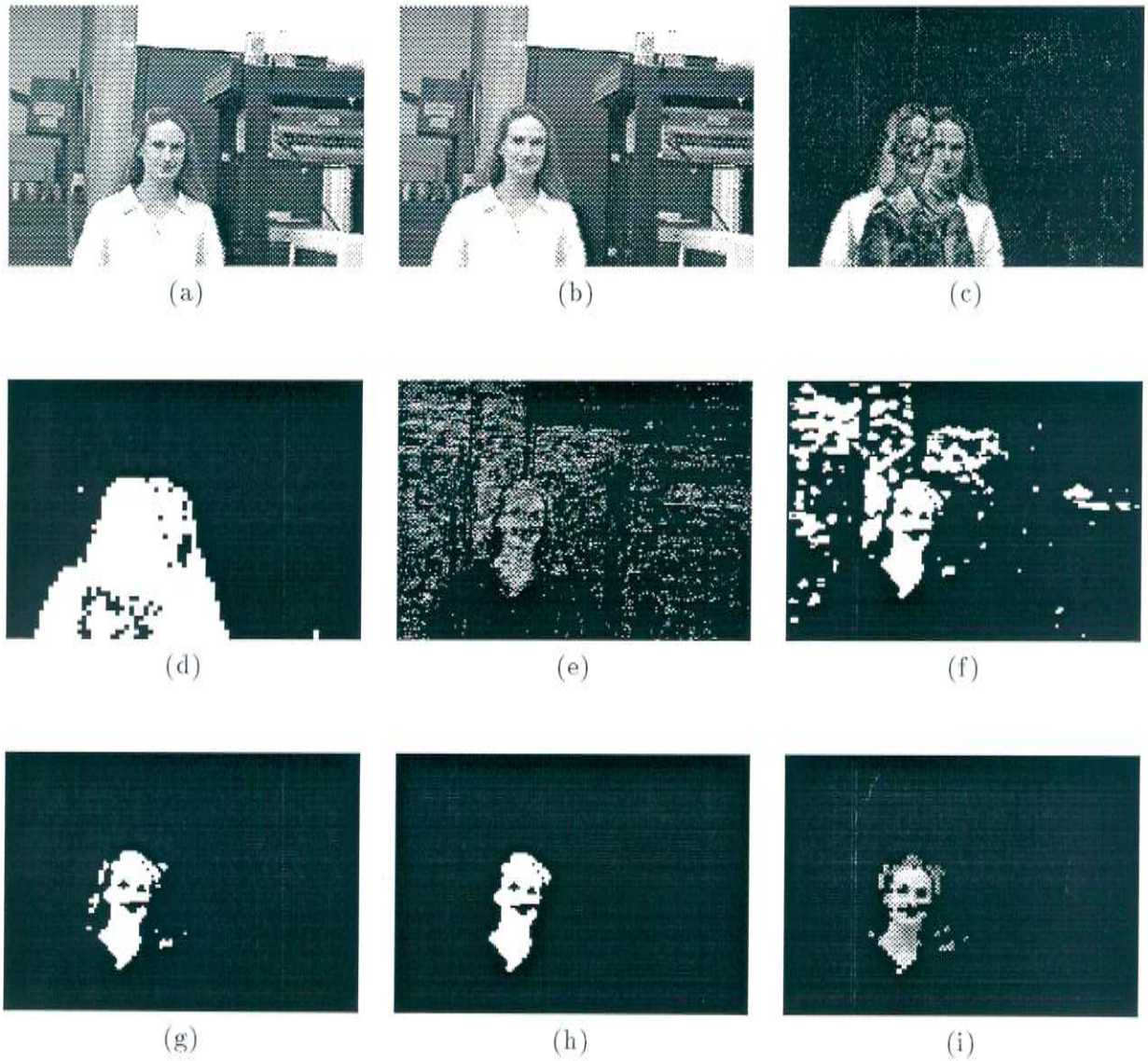


Abbildung 3.14: Lokalisierung eines Gesichtes

(a) und (b) Bilder aus einer Sequenz, (c) Differenzbild, (d) Pixelmenge mit Merkmal Bewegung, (e) Anwendung des GFCC, (f) Pixelmenge mit Merkmal Gesichtsfarbe, (g) alle Objekte mit den Merkmalen Bewegung und Gesichtsfarbe, (h) größtes Objekt, (i) lokalisiertes Gesicht

Kapitel 4

Aufbau des Gesamtsystems

Die Kamerasteuerung unterteilt sich in zwei Phasen mit unterschiedlichen Voraussetzungen:

1. *Lokalisieren eines Gesichtes*
   - Es soll das der Kamera am nächsten stehende Gesicht lokalisiert werden. Über das Aussehen des Gesichtes ist nichts bekannt.
   - Über die Position des Gesichtes im Bild ist nichts bekannt.
   - Die Kamera ist statisch und das Objektiv auf die kleinste Brennweite eingestellt, um für die Suche einen möglichst großen Blickwinkel zu haben.
2. *Nachführung der Kamera und Einstellung der Objektivbrennweite*
   - Das lokalisierte Gesicht soll im aktuellen Bild wiedergefunden werden. Das Aussehen des Gesichtes ist bekannt.
   - Die Position des Gesichtes kann als nahe der Position im vorigen Bild angenommen werden. Die Maximaldistanz hängt von der angenommenen maximalen Geschwindigkeit der Person relativ zur Kamera, der Objektivbrennweite und der Bildfrequenz ab.
   - Die Kameraposition und Objektivbrennweite werden laufend angepaßt.

Beide Phasen können trotz ihrer Unterschiedlichkeit die gleichen Verfahren aus dem Kapitel „Merkmale zur Gesichtslokalisierung" verwenden und werden nachstehend detailliert beschrieben. Es folgt eine Auflistung von Kriterien, die die Umschaltung zwischen den Phasen steuert. Eine Bildersequenz zur Demonstration des Systems schließt das Kapitel ab.

4.1 Lokalisieren eines Gesichtes

Das menschliche Auge hat nur in einem kleinen Bereich des Zentrums der Netzhaut, der Fovea centralis, eine hohe örtliche Auflösung und vermag nur in diesem Bereich Farbe zu sehen. Dieser Bereich wird deshalb auch „Ort des schärfsten Sehens" genannt. Die Randbereiche sind dafür wesentlich licht- und bewegungsempfindlicher [4]. Interessante Bereiche der Peripherie der Retina werden durch Augenbewegungen in den „Ort des schärfsten Sehens" geholt. Dies erlaubt eine wesentliche Verminderung der zu verarbeitenden Datenmenge ohne entscheidenden Informationsverlust. Das Verfahren wird im folgenden nachgebildet, indem zuerst das gesamte Kamerabild auf *interessante Bereiche* untersucht wird, und anschließend eine virtuelle Kamera

eingeführt wird, die den „Ort des schärfsten Sehens“ repräsentiert. Als *interessante Bereiche* sind alle Bereiche des Kamerabildes zu verstehen, die ein Gesicht enthalten. Von diesen Bereichen wird derjenige in den „Ort des schärfsten Sehens“ geholt, der das größte Gesicht enthält. Das größte Gesicht befindet sich der Kamera am nächsten und ist daher für die Kommunikation mit dem Rechner wahrscheinlich am wichtigsten. Andere Auswahlverfahren könnten die Richtung, aus der ein Sprachsignal zu orten ist, berücksichtigen.

4.1.1 Interessante Bereiche

Bei der Suche nach *interessanten Bereichen* werden die Merkmale Bewegung und Hautfarbe berücksichtigt. Da die Hautfarbe nicht bekannt ist, wird der Klassifikator GFCC benutzt, der alle Hautfarben erkennt. Die Bewegung wird durch Differenzbildung aufeinander folgender Bilder errechnet, da die Kamera in dieser Phase statisch ist. Die Merkmale Gesichtsfarbe und Bewegung werden konjunktiv verknüpft und auf zusammenhängende Objekte untersucht (s. Kapitel 3.4).

4.1.2 Auswahl eines Bereiches

Bild 4.1 zeigt für ein Bild alle gefundenen, *interessanten Bereiche*. Da der linke Bereich des Bildes einen hohen Anteil Farben aufweist, der vom GFCC als Gesichtsfarben erkannt werden, wurde das linke Gesicht größer eingeschätzt als es tatsächlich ist. Die Nutzung der Bewegungsinformation hilft diesen Fehler einzuschränken. Da sowohl die Position des sich bewegenden Objektes im aktuellen sowie im vorhergehenden Bild als bewegender Bereich gilt, bleibt ein Restfehler erhalten. Nach der später erfolgenden Anpassung des Farbenklassifikators auf dieses Gesicht wird der Fehler weitgehend eliminiert, da die Zahl der berücksichtigten Farben wesentlich eingeschränkt werden kann.

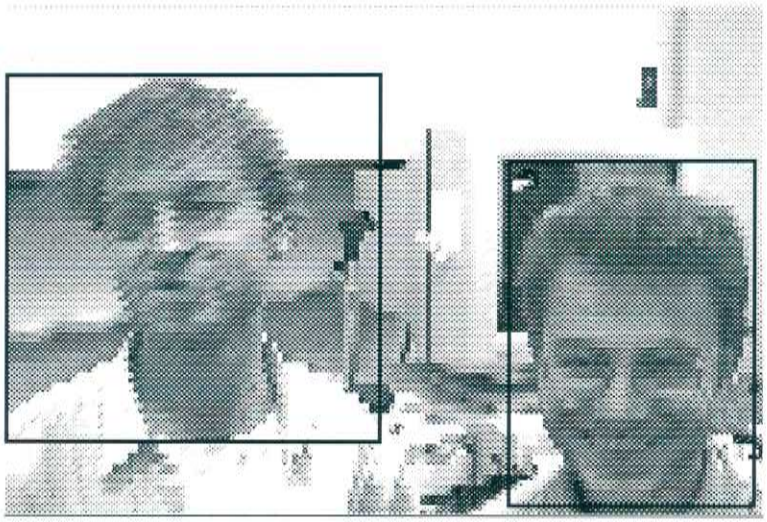


Abbildung 4.1: Lokalisierung

Um Fehler bei der Auswahl eines Gesichtes zu minimieren, werden diese Bereiche über einen Zeitraum von mehreren Bildern beobachtet. Wird das größte Objekt jeweils in naher Umgebung des größten Objektes des vorhergehenden Bildes gefunden, kann die Gesichtssuche erfolgreich abgeschlossen werden und die Phase der Kameranachführung beginnt.

4.2 Nachführung der Kamera

Die Suche eines Gesichtes im gesamten Kamerabild ist sehr zeitaufwendig. Das Wiederfinden des Gesichtes im zeitlich nächsten Bild wird deshalb auf einen Ausschnitt begrenzt, der in geeigneter Größe um die alte Position gelegt wird, ähnlich der Konzentration auf den „Ort des schärfsten Sehens" durch Augenbewegungen beim Menschen. Bei Annahme einer Maximalgeschwindigkeit einer Person und bekannter Objektivbrennweite und Bildfrequenz kann auf eine maximale Positionsänderung in zwei aufeinanderfolgenden Bildern geschlossen werden. Ein kleinerer Bildausschnitt verkürzt die benötigte Berechnungszeit um die enthaltenen Pixel zu verarbeiten, führt damit zu geringeren zeitlichen Abständen zwischen den Bildern und verringert somit die maximale Positionsveränderung, die wiederum den Ausschnitt weiter einzugrenzen gestattet. Diese sich gegenseitig verstärkenden Auswirkungen kommen bei einer minimalen Ausschnittsgröße, die hier etwa der zweifachen Kopfgröße entspricht, zum Erliegen. Da der Bildausschnitt wie eine physikalische Kamera verschoben und die Objektiveinstellung durch Größenveränderung des Ausschnitts nachgebildet werden kann, wird er als „virtuelle Kamera" bezeichnet.

4.2.1 Virtuelle Kamera

Bild 4.2 beschreibt das Konzept einer virtuellen Kamera. Während bei der physikalischen Kamera Positionsänderungen und Veränderungen der Brennweite mit zeitaufwendigen mechanischen Bewegungen gekoppelt sind, läßt sich die virtuelle Kamera verzögerungsfrei von einem Bild zum nächsten auf die gewünschten Werte einstellen. Wird die virtuelle Kamera nahe an die Begrenzung der physikalischen Kamera verschoben (im Bild als „movement zone" bezeichnet), wird diese mechanisch so bewegt, daß die virtuelle Kamera in der Mitte des Bildes zu liegen kommt. Ähnlich wird die Objektivbrennweite der physikalischen Kamera nachgestellt, falls die virtuelle Kamera eine bestimmte Größe über- oder unterschreitet. Auf diese Weise lassen sich die mechanischen Bewegungen minimieren.

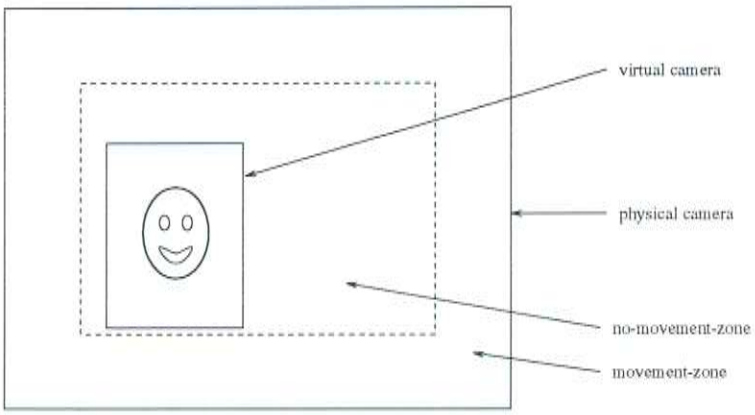


Abbildung 4.2: Virtuelle Kamera

Die Vorteile der virtuellen Kamera auf einen Blick:

- Höhere Bildfrequenz, da die Reduzierung der Datenmenge zu höherer Verarbeitungsgeschwindigkeit führt.
- Mechanische Bewegungen der physikalischen Kamera werden auf ein Minimum reduziert. Dadurch wird der Hintergrund möglichst lange konstant gehalten und kann zur Bewegungserkennung verwendet werden.

- Der mittlere Bereich der virtuellen Kamera entspricht bereits der in der Einführung geforderten Ausgabe der exakten Position und Größe des Gesichtes. Da die Positions- und Größenveränderungen der virtuellen Kamera verzögerungsfrei sind, können nachfolgende Systeme (z.B. zum Lippenlesen) jederzeit auf ein stabiles Bild zurückgreifen. Falls die Auflösung der virtuellen Kamera für die Anwendung nicht ausreicht, kann ein zweiter Framegrabber dasselbe Bild in höherer Auflösung einlesen. Die Ausgabe des Sytems bestünde dann in den relativen Koordinaten der virtuellen Kamera, die die Position des Gesichtes im Bild angeben. Mit diesem zweiten Framegrabber können die Bilder auch mit einer unabhängigen Bildfrequenz eingelesen werden. Die ausgegebenen Lokalisierungsdaten müßten dann interpoliert werden.

Die Größe und Position des lokalisierten Gesichtes im aktuellen Bild wird zur Einstellung der virtuellen Kamera verwendet. Von der Breite und Höhe des Gesichtes wird der Median von den jeweils letzten drei Werten ermittelt, um die Auswirkungen einer einzelnen schlechten Lokalisierung zu vermindern. Diese Mittelung ist begründet, da sich die Gesichtsgröße in zwei aufeinanderfolgenden Bildern nicht beliebig ändern kann. Bei der Position wurde auf eine Mittelung oder vorausschauende Berechnung verzichtet. Diese Methoden führen nur bei einer gleichmäßigen Bewegung zu einem stabileren Verhalten. In kritischen Situationen, in denen ein starker Richtungswechsel durchgeführt wird, wird eine falsche Position gemittelt oder vorhergesagt. Die virtuelle Kamera wird auf die Position des Gesichtes verschoben und in Breite und Höhe gemäß den doppelten Werten der Gesichtsgröße eingestellt.

Um Bewegungen zu erkennen, muß das zeitlich vorhergehende Bild bekannt sein. Da jedoch nicht das Gesamtbild, sondern nur die virtuelle Kamera eingelesen wird, ensteht bei einer Verschiebung derselben eine Informationslücke. Solange die physikalische Kamera in ihren Einstellungen nicht verändert wird, wird der Hintergrund als konstant angenommen und ein einmaliges Einlesen des Gesamtbildes genügt zur Bewegungserkennung. Dies wird zu Beginn der Phase *Kameranachführung* durchgeführt. In bestimmten Situationen muß dieses Hintergrundbild aufgefrischt werden:

- Die physikalische Kamera wird bewegt oder die Objektivbrennweite verändert. Die verwendete Hardware erlaubt dazu jederzeit die Abfrage, ob ein mechanischer Teil der Kamera in Bewegung ist.
- Der Hintergrund verändert sich. Wird die virtuelle Kamera in einen Bereich verschoben, dessen Hintergrund sich gegenüber dem gespeicherten Gesamtbild verändert hat, wird Bewegung nicht nur bei der Person, sondern durch die Veränderung auch im Hintergrund detektiert. Eine sehr hoher Prozentsatz veränderter Pixel deutet daher auf eine Inkonsistenz im gespeicherten Gesamtbild und veranlaßt ein neues Einlesen desselben. Dies ist insbesondere wichtig, wenn anstelle des Kamerabildes eine gespeicherte Sequenz verwendet wird, da die Information, ob sich die Kamera bewegt, nicht mehr abgefragt werden kann.

4.2.2 Kalibrierung der Kamera

Eine Verschiebung der virtuellen Kamera an den Bildrand der physikalischen Kamera erzwingt eine mechanische Nachführung, um die virtuelle Kamera wieder zu zentrieren. Die erforderlichen Verschiebungskoordinaten können relativ zum Gesamtbild angegeben werden. Die Motorik der Kameranachführung erfordert Winkelkoordinaten, deren Berechnung aus den Bildkoordinaten von der Einstellung der Objektivbrennweite abhängt. Da bei der verwendeten Hardware (s. Anhang) keine Rückmeldung über die Brennweiteneinstellung und nur eine relative Veränderung *zoom in* und *zoom out* möglich ist, wird eine Approximation durch Protokollierung der

Veränderungen durchgeführt. Zu Beginn der Phase *Lokalisierung* wird die Kamera durch längere Ausgabe des Befehls *zoom out* auf die kleinste Brennweite eingestellt. Alle folgenden Befehle werden entsprechend ihrer Länge aufsummiert, so daß die Summe in etwa der Zeit entspricht, die ausgehend von der kleinsten Objektivbrennweite der Befehl *zoom in* gegeben werden muß, um die aktuelle Brennweite zu erreichen.

Um eine Umrechnung der Bildkoordinaten in Winkelkoordinaten zu ermöglichen, wird eine Tabelle mit zehn Stützwerten erstellt, zwischen denen approximiert wird. Die Stützwerte lassen sich automatisch ermitteln, indem ein Programm die Farbe einer rechteckigen, einfarbigen Fläche mißt, die in angemessener Entfernung von der Kamera aufgestellt wird, und ausgehend von der kleinsten Brennweite eine Messung durchführt, die bei verschiedenen Brennweiten wiederholt wird. Eine Messung besteht in einer Bewegung der Kamera, bei der die entsprechende Verschiebung der beobachteten Fläche in Bildkoordinaten protokolliert wird. Die Beobachtung einer farbigen Fläche ist mit den bereits vorgestellten Methoden einfach durchzuführen. Es werden getrennte Werte für horizontale und vertikale Bewegungen ermittelt. Die Tabelle enthält schließlich für die verschiedenen Brennweiten die Winkelkoordinaten, die der Breite und der Höhe des Bildes entsprechen.

4.2.3 Wiederfinden eines Gesichtes

Nach Auswahl eines der bei der Suche lokalisierten Gesichter soll dasselbe Gesicht in allen nachfolgend von der Kamera gelieferten Bildern gefunden werden. Dazu wird im nächsten Bild der Bildausschnitt der virtuellen Kamera auf die Merkmale Gesichtsfarbe und Bewegung untersucht. Wird eine Positionsänderung der Kamera oder eine Veränderung der Objektivbrennweite durchgeführt, wird die Bewegungsinformation nicht berücksichtigt, da der Hintergrund sich mitbewegt. Wird keine Bewegung erkannt, wird nur die Information über Farbe verwendet, um auch sich nicht bewegende Gesichter erkennen zu können. Die Erkennung von Bewegung als zusätzliches, nicht notwendiges Merkmal bietet zusätzliche Information, um das Gesicht vom Hintergrund zu trennen.

4.2.4 Anpassung des Farbenklassifikators

Nach erfolgreicher Lokalisierung sollte die gewonnene Zusatzinformation über das Gesicht verwendet werden, indem der Farbenklassifikator auf die im Gesicht tatsächlich enthaltenen Farben angepaßt wird (s. Kapitel 3.1.4). Dabei werden alle zum Gesicht gehörenden Pixel, die die Merkmale Gesichtsfarbe und Bewegung aufweisen, zur Anpassung verwendet. Das gleiche Verfahren wird bei jedem neuen Bild wiederholt, um den Klassifikator sich ändernden Beleuchtungsverhältnissen anzupassen.

4.3 Das Gesamtsystem

In Bild 4.3 ist der Aufbau des Gesamtsystems illustriert. Auf das aktuelle, von der Kamera gelieferte Bild, wird ein Farbenklassifikator angewendet. Während der Lokalisierungsphase ist dies der GFCC, nachher der angepaßte IFCC. Die Bewegungsinformation wird aus den letzten beiden Bildern gewonnen. Beide Informationen werden verknüpft und das größte, zusammenhängende Objekt mit beiden Merkmalen gesucht. Bei sich bewegendem Hintergrund oder bei fehlender Bewegung des Gesichtes wird nur die Farbinformation verwendet. Die virtuelle Kamera wird entsprechend der Position und Größe des Objektes eingestellt. Erreicht diese Einstellung einen kritischen Wert, wird die physikalische Kamera bewegt oder die Objektivbrennweite angepaßt.

Die Information über das lokalisierte Gesicht, wird zur Anpassung des Klassifikators verwendet. Alternativ zur Routine „suche größtes zusammenhängendes Objekt" kann ein künstliches neuronales Netz verwendet werden, das zusätzlich auch in beschränktem Umfang eine Formerkennung durchführt. Diese Methode wird im Kapitel „Künstliche neuronale Netze" beschrieben.

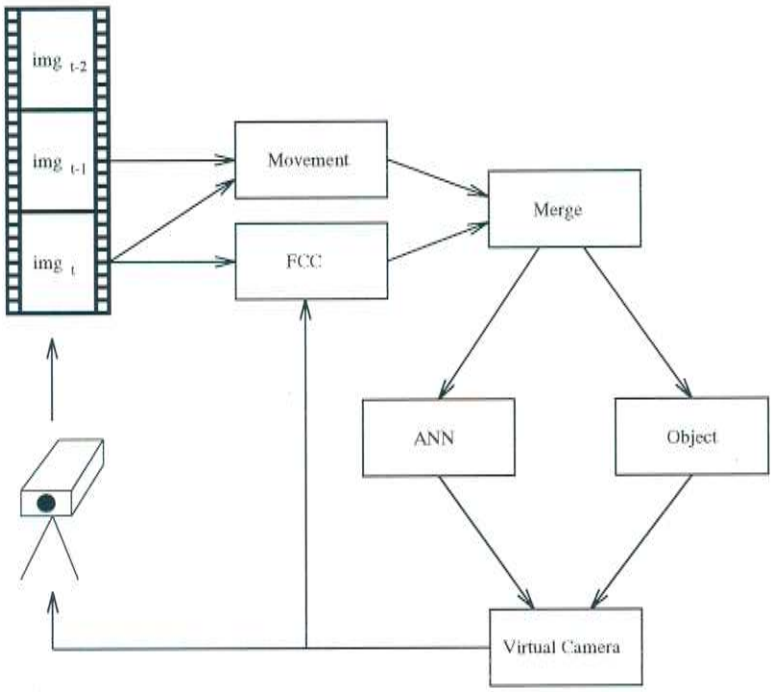


Abbildung 4.3: Gesamtsystem

4.4 Umschaltung zwischen den Phasen

Das System wird in der Phase *Lokalisierung* gestartet und schaltet nach erfolgreicher Lokalisierung auf die *Kameranachführung* um. In folgenden Situationen gilt das Gesicht als verloren und es wird auf die erste Phase zurückgeschaltet:

- Im Bild der virtuellen Kamera kann kein Objekt mit den Merkmalen Bewegung und Gesichtsfarbe lokalisiert werden.
- Das größte, gefundene Objekt unterschreitet eine Mindestgröße.
- Mehr als die Hälfte der Pixel, die zum gefundenen Objekt gehören, sind nicht im IFCC enthalten. Eine derart drastische Farbveränderung würde nur bei extremer Veränderung der Beleuchtungsverhältnisse auftreten, z.B. bei einem Wechsel von Innenbeleuchtung zu Außenbeleuchtung. Es wird daher angenommen, daß es sich nicht mehr um das gleiche Objekt handelt.

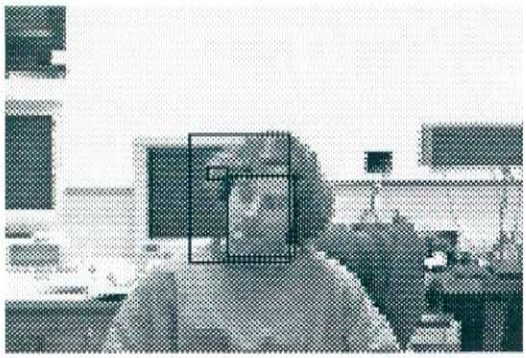
Damit einzelne Fehlklassifikationen nicht zum Umschalten zur ersten Phase führen, wird ein Zutreffen jeweils mindestens eines der Kriterien über einen Zeitraum von drei Bildern abgewartet.

4.5 Beispiel einer Kameranachführung

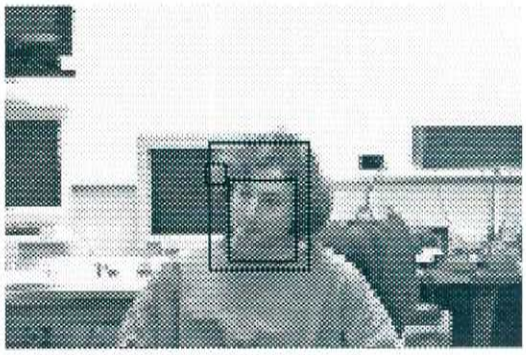
Auf den folgenden Seiten ist eine Beispielsequenz abgedruckt, die die Arbeitsweise der Kameranachführung verdeutlicht. Die erste Seite der Sequenz zeigt in Bild 4.4 (1–4) die Phase *Lokalisieren*. Der größte Rahmen umfaßt den gefundenen *interessanten Bereich*, der mittlere Rahmen wurde von Hand eingezeichnet, um das Gesicht zu markieren. Dies wird im Kapitel „Auswertung" näher erläutert. In den ersten drei Bildern tritt noch ein weiterer, sehr kleiner Rahmen auf, der einen weiteren *interessanten Bereich* markiert. Dieser entstand durch einen Teil der Frisur, der Farben des GFCC enthielt. Da in den ersten vier Bildern der größte Bereich jeweils in naher Distanz zum vorhergehenden Bild gefunden wurde, wird er als das zu beobachtende Gesicht markiert (Bild (5)).

Anschließend wird zur Phase *Kameranachführung* gewechselt. Die Rahmen zeigen der Größe nach den Ausschnitt der virtuellen Kamera, das vom Programm lokalisierte Gesicht und die manuelle Markierung. Obwohl nur die virtuelle Kamera eingelesen wird, kann das gesamte Bild durch Überlagerung mit dem statischen Hintergrundbild dargestellt werden. Der vom Programm ermittelte Gesichtsrahmen bestimmt im nächsten Bild Position und Größe der virtuellen Kamera. Sie wird zentriert um diesen Rahmen gelegt und in der Größe auf doppelte Breite und Höhe des Gesichtes eingestellt. In den Bildern (15–19) wird die sich verstärkende Inkonsistenz mit dem gespeicherten Hintergrundbild deutlich. In Bild (19) wird schließlich dieser Fehler erkannt und das Gesamtbild neu eingelesen, wie in Bild (20) erkennbar.

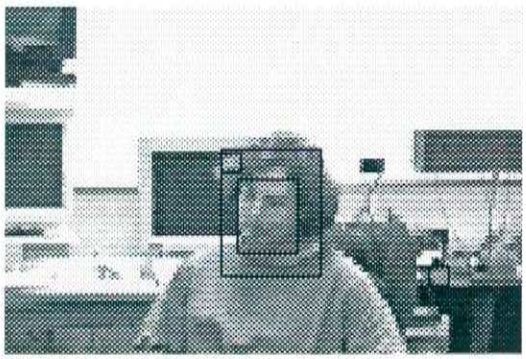
Die Bildersequenz 4.7 zeigt die Ausgabe des Sytems, das stabile Bild des beobachteten Gesichtes. Jedes Bild entstand durch automatische Ausschnittsvergrößerung des lokalisierten Bereiches. Die unter den Bildern angegebenen Nummern beziehen sich auf die Nummern der Bildsequenz. Falls die Auflösung dieser Bilder für die angestrebte Anwendung nicht ausreicht, ist ein paralleles Einlesen derselben Bilder mit einem weiteren Framegrabber und Computer in höherer Auflösung erforderlich. Die Position des Gesichtes wird dabei vom Programm zum zweiten Rechner übermittelt und das entsprechende Teilbild kann von diesem in höherer Auflösung eingelesen werden. Wird zwischen den Positionen interpoliert, läßt sich das Gesicht vom zweiten Rechner auch mit einer unabhängigen Bildfrequenz einlesen.



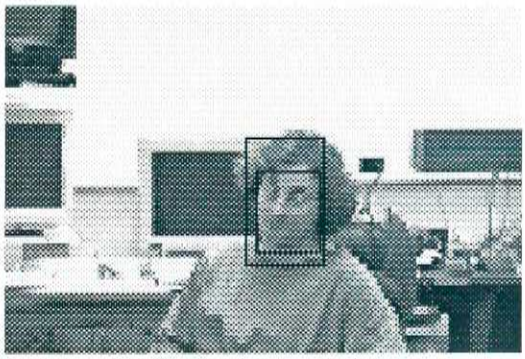
(1)



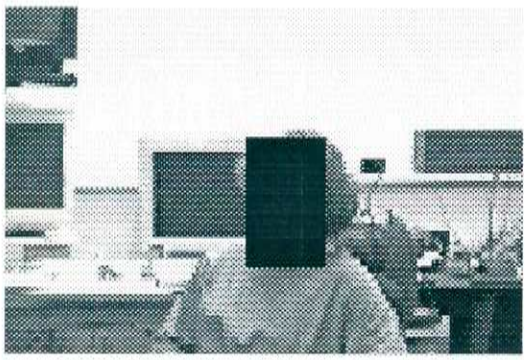
(2)



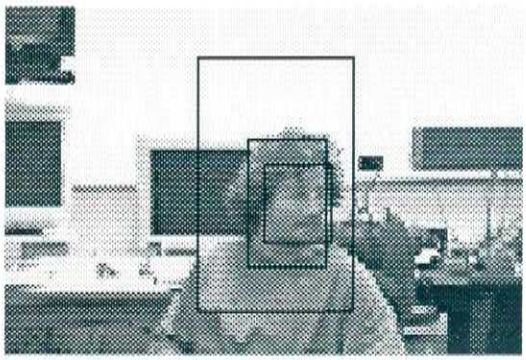
(3)



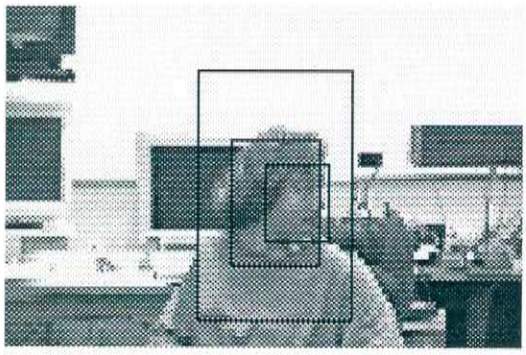
(4)



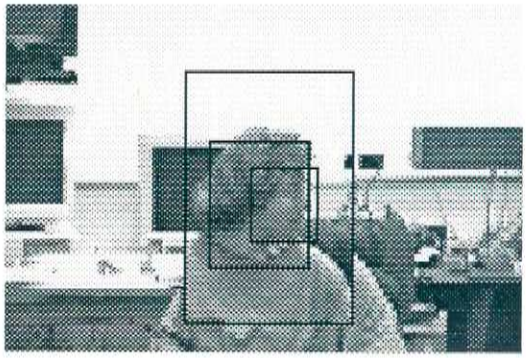
(5)



(6)

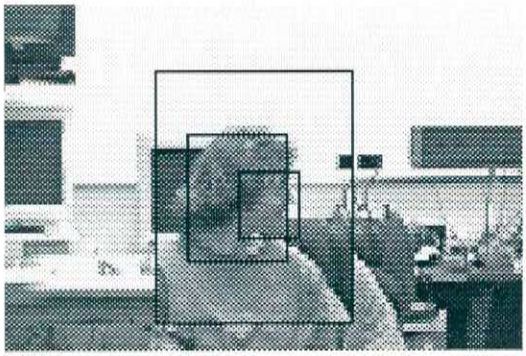


(7)

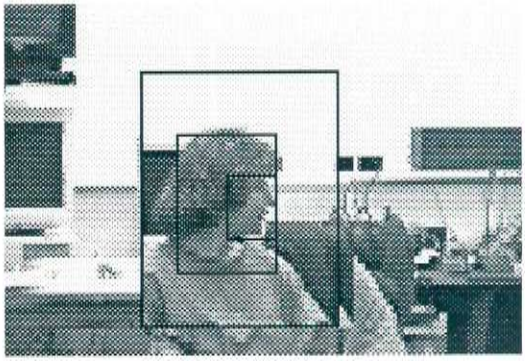


(8)

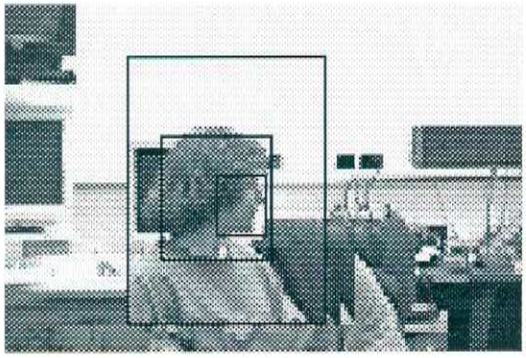
Abbildung 4.4: Beispielsequenz, Teil 1



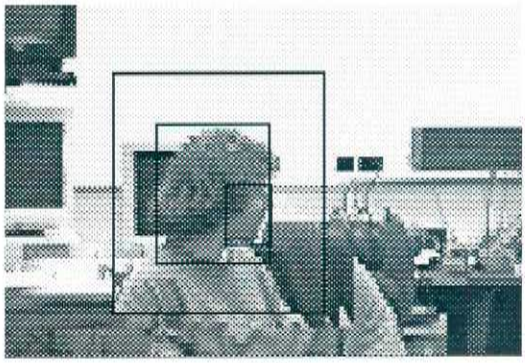
(9)



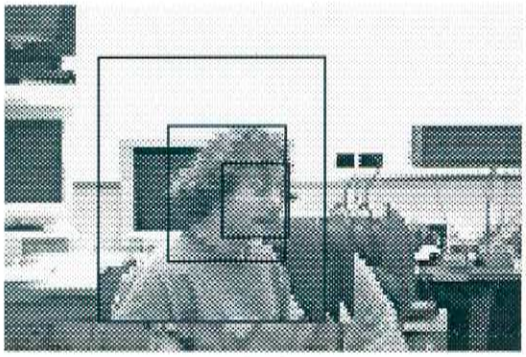
(10)



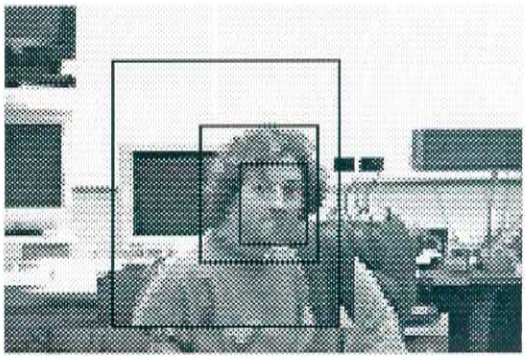
(11)



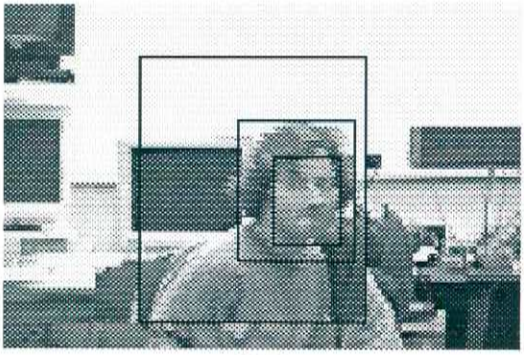
(12)



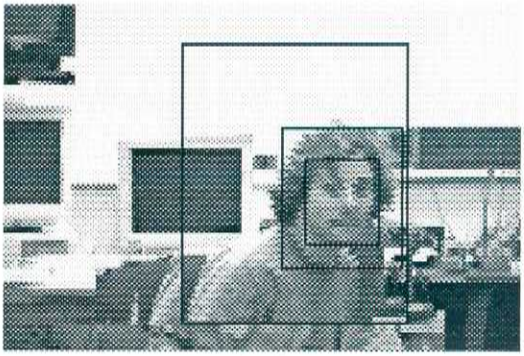
(13)



(14)

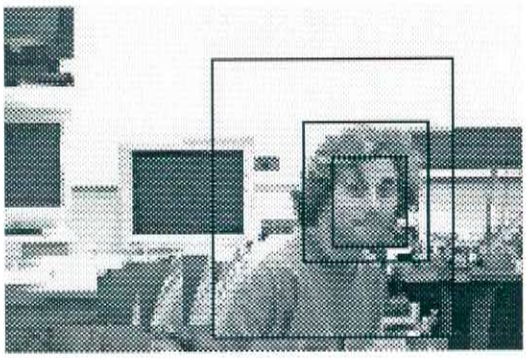


(15)

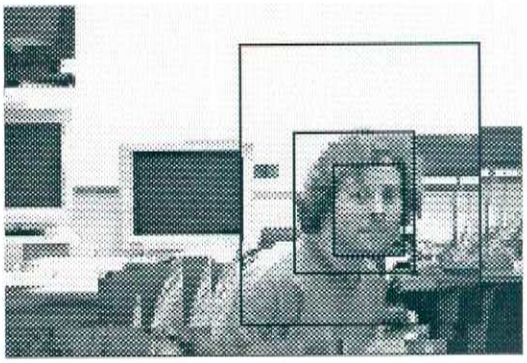


(16)

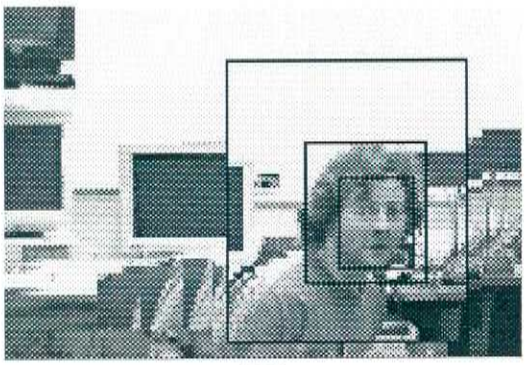
Abbildung 4.5: Beispielsequenz, Teil 2



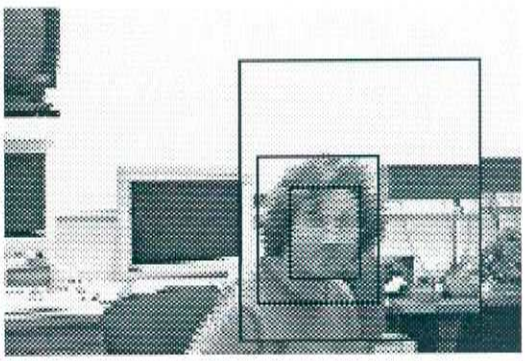
(17)



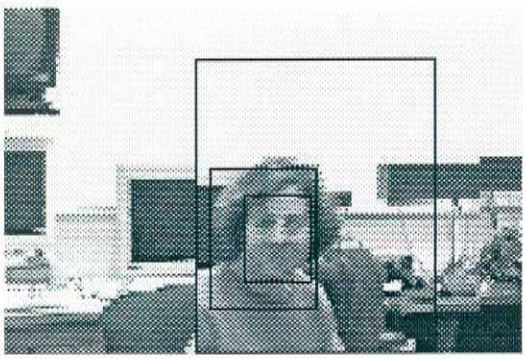
(18)



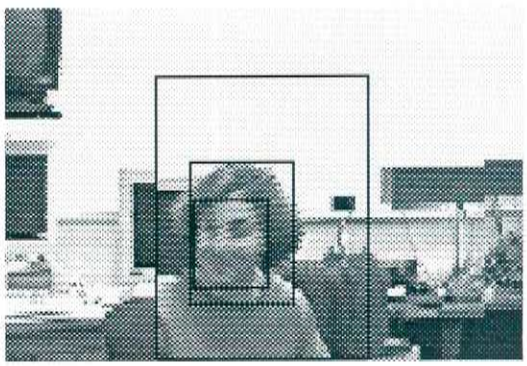
(19)



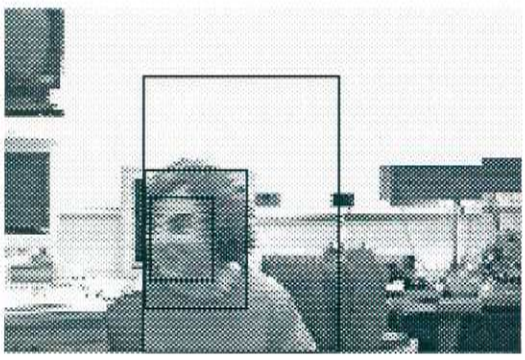
(20)



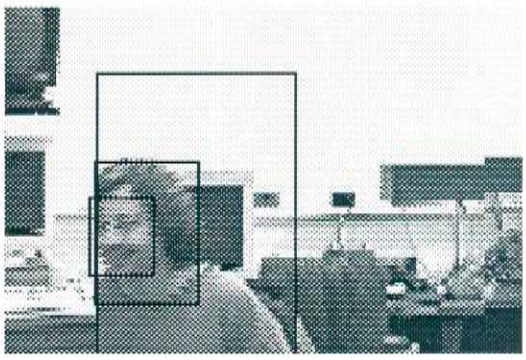
(21)



(22)



(23)



(24)

Abbildung 4.6: Beispielsequenz, Teil 3

Abbildung 4.7: Lokalisierte Gesichter

Kapitel 5

Künstliche neuronale Netze

Die Routine *suche größtes zusammenhängendes Objekt* beinhaltet keine Formerkennung des betrachteten Objektes. Alle Objekte mit den Merkmalen Gesichtsfarbe und Bewegung werden als Gesichter betrachtet. Um die Wahrscheinlichkeit einer Fehlklassifikation zu verringern, könnte eine Formerkennung z.B. Arme als nicht runde, aber längliche Objeke von einer Erkennung als Gesicht ausschließen. Die in diesem Kapitel vorgestellten neuronalen Netze werden zu diesem Zweck in das Gesamtsytem eingebunden und ersetzen oben erwähnte Routine zur Objektsuche.

Alle hier betrachteten neuronalen Netze sind mehrschichtige Perzeptronen, die mit einem Gradientenabstiegsverfahren trainiert werden, allgemein als Back–Propagation bekannt. Nach einer Einführung in die Aufgabe, Grundstrukturen und das Training dieser Netze werden die verwendeten Netztopologien im Einzelnen diskutiert. Während die Netze ohne oder mit einfacher Vorverarbeitung, die in einer Normierung der Eingabe besteht, schlechte Erkennungsraten aufweisen, die für die vorgesehene Aufgabe unbefriedigend sind, bringt die Verwendung der Farbenklassifikatoren aus Kapitel 3.1.4 nicht nur eine wesentliche Verbesserung der Erkennungsleistung, sondern auch die Unabhängigkeit des Netzes von Beleuchtungssituationen und der verwendeten Hardware, so daß eine Veränderung dieser Parameter ein erneutes Trainieren der Netze nicht mehr erfordert.

5.1 Aufgabe und Training der Netze

Alle im folgenden näher betrachteten Netze besitzen eine als Retina bezeichnete Eingangsschicht, auf die das Bild der virtuellen Kamera direkt oder nach einer Vorverarbeitungsstufe projeziert wird, eine Zwischen- und eine Ausgabeschicht. Die Ausgabeschicht enthält Informationen über Position und Größe eines innerhalb der Retina lokalisierten Gesichtes, die zur Steuerung der virtuellen Kamera verwendet werden, oder die Information, daß kein Gesicht erkannt wurde, welche für die Umschaltung zwichen den Phasen *Lokalisierung* und *Kameranachführung* des Gesamtsystems entscheidend ist.

Eine einfache Struktur eines mehrschichtigen Perzeptrons für diese Aufgabe zeigt Bild 5.1. Im ALVINN–Projekt [13] wurde gezeigt, daß eine ähnliche Netzstruktur in der Lage ist, die Lenkung eines Autos selbständig zu übernehmen.

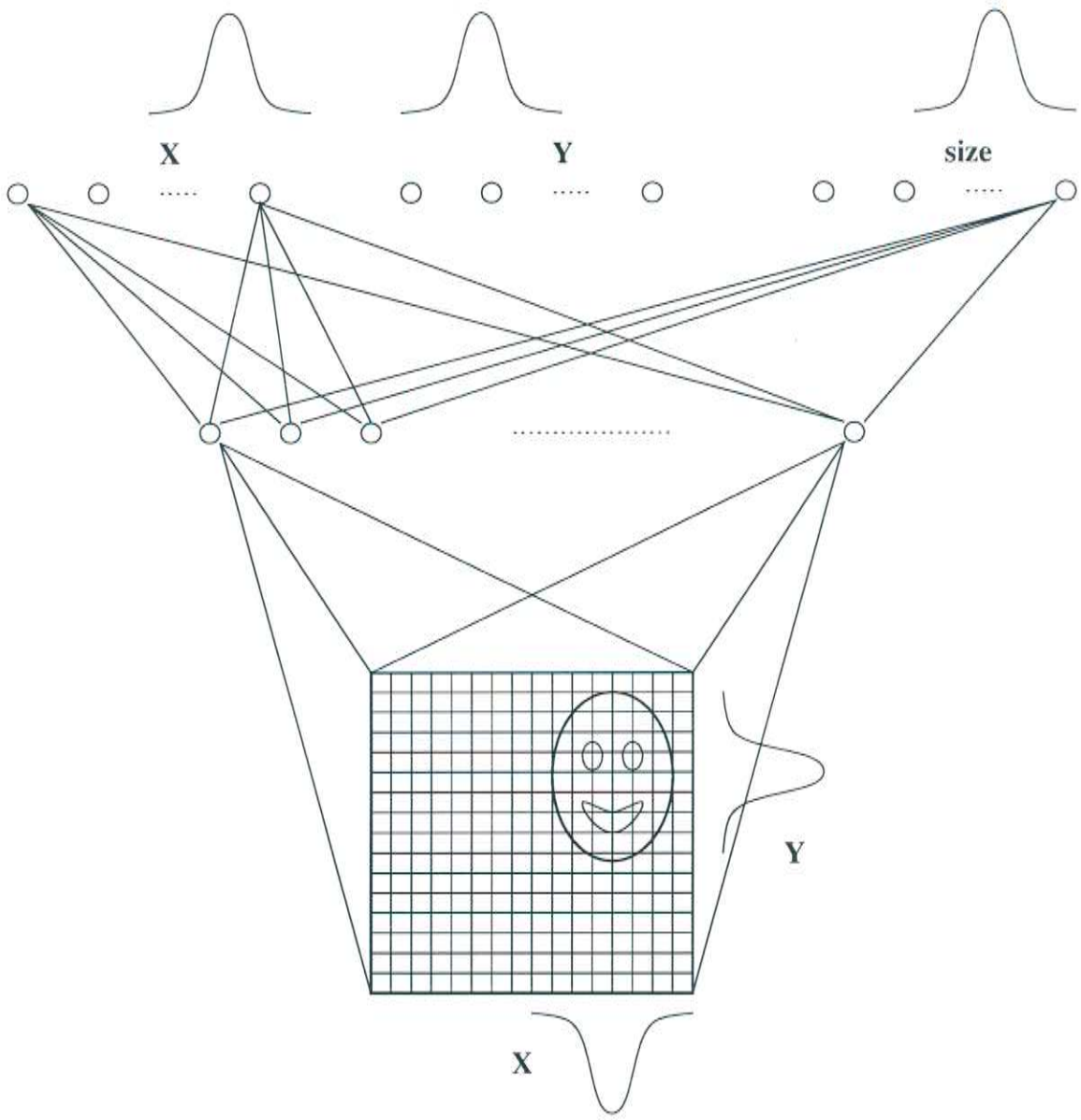


Abbildung 5.1: Künstliches neuronales Netz zur Kameranachführung

Die zum Training der Netze notwendige Erzeugung von Trainingsdaten wird im Kapitel 6.1 beschrieben.

5.1.1 Repräsentation der Eingabe

Die Aktivierungswerte der Retina bestimmen sich aus den Farbwerten der virtuellen Kamera, deren Bild auf die Größe der Retina skaliert wurde. Folgende Vorverarbeitungsmethoden wurden untersucht:

- *Normierte Grauwerte:* Die Grauwerte werden durch Mittelung der drei Farbwerte berechnet und linear so skaliert, daß der geringste Grauwert die Aktivierung -1 und der höchste den Wert 1 erhält. Diese Normierung ist notwendig, da sich die Grauwerte aller Bilder in einem kleinen Bereich konzentrieren.

- *Normierte Farbwerte:* Die Farbwerte werden auf das Chromatikdiagramm abgebildet und liefern pro Pixel zwei Aktivierungswerte $q = (r, g)$, so daß die Zahl der Eingangsneuronen verdoppelt werden muß.

- *Farbenklassifikation:* Die Farbwerte werden von einem FCC als Gesichts- oder Hintergrundfarbe klassifiziert. Die Ausgabe des Klassifikators wird linear auf den Bereich $[-1, 1]$ skaliert.

5.1.2 Repräsentation der Ausgabe

Einige grundlegende Repräsentationen werden hinsichtlich ihrer Eignung für die Lokalisierung von Gesichtern verglichen. Die folgenden Betrachtungen beziehen sich auf die idealen, beim Training vorgegebenen Ausgaben.

Einzelnes Ausgabeneuron

Diese Darstellung läßt wesentliche Informationen aus der Zwischenschicht unberücksichtigt und führt daher zu fehlerhaften Interpretationen. Erhält ein Neuron der Zwischenschicht beispielsweise Hinweise für eine Gesicht links im Bild und ein weiteres für ein Gesicht rechts im Bild, versucht das eine Neuron, die Ausgabeaktivierung klein zu halten, während das andere eine hohe Aktivierung anstrebt. Da diese Aktivierungen aufaddiert werden, ergeben sie einen Ausgangswert, der ein Gesicht in der Mitte des Bildes vermuten läßt und führen daher zu einer falschen Reaktion des Gesamtsytems.

Diskrete Verteilung

Mehrere Neuronen, die diskrete Ausgabewerte repräsentieren, ergeben zusammen die Information für einen Ausgabewert. Beim Training erhält nur das Neuron den Sollwert 1, das dem gewünschten Ausgabewert am nächsten liegt, alle anderen den Wert 0. Für die Größe eines Gesichtes könnten z.B. drei Neuronen mit den Bedeutungen *zu klein*, *optimal* und *zu groß* verwendet werden. Das Neuron mit der höchsten Ausgabeaktivierung bestimmt die Gesamtausgabe. Bleiben die Aktivierungen aller Neuronen unter einem Schwellwert, kann auf das Nichtvorhandensein eines Gesichtes geschlossen werden. Mit dieser Repräsentationsmethode wurden wesentlich schlechtere Ergebnisse erzielt als mit der nachstehend beschriebenen normalverteilten Darstellung. Dies erklärt sich durch die Möglichkeit, daß kleine Änderungen an der Eingabe eine große Änderung der Ausgabe bewirken können (Zuordnung zu einer anderen Ausgabeklasse). Dieses bei den gewünschten Ausgaben vorgegebene Verhalten kann vom realen Netz nicht nachgebildet werden.

Normalverteilte Darstellung

Anstatt die Sollwerte beim Training nach einer 1 aus N Auswahl vorzugeben, werden die Aktivitäten der Neuronen mit einer Normalverteilung berechnet (s. Bild 5.2). Die Bedeutung des Ausgabeneurons mit der höchsten Aktivierung bestimmt die Gesamtausgabe. Durch die gleitenden Übergänge führen kleine Änderungen der Eingabe auch zu kleinen Änderungen der Aktivierungen in der Ausgabeschicht.

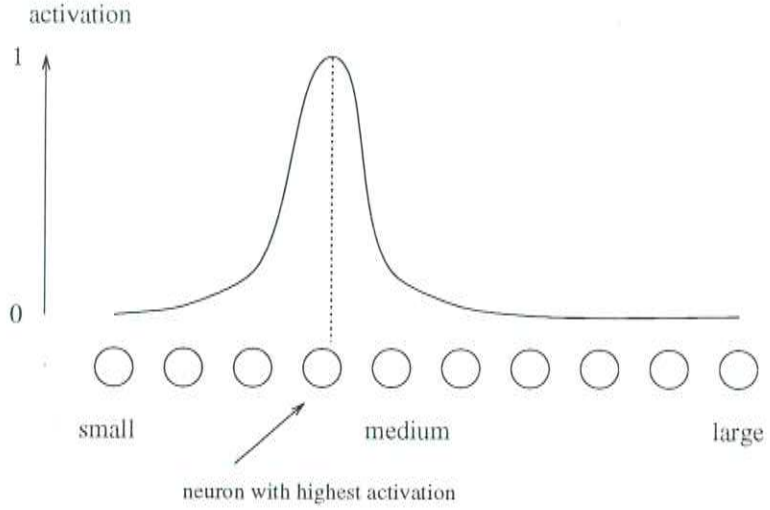


Abbildung 5.2: Neuronen eine Ausgabefeldes

In allen folgenden Netzen wurde daher die normalverteilte Darstellung gewählt.

5.1.3 Das Back–Propagation Verfahren

Eine ausführlichere Darstellung des nachstehenden Sachverhaltes findet sich z.B. in [10]. Ein Trainingsschritt besteht in der Auswahl eines Trainingsbeispieles aus einer Trainingsmenge und seiner Anwendung auf das Netz. Jedes Beispiel enthält eine Netzeingabe, die auf die Eingangsschicht kopiert und bis zur Netzausgabe durchgerechnet wird, und eine gewünschte Netzausgabe, die mit der errechneten verglichen wird und den Fehler für dieses Beispiel bestimmt. Die Gewichte des Netzes werden so verändert, daß der Fehler verringert wird. Dieses Verfahren wird mit allen Beispielen in zufälliger Reihenfolge aus der Trainingsmenge wiederholt.

Ein unabhängiges Testset zeigt nach jedem Durchlauf der Trainingsmenge die Generalisierungsfähigkeit des Netzes an. Das Training wird abgebrochen, wenn das Netz beginnt, die gezeigten Trainingsbeispiele auswendig zu lernen. Die Leistungsfähigkeit des Netzes wird anschließend von einem weiteren unabhängigen Validitätsset ermittelt (Crossvalidation, s. [17]).

Das neuronale Netz besteht aus in mehreren Schichten angeordneten Neuronen. Jedes Neuron j ist auf der Eingangsseite mit den Ausgängen o_k aller Neuronen k der vorherigen Schicht über Gewichte w_{kj} verbunden und erhält dadurch eine Aktivierung a_j gemäß 5.1. Der Ausgangswert o_j dieses Neurons wird nach 5.2 berechnet.

$$a_j = \sum_k o_k w_{kj} - \Theta_j \tag{5.1}$$

$$o_j = f(a_j) \tag{5.2}$$

Je nach Vorzeichen des Gewichtes kann eine Verbindung verstärkend oder hemmend wirken. Der Schwellwert Θ_j kann als zusätzliches Gewicht betrachtet werden, das zu einem Neuron mit konstantem Ausgangswert von -1 führt. Als Funktion f wird die sigmoide Funktion

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.3}$$

verwendet. An die Eingänge der Neuronen in der untersten Schicht werden Trainingsbeispiele gelegt und die Aktivierungen und Ausgänge der Neuronen der nachfolgenden Schichten sukzessiv berechnet. Die Ausgänge o_j der letzten Schicht werden mit zu den Trainingsbeispielen gewünschten Assoziationen d_j verglichen und ein der Differenz entsprechender Fehler, hier der Mean Square Error, berechnet:

$$E = \frac{1}{2}\sum_j (o_j - d_j)^2 \tag{5.4}$$

Um die Differenz zwischen den gewünschten und den tatsächlichen Ausgängen der Neuronen der letzten Schicht zu verringern, werden ein Fehlergradient berechnet und die Gewichte aller Verbindungen entsprechend geändert:

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial o_j} \cdot \frac{\partial o_j}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{kj}} \tag{5.5}$$

Die einzelnen Gradienten lassen sich leicht berechnen:

$$\frac{\partial E}{\partial o_j} = o_j - d_j \tag{5.6}$$

$$\frac{\partial o_j}{\partial a_j} = f'(a_j) = o_j \cdot (1 - o_j) \tag{5.7}$$

$$\frac{\partial a_j}{\partial w_{kj}} = o_k \tag{5.8}$$

Damit ergibt sich:

$$\frac{\partial E}{\partial w_{kj}} = (o_j - d_j) \cdot o_j \cdot (1 - o_j) \cdot o_k = \delta_j \cdot o_k \tag{5.9}$$

Dabei gibt der Fehlergradient δ_j an, wie stark sich eine Änderung der Eingangsaktivierung des Neurons j auf den Gesamtfehler auswirkt.

$$\delta_j = \frac{\partial E}{\partial a_j} \tag{5.10}$$

Der Anteil, den ein Gewicht w_{kj} zum Gesamtfehler beiträgt, hängt somit vom Fehlergradienten δ_j und vom Ausgabewert o_k des darunterliegenden Neurons ab, das die Aktivierung des Neurons j mitbestimmt.

Um den Fehlergradienten δ_j auch für Neuronen einer tiefer liegenden Schicht bestimmen zu können, wird der Teilterm 5.6 aus den Fehlern der Neuronen der höheren Schicht berechnet:

$$\frac{\partial E}{\partial o_j} = \sum_h \frac{\partial E}{\partial a_h} \cdot \frac{\partial a_h}{\partial o_j} = \sum_h \delta_h w_{jh} \tag{5.11}$$

Durch rekursives Anwenden der Gleichungen läßt sich der Fehlergradient für jedes Neuron berechnen. Die Gewichte werden in Richtung des negativen Gradienten verändert:

$$\triangle w_{kj}(t+1) = -\eta \frac{\partial E}{\partial w_{kj}(t)} \tag{5.12}$$

Im Allgemeinen wird zusätzlich ein Momentum α verwendet, das zu jedem Trainingsschritt einen Teil der letzten Gewichtsänderung aufaddiert.

$$\triangle w_{kj}(t+1) = \alpha \cdot \triangle w_{kj}(t) - \eta \frac{\partial E}{\partial w_{kj}(t)} \tag{5.13}$$

Dadurch wird bei flachem Gradienten die Konvergenzeigenschaft des Verfahrens wesentlich verbessert.

Nach obigen Gleichungen wird in zufälliger Reihenfolge mit Trainingsbeispielen verfahren, bis der Fehler auf einen akzeptablen Wert gesunken ist. Dabei wird die Lernrate η im Verlauf des Trainings verringert und das Momentum α bis auf einen Wert von 0.9 erhöht. Wie bei allen Gradientenabstiegsverfahren besteht die Gefahr, daß der Fehler nur bis zu einem lokalen Minimum verringert wird. Deshalb wird das Netz mehrfach mit jeweils zufällig initialisierten Gewichten trainiert und das beste Ergebnis verwendet.

5.2 Neuronale Netze ohne FCC

In Tabelle 5.1 werden die höchstenerzielten Erkennungsraten für die Netzstruktur aus Bild 5.1 aufgelistet. Die Erkennungsrate ergibt sich aus dem prozentualen Anteil der richtig klassifizierten Bilder aus dem Validitätsset, das 3000 Bilder von Personen enthält, die nicht in den Bildern enthalten waren, die während des Trainings verwendet wurden. Ein Bild gilt als richtig klassifiziert, wenn die geschätze Position des Gesichtes in jeder Koordinate um maximal $\pm 10\%$ und die Größe um maximal $\pm 20\%$ der Gesichtsgröße von den tatsächlichen Daten abweicht. Alle Netze wurden mit Gesichtern an verschiedenen Positionen und Größen von $\frac{1}{4} \ldots \frac{1}{2}$ der Retinahöhe trainiert. Einige Beispiele für Grauwertbilder in dieser Auflösung sind in Bild 6.6 (a) abgebildet. Folgende Netze wurden mit unteschiedlich großen Trainingsmengen von bis zu 30000 Bildern trainiert und erbrachten die besten Ergebnisse mit jeweils 40 Neuronen in der Zwischenschicht:

1. *Netz 1:* Das Training mit *normierten Grauwerten* als Eingabe ergab ein sehr schlechtes Konvergenzverhalten.
2. *Netz 2:* Um die Anzahl der Gewichte zu verringern, wurde die Retina für dieses Netz verkleinert. Die Aufgabe des Netzes wurde dahingehend vereinfacht, daß nur die Größe eines Gesichtes ausgegeben wird. Dazu wurde es nur mit Gesichtern trainiert, deren Position um maximal $\pm 50\%$ der Gesichtsgröße vom Mittelpunkt der Retina abwichen.
3. *Netz 3:* Wie Netz 2, jedoch mit *normierten Farbwerten* als Eingang.

Netz	Retina	Ausgabe	Anzahl Gewichte	Erkennungsrate
1	32x32	X, Y, Größe	40000	32%
2	24x24	Größe	23000	65%
3	2x24x24	Größe	46000	68%

Tabelle 5.1: Neuronales Netz ohne FCC

Die überraschend geringe Verbesserung bei Verwendung von farbigen Eingangsbildern kann nur durch die Redundanz der Zusatzinformation oder der Unfähigkeit des Netzes, die Zusatzinformation auszuwerten, erklärt werden. Daß Farbinformationen zum Finden eines Gesichtes

wenig hilfreich sind, erscheint unwahrscheinlich. Offenbar ist das Netz mit der hohen Informationsmenge der Eingangsbilder überfordert. Diese Erklärung wird vom Konvergenzverhalten des Netzes unterstützt. Wird das Netz mit einer kleinen Trainingsmenge mit weniger als 10000 Beispielen trainiert, werden die Beispiele auswendig gelernt und nur eine geringe Generalierungsfähigkeit entwickelt. Bei größeren Trainingsmengen nimmt diese zu, das Netz konvergiert jedoch wesentlich schlechter oder gar nicht mehr.

5.3 Neuronale Netze mit FCC

Selbst wenn das Netz mit den *normierten Farbwerten* akzeptable Ergebnisse gezeigt hätte, wäre es mit einem wesentlichen Nachteil verbunden gewesen. Die Abhängigkeit der Farbwerte von der verwendeten Hardware wäre auf die Gewichte des Netzes ausgedehnt worden, so daß bei einer Veränderung der Hardware das Netz neu trainiert werden müßte. Die in Tabelle 3.1 demonstrierte Abhängigkeit von der Beleuchtungssituation müßte von dem Netz gemeistert werden oder es wären für verschiedene Beleuchtungsverhältnisse unterschiedliche Netze erforderlich.

Diese Probleme treten bei Verwendung eines FCC's nicht auf, da alle Abhängigkeiten von den RGB–Farbwerten in der Vorverarbeitung zusammengefaßt werden. Das Netz wird als Alternative zu der Prozedur *„suche größtes zusammenhängendes Objekt"* an der in Bild 4.3 beschriebenen Stelle eingebunden. Dadurch erhält es anstelle der *normierten Farbwerte* die Verknüpfung der vom FCC und der Bewegungsanalyse gelieferten Informationen. Einen Vergleich der Trainingsbilder mit und ohne FCC ermöglichen die Bilder 6.6 (a) und (b). Alle Abhängigkeiten von der Hardware und der Beleuchtungssituation können somit auf den FCC übertragen werden.

Im Gegensatz zur einfachen Suche nach einem zusammenhängenden Objekt führt das Netz auch eine Formanalyse durch, die in einigen Situationen entscheidende Verbesserungen erzielt (s. Kapitel 7.2.2). Dazu erhält jedes Pixel der Retina nicht nur die Ausgabe des FCC sondern auch den *normierten Grauwert* als Eingabe.

Bild 5.3 zeigt die verwendete Netztopologie, die tatsächlich zwei Netze enthält. Die Retina des unteren Netzes wurde auf die Größe von 16x16 Pixeln verkleinert. Dieses Netz bestimmt die X- und Y–Position eines Gesichtes im Bild. Falls ein Gesicht lokalisiert wurde, wird ein Ausschnitt von 12x12 Pixeln um dieses Gesicht auf die Retina des zweiten Netzes abgebildet, das die Größe des Kopfes bestimmt. Der Vorteil der Aufspaltung in zwei Netze besteht in einer wesentlichen Vereinfachung der Aufgabe des Netzes. Ein kombiniertes Netz müßte für jede Position eines Gesichtes in der Lage sein, seine Größe zu schätzen, während das nachgeschaltete Netz in der aufgeteilten Version nur die Größen von zentrierten Gesichtern erkennen muß.

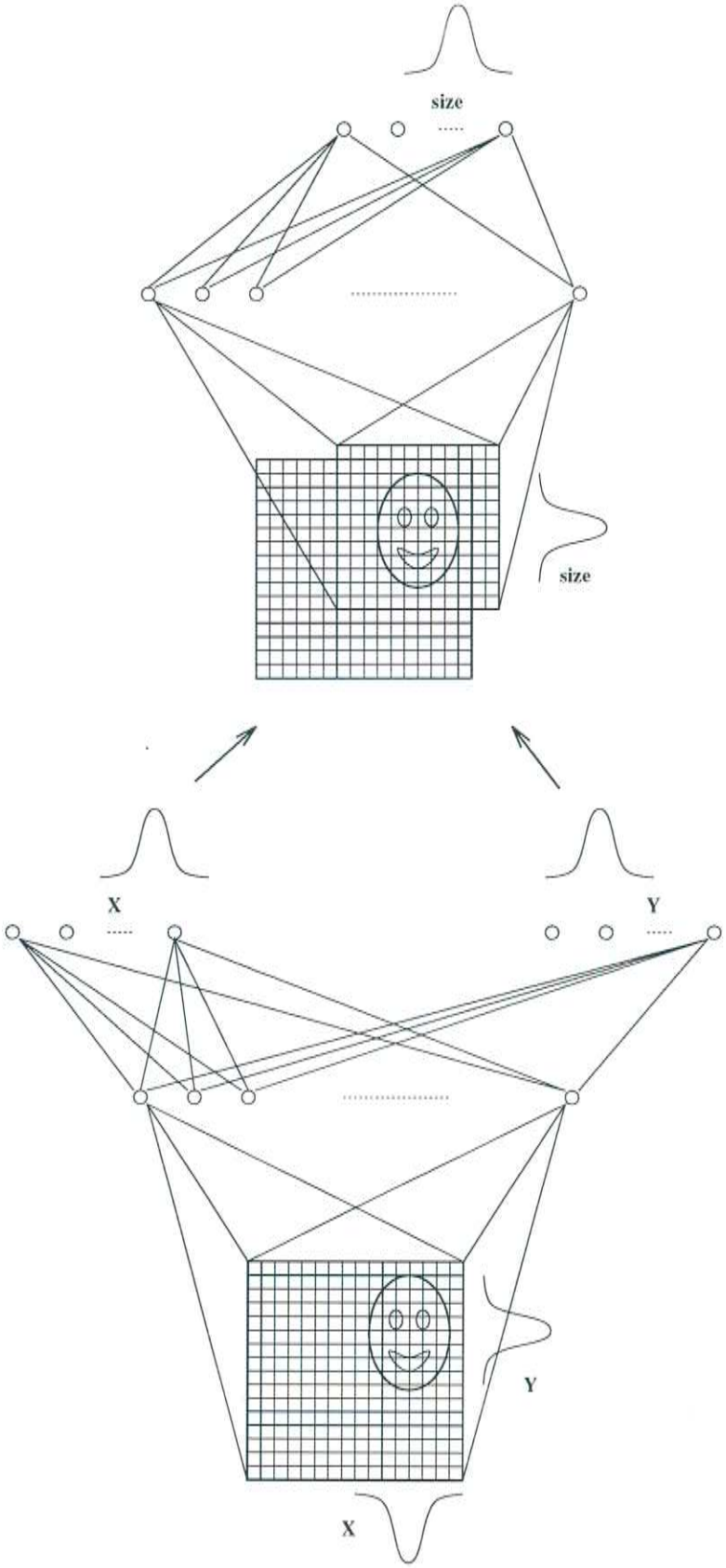


Abbildung 5.3: Aufgeteiltes Netz zur Kameranachführung

Das Netz konvergierte wesentlich schneller als die Netze ohne FCC und erreichte die besten Ergebnisse bereits mit 5000 Trainingsbildern. Tabelle 5.2 zeigt die erzielten Erkennungsraten, die wieder mit 40 Neuronen in der Zwischenschicht erreicht wurden:

Netz	Retina	Ausgabe	Anzahl Gewichte	Erkennungsrate
1	16x16	X, Y	21000	93.6%
2	12x12	Größe	12000	95.2%

Tabelle 5.2: Neuronales Netz mit FCC

Kapitel 6

Künstliche Bilder und Filmsequenzen

Die Möglichkeit, Bilder und Filmsequenzen künstlich zu erzeugen, so daß zu jedem Bild bekannt ist, ob, wo, und in welcher Größe es einen Kopf enthält, ist sowohl für das Training der neuronalen Netze als auch die Untersuchung der Leistungsfähigkeit in bestimmten Situationen sehr wertvoll.

Im folgenden wird die Bedeutung der Generierung von Bildern für beide Anwendungen erläutert und anschließend die Methode illustriert.

6.1 Trainieren von neuronalen Netzen mit künstlichen Bildern

Die hier verwendeten neuronalen Netze werden mit einer großen Menge von Beispielen trainiert, die jeweils eine Netzeingabe und die zugehörige, gewünschte Netzausgabe enthalten. Die Eingabeschicht des Netzes ist als zweidimensionale Retina strukturiert, auf die ein evtl. vorverarbeitetes Kamerabild projeziert wird. Die Ausgabe enthält Informationen über die Position und Größe eines im Bild enthaltenen Gesichtes. Funktionsweise und Aufbau der Netze wurden im Kapitel 5 erklärt.

6.1.1 Anforderungen an die Trainingsmenge

Um ein generalisierendes Verhalten des Netzes zu ermöglichen, sollte die Trainingsmenge folgende Bedingungen erfüllen:

1. **Anzahl der Trainingsbeispiele:** Es sollten mehr Beispiele vorhanden sein, als Verbindungen im gesamten Netz existieren. In den hier verwendeten Netzen führt dies zu mehreren tausend Beispielen.

2. **Repräsentative Netzeingaben:** Um das Erlernen eines einzelnen Gesichtes oder einer Gruppe von Gesichtern mit gemeinsamen Merkmalen zu vermeiden, muß die Trainingsmenge Gesichter unterschiedlichen Alters, Geschlechts, Hautfarbe, Frisur, Barttracht, Blickrichtung, Kopfhaltung usw. enthalten.

3. **Abdeckung der möglichen Netzausgaben:** Die Trainingsmenge muß für jede mögliche Ausgabeklassen Repräsentanten enthalten.

6.1.2 Zuordnung von Netzein- und Ausgabe

Für die Generierung von Trainingsbeispielen sind prinzipiell zwei Verfahren möglich, die Paare von Netzeingaben und -ausgaben liefern:

1. Vorgabe der Netzeingänge und Ermittlung der Netzausgänge
2. Vorgabe der Netzausgänge und Generierung der Netzeingänge

Beide Verfahren werden auf ihre Anwendbarkeit im Hinblick auf die genannten Bedingungen an die Trainingsmenge untersucht.

Vorgabe der Netzeingänge

1. **Anzahl der Trainingsbeispiele:** Eine hohe Anzahl von Bildern läßt sich durch die Aufnahme von Bildsequenzen mit der Kamera erreichen. Die Schwierigkeit bei dieser Vorgehensweise liegt darin, die Größe und Position der Gesichter in den aufgenommenen Sequenzen auf praktikable Weise zu bestimmen. Optische Markierungen, die während der Aufnahme zur Positions- und Größenbestimmung dienen, würden auf den Trainingsbildern erkennbar sein und damit während des Trainings als Merkmal verwendet werden. Derartige Markierungen müßten folglich nichtoptischer Art sein oder vor der Verwendung als Netzeingang vollständig aus den Bildsequenzen herausgefiltert werden.

2. **Repräsentative Netzeingaben:** Die Sequenzen müssen mit einer Vielzahl von Personen vor unterschiedlichen Hintergründen aufgenommen werden. Ist die Markierungsmethode aus technischen Gründen auf einen Raum beschränkt, ergeben sich dadurch unerwünschte Gemeinsamkeiten der Trainingsbilder, z.B. hinsichtlich Beleuchtungssituation, Hintergrundfarben, usw.

3. **Abdeckung der möglichen Netzausgaben:** Alle vom Netz zu erkennenden Kopfgrößen und -positionen müssen in den Sequenzen in vergleichbarer Häufigkeit enthalten sein.

Vorgabe der Netzausgänge

Diese Methode erfordert die Erzeugung von künstlichen Bildern, die entsprechend der gewünschten Kopfpositionen und -größen generiert werden.

1. **Anzahl der Trainingsbeispiele:** Da künstliche Bilder erzeugt werden, können praktisch beliebig viele Trainingsbeispiele berechnet werden.

2. **Repräsentative Netzeingaben:** Die Komplexität der Datenbasis ist für diese Forderung entscheidend. Wie nachfolgend gezeigt wird, ist bei dieser Methode für jedes unterschiedliche Gesicht nur ein Bild notwendig, aus dem weitere Bilder erzeugt werden.

3. **Abdeckung der möglichen Netzausgaben:** Da die Netzausgaben vorgegeben werden, ist diese Forderung durch iteratives Durchlaufen aller Ausgabeklassen erfüllbar.

6.2 Filmsequenzen

Das im Kapitel 6.3 beschriebene Verfahren ermöglicht es, Hintergründe mit Gesichtern beliebiger Größe und Position zu überlagern. Dadurch ist es u.a. möglich,

- Bildsequenzen mit vorgegebener Bewegungsbahn von Gesichtern zu erzeugen.
- gleiche Bewegungsabläufe vor unterschiedlichen Hintergründen auszutesten.
- gleiche Bewegungsabläufe mit einer Vielzahl von Gesichtern auszutesten.
- beliebige Geschwindigkeiten der Bewegungen zu simulieren.

Dadurch bietet sich eine Fülle von Möglichkeiten, das Gesamtsystem unter speziellen Bedingungen zu testen und die Grenzen auszuloten. Die Aussagekraft von Tests mit künstlichen Bildern bleibt jedoch auf das Aufzeigen von Tendenzen beschränkt, da die Bilder sich in mehrfacher Hinsicht von realen Bildern unterscheiden. Auf die Unterschiede wird bei der Erklärung des Verfahrens noch genauer hingewiesen.

6.3 Generierung künstlicher Bilder

Die Vorgabe der Netzausgänge mit anschließender Erzeugung eines dazu passenden Bildes hat für unsere Zwecke entscheidende Vorteile. Die im folgenden beschriebene Methode erlaubt mit geringem manuellem Aufwand, eine Trainingsmenge zu generieren, die alle in Abschnitt 6.1.1 aufgeführten Forderungen erfüllt. Die Methode verwendet zwei Datenbasen, die repräsentative Gesichter und unterschiedliche Hintergründe enthält. Zu jeder vorgegebenen Kopfposition und -größe wird ein Gesicht der Datenbasis auf einen der Hintergründe projiziert.

6.3.1 Einrichtung der Datenbasen

Da zu jedem in der Datenbasis enthaltenen Gesicht die Position und Größe bekannt sein muß, werden diese Daten bereits bei der Aufnahme normiert. Dazu wird auf dem Bildschirm ein Ausschnitt des Kamerabildes dargestellt (s. Bild 6.1 a). Die Kameraposition und Objektivbrennweite werden so eingestellt, daß das aufzunehmende Gesicht mittig mit dem Kinn am unteren und dem Haaransatz am oberen Rand dargestellt wird. Abgespeichert wird ein Bereich doppelter Breite und Höhe (s. Bild 6.1 b). Die Aufnahme muß vor einem besonderen, im nächsten Abschnitt beschriebenen Hintergrund, dem Blue–Screen, durchgeführt werden.

(a)

(b)

Abbildung 6.1: Aufnahme eines Gesichtes

(a) Positionierung des Kopfes vor der Kamera, so daß das Kinn am unteren und der Haaransatz am oberen Rand zu liegen kommt, (b) tatsächlich aufgenommenes Bild

Bild 6.2 zeigt eine Anzahl so entstandener Bilder. Die Gesichter aller Bilder sind durch die Aufnahmetechnik in Position und Größe, genauer Höhe, genormt. Die hier verwendete Datenbasis enthält 24 Personen mit jeweils 3 unterschiedlichen Kopfhaltungen, um das Netz robust gegen Verdrehungen zu trainieren. Insgesamt werden somit 72 RGB–Bilder mit einer Auflösung von jeweils $256 \cdot 256$ Pixeln verwendet.

Abbildung 6.2: Beispielbilder aus der Datenbasis

Eine weitere Datenbasis wurde mit 10 Hintergrundbildern der Auflösung $600 \cdot 460$ Pixeln aus dem Versuchsraum gebildet. Die Hintergrundbilder enthielten keine Gesichter und deckten die gesamten Wände des Versuchsraumes ab. Der Verzicht auf zusätzliche Hintergründe außerhalb des Raumes liegt darin begründet, daß das zu trainierende Netz für die spezielle Arbeitsumgebung optimiert werden sollte und die generierten Trainingsbeispiele damit dem realen Szenarium entsprechen.

6.3.2 Das Blue–Screen Verfahren

Um die Gesichter mit einem neuen Hintergrund unterlegen zu können, bedarf es einer Möglichkeit, ein Gesicht aus einem bestehenden Hintergrund ausschneiden zu können. Das aus der Fernsehtechnik bekannte Blue–Screen Verfahren erlaubt dies auf einfache Weise. Die Gesichter

(a) (b)

Abbildung 6.6: Beispiele von künstlichen Bildern
(a) künstliche Bilder, (b) nach Anwendung des GFCC

der Datenbasis müssen dazu vor einem gleichmäßigen, blauen Hintergrund aufgenommen werden, da die Grundfarbe Blau in einem Gesicht gegenüber den Farben Rot und Grün nur schwach vertreten ist.

Mit den im Kapitel „Farbe als Merkmal" entwickelten Methoden kann die Hintergrundfarbe automatisch ausgeblendet werden (s. Bild 3.2). Bild 6.3 zeigt die Farbverteilung, die aus einer Aufnahme des hier verwendeten Hintergrundes, des Blue–Screens entstanden ist. Der die häufigste Farbe markierende schwarze Punkt repräsentiert die Grundfarbe, während die seltene-ren, benachbarten Farben durch Textur und Reflexionen entstehen. In dieser Aufnahme wurde der blaue Hintergrund aus kürzerer Distanz und nur in dem Bereich aufgenommen, der bei der Datensammlung tatsächlich als Hintergrund diente. Dadurch enthält diese Aufnahme weniger Farbverschiebungen durch Reflexionen als die Aufnahme in Bild 3.2. Dies ist an der stärker abgegrenzten Farbverteilung erkennbar, die mit Farbverteilungen von Gesichtern disjunkt ist (vgl. Bild 3.5 (a)). Dadurch können alle Bereiche als Hintergrund klassifiziert werden, die Farben enthalten, die in der Verteilung von Bild 6.3 auftreten. Bild 6.4 zeigt ein Beispiel aus der Datenbasis, bei dem die Hintergrundfarbe mit dieser Methode schwarz, alle anderen Farben weiß dargestellt wurden.

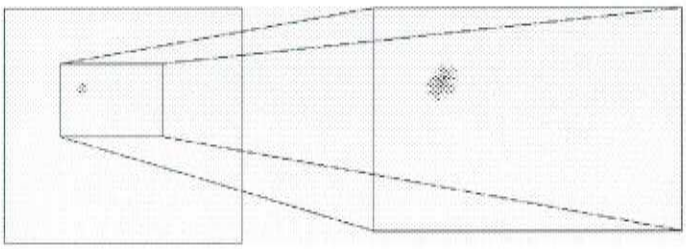


Abbildung 6.3: Farbverteilung des Hintergrundes

Abbildung 6.4: Ausschneiden eines Gesichtes

6.3.3 Berechnung eines Bildes

Um gemäß den Vorgaben Position und Größe ein Bild erzeugen zu können, bedarf es einer Skalierung und Verschiebung des mit dem Blue–Screen Verfahrens extrahierten Gesichtes. Dazu werden von jedem Gesicht der Datenbasis in einer logarithmischen Abstufung 10 Skalierungen im Bereich $\frac{1}{2} \ldots 1$ berechnet. Der Hintergrund des Bildes wird durch eine in Größe und Position zufällige Auswahl eines Auschnittes einer der Hintergrundbilder in der Datenbasis gebildet. Dieser Auschnitt wird entsprechend des Eingangsformates des Netzes skaliert. Der Vordergrund des Bildes entsteht durch Auswahl eines der skalierten Gesichter der Datenbasis, das gemäß der Positionsvorgabe verschoben wird. Da in den künstlichen Bildern nur Gesichter erzeugt werden, die scheinbar frei im Raum schweben, wird der Übergang zwischen Hals, Schultern und Hintergrund fließend berechnet, so daß kein scharfer Übergang als vermeintliches Merkmal beim Training der Netze erlernt wird. Die „rumpflose" Erzeugung der Bilder ist durchaus sinnvoll, da in realen Szenen Teile des Rumpfes häufig verdeckt sind, z.B. beim Sitzen auf einem Stuhl hinter einem Tisch. In Bild 6.5 wird das Verfahren bildlich dargestellt.

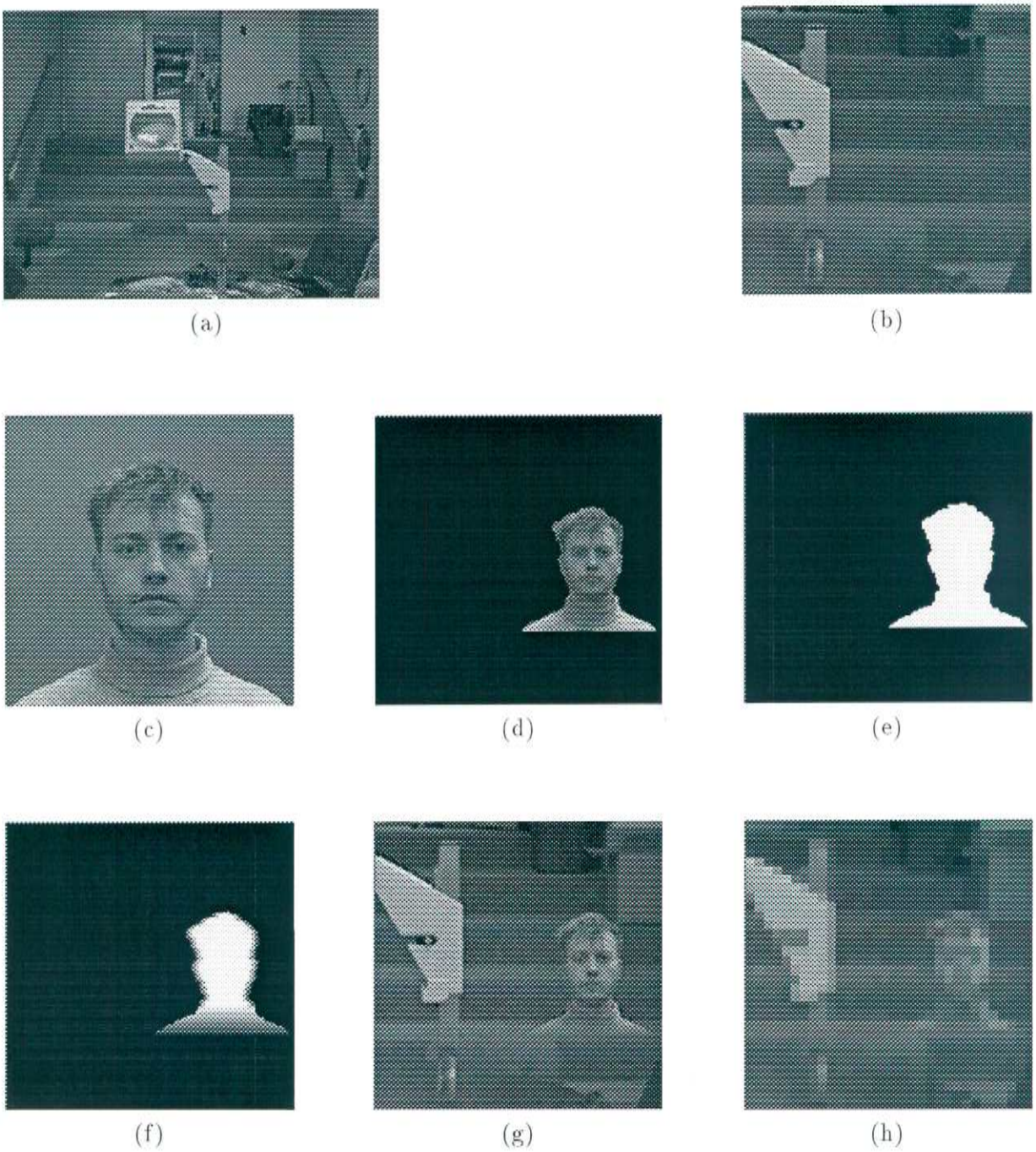


Abbildung 6.5: Berechnung eines künstlichen Bildes

(a) zufälliges Auswählen eines Hintergrundes aus der Datenbank, (b) zufälliges Auswählen eines Ausschnitts, (c) Gesicht aus der Datenbank, (d) Ausblenden des Hintergrundes, Skalierung und Verschiebung entsprechend der Vorgaben, (e) Bildung einer Vordergrundmaske, (f) Verwischung der Maske, (g) Übereinanderlegung von (b) und (d) gemäss der Maske, (h) Abbildung auf das Format der Retina

In Bild 6.6 (a) sind einige Beispielbilder mit verschiedenen Kopfgrößen bei einer Auflösung von $24 \cdot 24$ Pixeln dargestellt. Bild (b) zeigt die gleichen Bilder nach Anwendung des GFCC. Das Lokalisieren eines Gesichtes ist in diesen Bildern durch die hervorgehobene Farbinformation erheblich vereinfacht.

Kapitel 7

Auswertung

Eine objektive Bewertung des Systems erfordert einen Vergleich der tatsächlichen Kopfpositionen und -größen mit den vom Programm gelieferten Daten. Die Aufnahme von Testsequenzen, die diesen Vergleich ermöglichen, wird in diesem Kapitel beschrieben. Es folgt eine Bewertung des Gesamtsystems und ein Vergleich der Systeme mit und ohne Verwendung von neuronalen Netzen.

7.1 Testsequenzen

In Kapitel 6.2 wurde eine Möglichkeit vorgestellt, künstliche Filmsequenzen zu generieren, so daß zu jedem Einzelbild Position und Größe darin enthaltener Gesichter bekannt sind. Aus folgenden Gründen sind die Filmsequenzen nur begrenzt mit realen vergleichbar:

- Die künstlichen Bilder enthalten rumpflose Gesichter. Daher gibt es in den künstlichen Filmsequenzen nur sich bewegende Gesichter. In realen Filmsequenzen ist der sich bewegende Bereich größer und die Lokalisierung daher schwieriger.
- Es werden keine gleitenden Kopfbewegungen oder sich langsam veränderne Beleuchtungssituationen simuliert.

Um bei der Auswertung der Leistungsfähigkeit des Systems möglichst aussagekräftige Ergebnisse zu erhalten, wurden die Leistungsdaten daher basierend auf realen Filmsequenzen ermittelt und keine künstlichen Sequenzen verwendet.

7.1.1 Aufnahme von Testsequenzen

Um Testsequenzen zu erhalten, die den realen Anwendungen möglichst nahe kommen, wurden Filmsequenzen mit dem Rekorderteil der Kamera aufgenommen, während das System dieselbe Kamera einem Gesicht nachführte. Anschließend wurden die Sequenzen von der Kamera im Wiedergabemodus abgespielt und mit einer Bildfrequenz von 10 Bilder/sec in den Rechner eingelesen. Zu jedem Einzelbild wurden von Hand die Größe und Position eines enthaltenen Gesichtes markiert. Die eingelesene Sequenz wurde dem System Bild für Bild vorgeführt und die ausgegebenen Lokalisierungsdaten mit den manuell ermittelten verglichen.

Die so erhaltenen Testsequenzen enthalten Bilder aus realen Szenarien, sind aber aus einem anderen Grund nicht exakt mit einer Anwendung in Realzeit vergleichbar. Tatsächlich erbringt das System auf den Testsequenzen schlechtere Ergebnisse als während der Aufnahme mit der Kamera. Dieses Verhalten hat folgende Ursachen:

- Die Bildqualität von Einzelbildern leidet durch die Zwischenspeicherung auf einem Magnetband. Gelegentlich auftretende Streifen, insbesondere bei sich bewegenden Objekten, führen zu schlechteren Lokalisierungsergebnissen.

- Das System schaltet in die Phase *Lokalisieren* zurück, wenn es ein Gesicht nicht mehr finden kann. Dazu wird in einer Realzeit–Anwendung jede Kamerabewegung gestoppt um die Bewegungsinformation des Vordergrundes nutzen zu können. Bei einer abgespeicherten Sequenz kann die Kamerabewegung nicht mehr nachträglich beeinflußt werden. Daher muß ein Zeitpunkt der Bildsequenz abgewartet werden, bei der sich der Hintergrund nicht bewegt.

Im Realzeitbetrieb liest das System nach Bearbeitung des letzten Bildes sofort das nächste ein und arbeitet bei der verwendeten Hardware mit einer durchschnittlichen Bildfrequenz von 5 Bildern/sec. Eine gespeicherte Bildsequenz wird dagegen mit einer festen Bilderrate (10 Bilder/sec) aufgenommen. Um vergleichbare Ergebnisse zu erzielen, wird dem System nur jedes zweite Bild der gespeicherten Sequenz gezeigt. Werden alle Bilder gezeigt, können dadurch eine doppelte Rechengeschwindigkeit simuliert und die Verbesserungen abgeschätzt werden, die ein schnellerer Rechner im Realzeitbetrieb erzielen könnte.

7.1.2 Markierung der Testsequenzen

Um eine Bewertung der Genauigkeit der Lokalisierungsdaten zu ermöglichen, wurden die Gesichter in allen Bildern der gespeicherten Sequenzen von Hand markiert. Dazu wurde für jedes Gesicht die Eckpunkte eines Rechtecks eingegeben, so daß das gesamte Gesicht innerhalb des markierten Bereichs zu liegen kam. Als Gesicht wurde der Bereich eines Kopfes zwischen Kinn und Haaransatz definiert. Durch diese subjektive Markierung ist eine Differenz mit den vom System ermittelten Lokalisierungsdaten zwangsläufig. Eine von zwei verschiedenen Personen markierte Sequenz gibt einen Hinweis auf diesen subjektiven Fehler. Je nach Größe des Gesichtes varriert der subjektiv als Gesicht erkannte Bereich um 10% der Gesichtsgröße. Abweichungen in diesem Bereich werden daher nicht als mangelnde Genauigkeit betrachtet.

7.1.3 Auswahl der Testsequenzen

Die Genauigkeit einer Lokalisierung hängt im wesentlichen vom Schwierigkeitsgrad der Szenerie ab. Die Merkmale, die eine Lokalisierung erschweren, sind

- schnelle Kopfbewegungen,

- häufige Veränderungen des Hintergrundes durch Bewegung der Kamera oder Adjustierung der Brennweite,

- und Auftreten größerer Bereiche im Hintergrund, die Farben enthalten, die vom GFCC oder sogar vom IFCC als Gesichtsfarben klassifiziert werden.

Das Ergebnis einer Mittelung der Leistungsdaten über Testsequenzen verschiedenen Schwierigkeitsgrades ist abhängig vom Anteil schwieriger gegenüber leichter Szenarien. Die Bewertung wird daher auf zwei unterschiedlich schwierigen Testsequenzen getrennt durchgeführt:

- *Testsequenz 1:* Sechs Personen unterschiedlicher Hautfarbe wurden nacheinander vor einem „einfachen" Hintergrund gefilmt, der auch bei ausschließlicher Verwendung des GFCC gute Ergebnisse zuließ.

- *Testsequenz 2:* Eine weitere Person wurde vor einem Hintergrund aufgenommen, der viele Farben des GFCC enthielt. Die Testperson führte dafür langsamere Bewegungen durch.

Um einen Mindestschwierigkeitsgrad zu wahren, wurden alle Personen angewiesen, nur über kurze Zeiträume still zu verharren (weniger als 5% der Gesamtzeit), mindestens einmal während der Aufnahme aufzustehen, sich öfter mit der Hand ans Kinn zu fassen oder durch die Haare zu fahren und sich so stark zu bewegen, daß die physikalische Kamera fast ständig zu nachführenden Bewegungen (mehr als 70% der Gesamtzeit) gezwungen war. Derartige Bewegungen entsprechen Grenzsituationen in realen Szenarien. Die vergleichsweise geringen Bewegungen während eines tatsächlichen Gesprächs erfordern nur selten eine Nachführung der physikalischen Kamera und liefern fehlerfreie Ergebnisse.

Um die Stärke der Bewegungen auf eine bewertbare Grundlage zu stellen, wurden die durchschnittlichen und maximalen Bewegungen von Gesichtern relativ zur Gesichtsgröße zwischen zwei aufeinander folgenden Bildern jeder Testsequenz gemessen. Die in der Tabelle 7.1 angegebenen Meßwerte beziehen sich auf eine Bildfrequenz von 5 Bildern/sec. Der Eintrag $y_{max} = 1.88$ zeigt, daß die maximale Bewegung eines Gesichtes in dieser Testsequenz in vertikaler Richtung das 1.88 fache der Gesichtshöhe beträgt.

Testsequenz	Anzahl Bilder	$\overline{x}$	$\overline{y}$	x_{max}	y_{max}
1	1860	0.19	0.06	1.52	1.88
2	460	0.19	0.05	1.26	0.56

Tabelle 7.1: Verwendete Testsequenzen

Die Angabe der Bewegung relativ zur Gesichtsgröße gibt Aufschluß darüber, ob das Gesicht im nächsten Bild innerhalb des Blickfeldes der virtuellen Kamera erwartet werden kann. Da die virtuelle Kamera in Breite und Höhe die doppelten Ausmaße des Gesichtes besitzt, wird bei einer Bewegung größer als 0.5 im nächsten Bild nur ein Teil des Gesichtes sichtbar sein, bei Bewegungen größer 1.5 das Gesicht sogar ganz fehlen. Die Durchschnittswerte der Bewegungen in beiden Sequenzen liegen weit unter diesen Grenzwerten. Vertikale Bewegungen treten hauptsächlich beim Aufstehen auf, daher die große Diskrepanz zwischen Durchschnitts- und Maximalwerten bei vertikalen Bewegungen. In der Testsequenz 1 treten mehrfach Bewegungen größer 1.5 auf, so daß ein Verlust des Gesichtes zwangsläufig folgt.

7.1.4 Evaluation einer Testsequenz

Bei der Auswertung der Testsequenzen interessieren folgende Aussagen, die anschließend im Einzelnen behandelt werden:

1. Wie hoch ist der Anteil der Bilder, in denen das Gesicht erfolgreich lokalisiert wird?

2. Wie groß ist der systematische Fehler bei der Angabe der Position und Größe des Gesichtes, d.h. wie groß ist die durchschnittliche Abweichung von den tatsächlichen Daten?

3. Wie groß ist die Standardabweichung der Differenz zu den tatsächlichen Daten?

Anteil der richtigen Lokalisierungen

Bei sehr schnellen Bewegungen oder Hintergründen, die Bereiche enthalten, die als Gesichtsfarben klassifiziert werden, kann das System die Position des Gesichtes verlieren und muß eventuell zur Phase *Lokalisieren* zurückschalten. Der Wert *correct* gibt den Zeitanteil in Prozent an, den das Sytem in der Phase *Nachführen* verbringt und die Position des Gesichtes richtig erkennt. Die restliche Zeit wird mit erneutem Suchen nach einem Verlust des Gesichtes verbracht. Die erste Suche des Sytems bei Beginn der Testsequenz wird bei dieser Angabe nicht berücksichtigt. Für ein nachgeschaltetes System, z.B. zum Lippenlesen, gibt dieser Wert an, wieviel Prozent der Zeit ein stabiles Bild des Sprechers erwartet werden kann.

Systematischer Fehler

Es werden zu jeder Testsequenz vier systematische Fehler angegeben, s_x und s_y für die Schätzung der Position, sowie s_w und s_h für die Schätzung der Größe des Gesichts. Es bezeichnen x_{target}, y_{target}, w_{target} und h_{target} die tatsächliche Position und Größe des Gesichtes, sowie x_{output}, y_{output}, w_{output} und h_{output} die vom System ermittelten Werte. Dann berechnen sich die systematischen Fehler als artihmetische Mittel folgender Werte über alle Bilder einer Testsequenz:

$$s_x : \frac{x_{target} - x_{output}}{w_{target}} \qquad s_w : \frac{w_{output}}{w_{target}}$$

$$s_y : \frac{\|y_{target} - y_{output}\|}{h_{target}} \qquad s_h : \frac{h_{output}}{h_{target}}$$

Standardabweichung

Ein weiterer Gesichtspunkt der Leistungsfähigkeit des Systems bildet die Standardabweichung der vom System ermittelten Lokalisierungsdaten bezogen auf die systematischen Fehler als Mittelwerte. Ein hoher, aber bekannter systematischer Fehler läßt sich eliminieren; eine hohe Standardabweichung führt dagegen zu einer Ungenauigkeit des Systems. Da die Abweichung der Größenschätzung des Gesichtes sich auf Quotienten bezieht, werden die Differenzen über die Logarithmen der Quotienten gebildet. Die Standardabweichung vom jeweiligen systematischen Fehler berechnet sich als Mittelwert über alle quadratischen Abweichungen:

$$\sigma_x : \left(\frac{\|x_{target} - x_{output}\|}{w_{target}} - s_x\right)^2 \qquad \sigma_w : \left(\log_2 \frac{w_{output}}{w_{target}} - \log_2 s_w\right)^2$$

$$\sigma_y : \left(\frac{\|y_{target} - y_{output}\|}{h_{target}} - y_x\right)^2 \qquad \sigma_h : \left(\log_2 \frac{h_{output}}{h_{target}} - \log_2 s_h\right)^2$$

Unter Annahme einer Normalverteilung der Abweichungen der Lokalisierungsdaten kann die Genauigkeit berechnet werden, die für einen vorgegebenen Prozentsatz aller Lokalisierungdaten erwartet werden kann. In 95.5% aller Bilder kann eine Abweichung der Position im Bereich $[s - 2\sigma, s + 2\sigma]$ und in 68.3% aller Fälle im Bereich $[s - \sigma, s + \sigma]$ erwartet werden. Für die Schätzung der Größe ist in 95.5% aller Bilder ein Faktor von $[s2^{-2\sigma}, s2^{2\sigma}]$ und in 68.3% aller Fälle ein Faktor von $[s2^{-\sigma}, s2^{\sigma}]$ garantiert.

7.2 Ergebnisse

Die Testsequenzen wurden in verschiedenen Modi des Gesamtsystems ausgewertet, um einzelne Komponenten des Systems getrennt bewerten zu können. Im Einzelnen interessierte der Nutzen der Anpassung des Klassifikators GFCC zum IFCC, die Tiefpaßfilterung der Ausgabe des

Klassifikators und ein Vergleich der neuronalen Netze mit dem Algorithmus *suche größtes zusammenhängendes Objekt*.

7.2.1 Testsequenz 1

Folgende Modi wurden mit der Testsequenz 1 einzeln ausgewertet:

1. Verwendung des GFCC, keine Anpassung des Klassifikators
2. zusätzliche Tiefpaßfilterung der Ausgabe des GFCC
3. mit Anpassung des GFCC zum IFCC
4. Ersetzung der Objektsuche durch neuronale Netze

Tabelle 7.2 zeigt die Ergebnisse der Evaluierung.

Modus	correct	s_x	s_y	s_w	s_h	σ_x	σ_y	σ_w	σ_h
1	89.5%	0.11	0.10	1.31	1.23	0.08	0.07	0.13	0.14
2	96.1%	0.09	0.08	1.26	1.27	0.06	0.05	0.13	0.14
3	96.1%	0.07	0.09	1.24	1.29	0.06	0.06	0.13	0.13
4	68.6%	0.16	0.17	1.15	1.18	0.10	0.10	0.38	0.31

Tabelle 7.2: Ergebnisse für Testsequenz 1

Die Einführung der Tiefpaßfilterung bringt eine erhebliche Verbesserung der Zuverlässigkeit des Systems, die sich in einer starken Erhöhung des Wertes *correct* ausdrückt. Die zusätzliche Anpassung des GFCC auf die tatsächlich im Gesicht enthaltenen Farben bringt keine wesentliche Veränderung der Leistungsdaten mit sich. Dies liegt an der weitgehenden Vermeidung von Hintergrundfarben in der Testsequenz 1, die vom GFCC als Gesichtsfarben klassifiziert werden.

Daß trotz der Ausnutzung aller verwendeten Methoden im Modus 3 und dem farblich unproblematischen Hintergrund der Wert *correct* unter 100% bleibt, liegt an der bereits erwähnten hohen Maximalbewegung dieser Testsequenz, die einen Verlust des Gesichtes erzwingt.

Die Mittelwerte und Standardabweichungen erfahren in den ersten 3 Modi jeweils leichte Verbesserungen. Ausnahme ist der sich erhöhende Wert s_h. Grund ist das Heranführen eines Armes an das Gesicht. Die Routine *suche größtes zusammenhängendes Objekt* wird durch Einführung der Tiefpaßfilterung und Anpassung des GFCC darin unterstützt, den Arm, soweit Hautfarbe erkennbar ist, als Teil des Gesichtes anzusehen, und führt damit zu einer Vergrößerung des systematischen Fehlers für die Schätzung der Höhe des Gesichtes.

Die Verwendung von neuronalen Netzen führt zu einer gegenteiligen Tendenz. Die erhebliche Verschlechterung des Wertes *correct* fällt in der Testsequenz allerdings wesentlich stärker aus als bei Realzeitlokalisierungen. Dies liegt zum Einen daran, daß die Kamerabewegung bei den Testsequenzen nach einem Verlust des Gesichtes nicht mehr nachträglich gestoppt werden kann. Es muß daher auf einen Moment der Filmsequenz ohne Kamerabewegung gewartet werden. Eine zusätzliche Suche kann daher bereits zu einer wesentlichen Verschlechterung des Wertes *correct* führen. Zum Anderen wurden die Netze nur mit Gesichtern trainiert, die vollständig in der Retina enthalten waren. Ein Gesicht, daß nur zur Hälfte in der virtuellen Kamera sichtbar ist, wird von den Netzen nicht mehr als solches erkannt, von der Routine *suche größtes Objekt* aber noch geortet. Die Mittelwerte und Standardabweichungen der Lokalisierungen besitzen ebenfalls

höhere Werte bis auf die Mittelwerte der Größenschätzung des Gesichtes, die sich verbessert haben. Dies ist durch die Formerkennung erklärbar, die von den Netzen in begrenztem Maße durchgeführt wird und auf die in diesem Kapitel noch näher eingegangen wird.

In Tabelle 7.3 sind die maximalen Abweichungen angegeben, die für den bezeichneten Prozentsatz *Zeit* garantiert werden können. Der Wert *Zeit* entsteht aus der Multiplikation des Wertes *correct* mit einem der Werte 68.3 oder 95.5, für die Abweichungen im Bereich von σ bzw. $2 \cdot \sigma$ gelten. Z.B. kann für 91.7% der Zeit eine geringere Abweichung der X–Position als 19% der Kopfbreite und eine Schätzung der Kopfbreite um das 1.04 bis 1.48 fache des tatsächlichen Wertes erwartet werden.

Modus	Zeit	x	y	w	h
ohne NN	65.4%	<0.13	<0.15	[1.13,1.36]	[1.18,1.41]
ohne NN	91.7%	<0.19	<0.21	[1.04,1.48]	[1.08,1.54]
mit NN	46.7%	<0.26	<0.27	[0.88,1.50]	[0.95,1.46]
mit NN	65.5%	<0.36	<0.37	[0.70,1.89]	[0.77,1.81]

Tabelle 7.3: Genauigkeit bei Testsequenz 1

Um den Einfluß der Rechengeschwindigkeit auf die Zuverlässigkeit und Genauigkeit des Systems zu untersuchen, wurde der obige Modus 3 mit verschiedenen Bildfrequenzen wiederholt (s. Tabelle 7.4).

Bilder/sec	correct	s_x	s_y	s_w	s_h	σ_x	σ_y	σ_w	σ_h
2.5	89.5%	0.11	0.09	1.24	1.31	0.07	0.06	0.14	0.14
5	96.1%	0.07	0.09	1.24	1.29	0.06	0.06	0.13	0.13
10	98.8%	0.08	0.1	1.23	1.27	0.05	0.07	0.12	0.14

Tabelle 7.4: Auswirkungen der Rechengeschwindigkeit

Während die sich nur unwesentlich verändernden Mittelwerte und Standardabweichungen auf eine gleichbleibende Genauigkeit des Systems bei unterschiedlichen Rechengeschwindigkeiten schließen lassen, deutet die starke Erhöhung des Wertes *correct* auf eine deutliche Steigerung der Zuverlässigkeit hin. Solange das Gesicht vollständig im Bereich der virtuellen Kamera sichtbar ist, bringt eine Erhöhung der Rechengeschwindigkeit keine Verbesserung. Sie wirkt sich nur dann positiv aus, wenn sich das Gesicht stark genug bewegt, um bei einer zu geringen Rechengeschwindigkeit teilweise oder vollständig aus dem virtuellen Bildbereich zu verschwinden. Die bei Erhöhung der Rechengeschwindigkeit seltener auftretenden Situationen, in denen das System das Gesicht verliert, wirken sich in der Erhöhung des Wertes *correct* aus.

7.2.2 Testsequenz 2

Testsequenz 2 wird verwendet, um die Leistungsfähigkeit des Systems bei einem „schwierigen" Hintergrund, aber dafür langsameren Bewegungen zu testen. Die Testssequenz wurde unter folgenden Modi bewertet:

1. Verwendung des GFCC, keine Anpassung des Klassifikators
2. Verwendung des GFCC, mit Anpassung des Klassifikators

3. Verwendung eines Klassifikators, der manuell auf das Gesicht in der Testsequenz angepaßt wurde

Der dritte Modus verwendet einen Klassifikator, der mit einem manuell angegebenen Ausschnitt des Gesichtes gebildet und während der Testsequenz nicht mehr verändert wurde. Die Ergebnisse dieses Modus können daher als oberere Grenzwerte für eine automatische Anpassung des Klassifikators betrachtet werden.

Tabelle 7.5 zeigt die Ergebnisse für die einzelnen Modi unter Verwendung der Routine *suche größtes zusammenhängendes Objekt*. Die automatische Anpassung des Klassifikators bringt hinsichtlich der Zuverlässigkeit und der Genauigkeit wesentliche Verbesserungen, die etwa mit den Resultaten übereinstimmen, die bei manueller Anpassung des Klassifikators erzielt werden konnten. Das Verfahren der automatischen Anpassung kann daher als zufriedenstellend betrachtet werden.

Modus	correct	s_x	s_y	s_w	s_h	σ_x	σ_y	σ_w	σ_h
1	81.7%	0.85	0.38	2.79	2.06	0.67	0.32	0.37	0.4
2	100.0%	0.09	0.1	1.39	2.06	0.05	0.07	0.13	0.14
3	100.0%	0.08	0.12	1.32	2.01	0.04	0.06	0.08	0.14

Tabelle 7.5: Ergebnisse für Testsequenz 2

Tabelle 7.6 zeigt die entsprechenden Werte bei Verwendung von neuronalen Netzen. In Modus 1 wurden sowohl Zuverlässigkeit als auch Genauigkeit durch die Netze entscheidend verbessert. Insbesondere die Mittelwerte der Abweichungen bzgl. der Größenschätzung des Gesichtes wurden drastisch reduziert. Bei einer manuellen oder automatischen Anpassung des Klassifikators ist die Genauigkeit des Systems geringfügig schlechter als ohne Verwendung der neuronalen Netze. Ausnahme bildet der Wert s_h, der mit oder ohne Anpassung des Klassifikators wesentlich verbessert wurde. Diese Daten lassen auf eine vom Netz durchgeführte Formerkennung schließen, da das Netz wesentlich besser mit schlechten Ergebnissen des Farbklassifizierers zurechtkommt.

Modus	correct	s_x	s_y	s_w	s_h	σ_x	σ_y	σ_w	σ_h
1	94.8%	0.27	0.31	1.43	1.44	0.18	0.24	0.42	0.34
2	100.0%	0.13	0.15	1.21	1.31	0.08	0.07	0.21	0.18
3	100.0%	0.12	0.1	1.13	1.28	0.08	0.07	0.17	0.16

Tabelle 7.6: Ergebnisse für Testsequenz 2 mit neuronalen Netzen

Tabelle 7.7 zeigt die maximalen Abweichungen, die für die angegebenen Zeiträume garantiert werden können.

Modus	Zeit	x	y	w	h
ohne NN	86.3%	<0.14	<0.17	[1.27,1.52]	[1.87,2.27]
ohne NN	95.5%	<0.19	<0.24	[1.16,1.66]	[1.70,2.50]
mit NN	86.3%	<0.21	<0.22	[1.05,1.40]	[1.16,1.48]
mit NN	95.5%	<0.29	<0.29	[0.90,1.62]	[1.02,1.68]

Tabelle 7.7: Genauigkeit bei Testsequenz 2

Woher kommt die Sonderstellung des Wertes s_h? Bild 7.1 zeigt eine Situationsstudie, die einen Teil der Testsequenz 2 im Modus mit automatischer Anpassung des Klassifikators enthält. Die Versuchsperson führt in dieser Teilsequenz die rechte Hand vor den Mund, so daß der Algorithmus *suche größtes zusammenhängendes Objekt* den unbedeckten Arm und das Gesicht zusammen als Gesicht einsstuft und kurzzeitig sogar nur den Arm als Gesicht erkennt. Bild 7.3 zeigt jeweils das lokalisierte Gesicht in den Einzelbildern. Diese Bilder zeigen deutlich die Herkunft des hohen Wertes s_h.

Bild 7.2 zeigt die gleiche Studie unter Verwendung von neuronalen Netzen. Die Lokalisierungen des Gesichtes werden in Bild 7.4 dargestellt. Obwohl der Arm auch hier deutlich in den Bereich der virtuellen Kamera eintritt, ist das Netz in der Lage, das Gesicht richtig zu lokalisieren und den Arm als vom Gesicht getrennt zu betrachten. Dieses Verhalten läßt auf eine Formerkennung schließen, zu der der Algorithmus *suche größtes Objekt* nicht in der Lage war.

In Situationen, die eine Formerkennung nicht erfordern, führt die Verwendung des einfachen Algorithmus zu höherer Zuverlässigkeit und Genauigkeit, insbesondere bei starken Bewegungen. Dies ist der Fall, wenn der Hintergrund keine Gesichtsfarben aufweist und das Gesicht von anderen Objekten mit Hautfarbe deutlich abgegrenzt ist. Ist diese Abgrenzung nicht gegeben, und ist für eine korrekte Lokalisierung eine Formerkennung unerläßlich, bringt die Verwendung der neuronalen Netze bei geringen Bewegungen deutliche Verbesserungen.

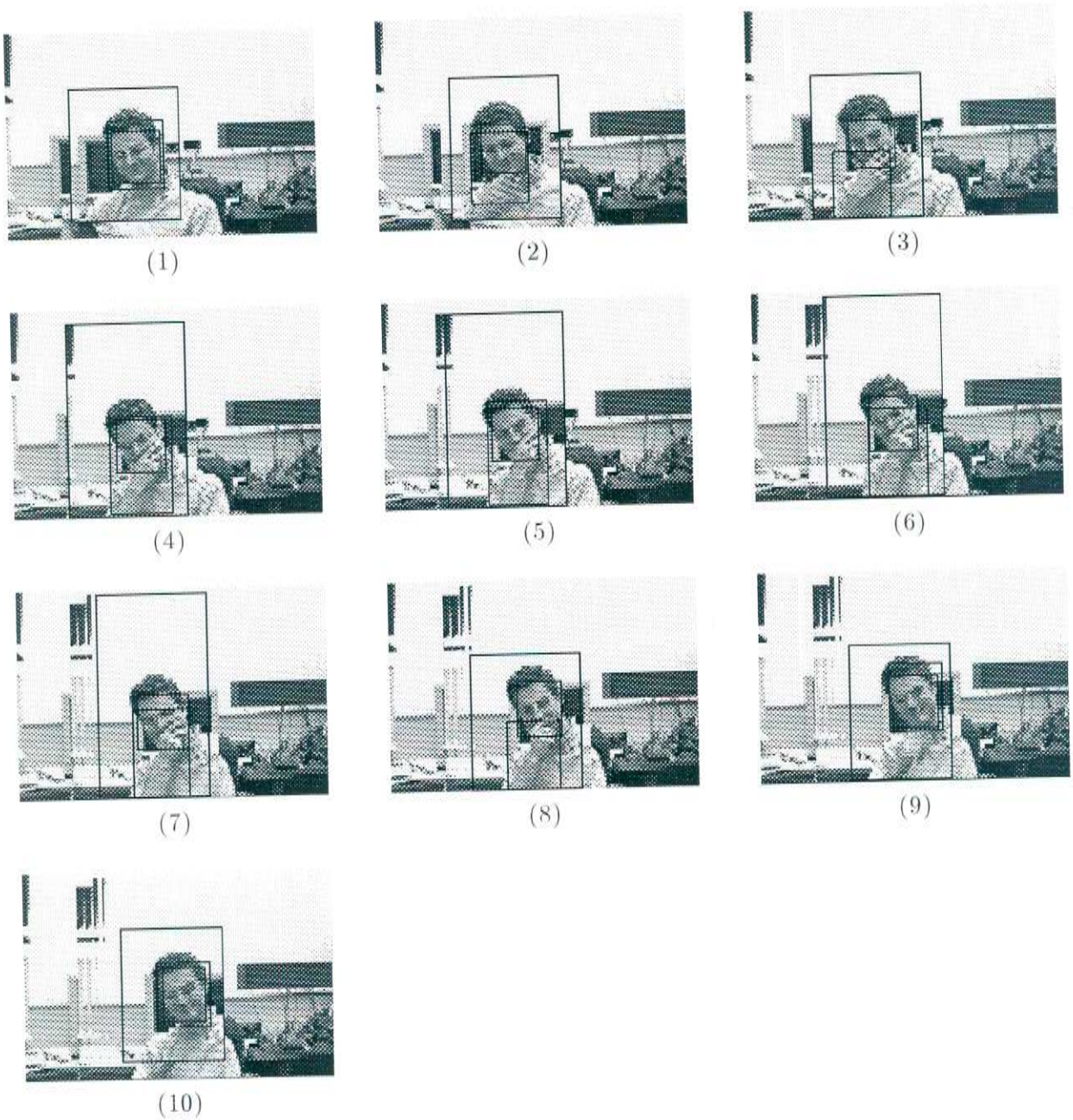


Abbildung 7.1: Situationsstudie - suche größtes Objekt

Abbildung 7.2: Situationsstudie - mit neuronalen Netzen

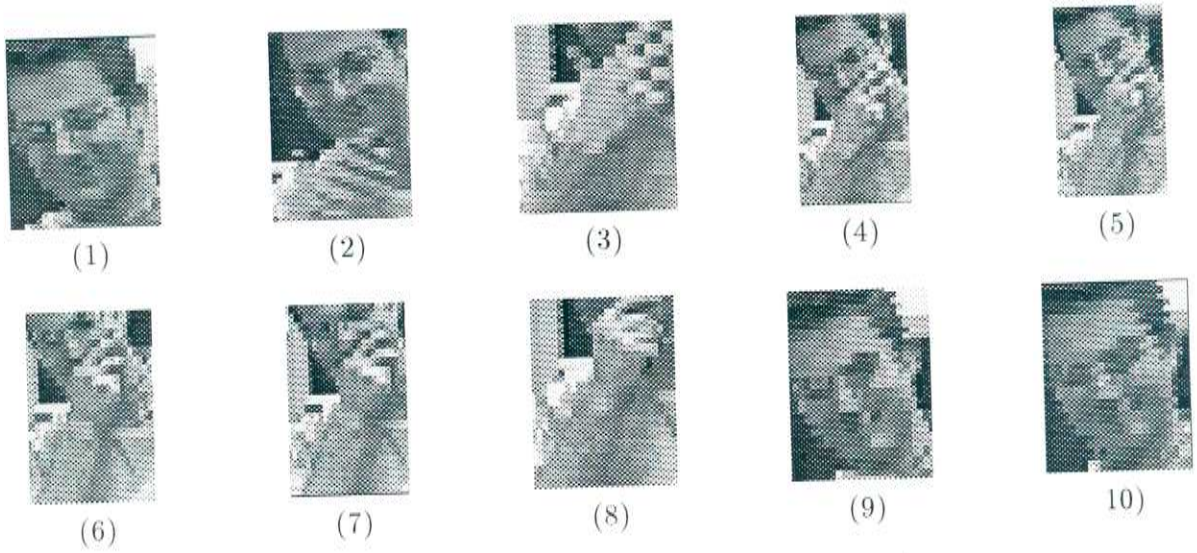


Abbildung 7.3: Ausgabe - suche größtes Objekt

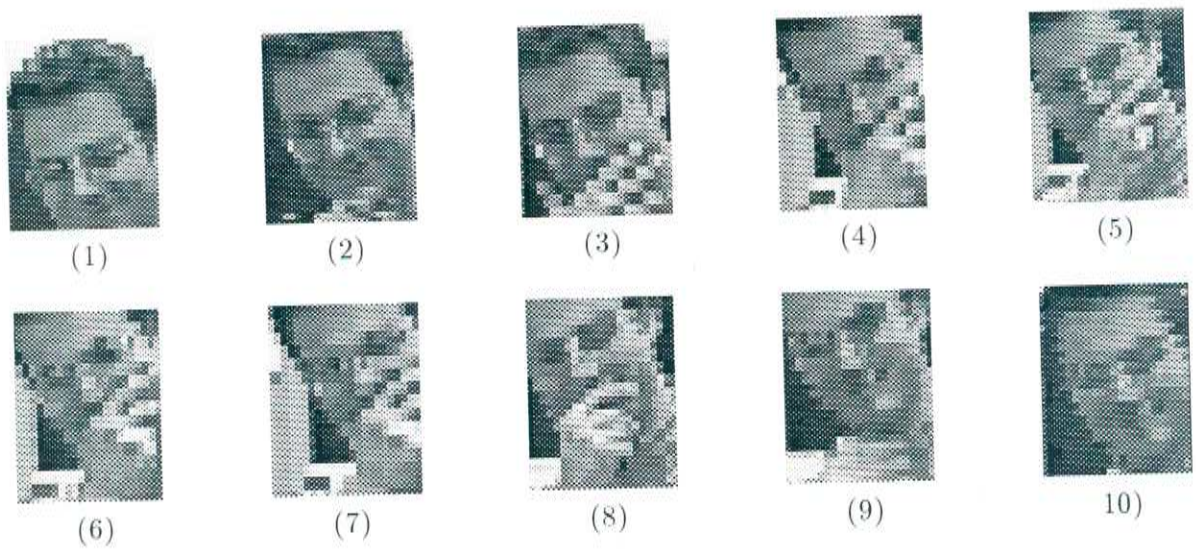


Abbildung 7.4: Ausgabe - mit neuronalen Netzen

Kapitel 8

Zusammenfassung

In dieser Arbeit wurde gezeigt, daß das Lokalisieren von Gesichtern in Echtzeit auf einer Workstation möglich ist. Die in der Einführung genannten Forderungen an ein System konnten erfüllt werden. Die zuverlässige und genaue Lokalisierung von Gesichtern ist durch das Auffinden von zusammenhängenden Bereichen, die Bewegung und Gesichtsfarbe enthalten, durchführbar.

Die Forderung der Echtzeit–Fähigkeit führte zur Beschränkung auf Bilder mit geringen Auflösungen, um eine Bildfrequenz von mindestens 5 Bildern pro Sekunde zu garantieren. Eine Lokalisierung von Gesichtern in diesen Bildern wurde erst durch die Verwendung von Farbe zuverlässig möglich. Die Einführung eines Gesichtsfarbenklassifikators, der eine Abstraktion von RGB–Werten auf ein Maß für Gesichtsfarbe durchführt, ermöglichte die Zusammenfassung aller Farbabhängigkeiten an zentraler Stelle, so daß andere Teilsysteme trotz Verwendung von Farbe unabhängig von Beleuchtungsverhältnissen, Hautfarben sowie der verwendeten Hardware erstellt werden konnten. Ein Austausch von Hardwarekomponenten, z.B. der Kamera oder des Framegrabbers, die einen Einfluß auf die gemessenen Farbwerte haben, beeinflußt dadurch nur noch diesen Klassifikator, erfordert aber keine Änderung eines anderen Teilsystemes mehr. Insbesondere ist dadurch kein erneutes Training der Netzwerke mehr erforderlich. Dieser Farbklassifikator kann mit den vorgestellten Methoden automatisch während der Lokalisierung auf unterschiedliche Beleuchtungsverhältnisse sowie Hautfarben angepaßt werden.

Die verwendeten Merkmale zur Gesichtslokalisierung ermöglichen die zuverlässige Lokalisierung von beliebigen Gesichtern, unabhängig von Hautfarbe, Frisur oder Barttracht. Fehlerhafte Lokalisierungen sind nur in folgenden Fällen zu erwarten:

- Mehrere Gesichter mit ähnlicher Gesichtsfarbe überlappen sich im Kamerabild. Bei Verwendung des Systems ohne neuronale Netze werden beide Gesichter als ein zusammenhängendes Objekt fehlklassifiziert. Die neuronalen Netze wählen dagegen eines der Gesichter aus. Die gleiche Problematik entsteht auch bei starken Gestikbewegungen, die Bereiche der Hände oder Arme vor das Gesicht bringen.
- Das Gesicht befindet sich vor einem Hintergrund, der viele der im Gesicht auftretenden Farben aufweist. Die neuronalen Netze führen durch die Formerkennung zu einer Verbesserung in diesen Situationen.
- Die Bewegung des Gesichtes überschreitet einen Maximalwert, der bei Verwendung der neuronalen Netze geringer liegt als ohne deren Verwendung.

Die Einführung einer virtuellen Kamera ermöglicht das verzögerungsfreie Ausschneiden und Skalieren des das Gesicht enthaltenden Bildausschnittes, so daß die Trägheit der Bewegung der

physikalischen Kamera sowie der Objektiveinstellung keine Limitierung des Systems mehr bedeuten. Die bei einem Gespräch auftretenden geringen Bewegungen durch Verrücken eines Stuhles, Zurück- oder Vorlehnen oder Drehen des Kopfes werden problemlos vom System eliminiert.

Kapitel 9

Ausblick

Das in dieser Diplomarbeit entwickelte Systems bietet viele Möglichkeiten zu Erweiterungen und zu Anwendungen in Kombination mit anderen Systemen, die die optische Information oder die Angaben über die Position eines Gesichtes bei der Mensch–Computer Kommunikation verwenden können.

9.1 Erweiterungen des Systems

9.1.1 Stereosehen

Die Verwendung einer zweiten Kamera würde eine Entfernungschätzung ermöglichen. Da die Größe eines Gesichtes im Bild von der Entfernung und der Objektivbrennweite abhängt, erlaubt die zusätzliche Information die Überprüfung der gefundenen Gesichtsgröße. Zusätzlich erlaubt das Stereobild eine bessere Trennung des Gesichtes vom Hintergrund. Eine Verbesserung ist auch in Situationen zu erwarten, in denen das Gesicht von Objekten mit Hautfarbe teilweise verdeckt wird.

9.1.2 Lokalisieren von Lippen und Augen

Die Methoden der Farbklassifizierung erlaubt auch die Lokalisierung von Bereichen, die Farben enthalten, die wesentlich seltener im Gesicht auftreten als andere Farben. Das Rot der Lippen und das Weiß der Augen treten z.B. wesentlich seltener als die Hautfarbe im Gesicht auf.

9.1.3 Lokalisieren mehrerer Gesichter

Die Verwendung mehrerer virtueller Kameras erlaubt die gleichzeitige unabhängige Gewinnung stabiler Bilder von mehreren Personen bei Verwendung von nur einer physikalischen Kamera.

9.2 Anwendungen des Systems

9.2.1 Bildtelefon

Die stabile Bildübertragung der Gesprächsteilnehmer erforderte bisher die Einhaltung einer vorgegebenen Position vor der Kamera. Die in dieser Arbeit entwickelten Techniken erlaubt dagegen eine freie Bewegung der Teilnehmer.

9.2.2 Gesichtsidentifizierung

Eine Verknüpfung mit bereits bestehenden Systemen zur Gesichtsidentifizierung ermöglicht das automatische Erkennen von Personen, die in das Blickfeld der Kamera treten. Ein exaktes Positionieren des Gesichtes vor der Kamera ist dazu nicht mehr notwendig.

9.2.3 Auswertung der Sprecherposition

In Verbindung mit einem Mikrofon–Array können akustische Sprachsignale aus der Richtung herausselektiert werden, in der das visuelle Bild eines Sprechers oder einer Sprecherin lokalisiert wurde. Bei Videokonferenzen kann dadurch die Kamera automatisch das Bild derjenigen Person übertragen, die gerade spricht. Umgekehrt ermöglicht das visuelle Lokalisieren einer sich bewegenden Person eine ständige Positionsangabe für das Mikrofon–Array, das als Richtmikrofon dienend das Sprachsignal der beobachteten Person gezielt herausfiltern kann.

Literaturverzeichnis

[1] S. Baluja and D. Pomerleau. Non–Intrusive Gaze Tracking Using Artificial Neural Networks. In *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, 1993.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving Connected Letter Recognition by Lipreading. In *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, 1993.

[3] G. W. Cottrell. Extracting Features from Faces Using Compression Networks: Face, Identity, Emotion, and Gender Recognition Using Holons. In *Connectionist Models: Proceedings of the 1990 Summer School*, pages 328–337, San Mateo, 1990. Morgan Kaufmann.

[4] A. Despopoulos. *Color Atlas of Physiology.* New York, 1991.

[5] P. Duchnowski, U. Meier, and A. Waibel. See Me, Hear Me: Integrating Automatic Speech Recognition And Lip–Reading. In *ICSLP*, 1994.

[6] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko. Computer Steered Microphone Arrays for Sound Transduction in Large Rooms. *Journal of the Acoustical Society of America*, 78:236–255, November 1985.

[7] M. K. Fleming and G. W. Cottrell. Categorization of Faces Using Unsupervised Feature Extraction. In *International Joint Conference on Neural Networks*, volume 2, pages 62–70, San Diego, CA., 1990.

[8] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A Neural Network Identifies Sex From Human Faces. In David Touretzky, editor, *Neural Information Processing Systems*, volume 3, pages 572–577, San Mateo, CA., 1991. Morgan Kaufmann.

[9] V. Govindaraju, D. B. Sher, and S. N. Srihari. Locating human faces in newspaper photographs. In *Proc. of IEEE–CS Conf. Computer Vision and Pattern Recognition*, pages 278–285, San Diego, CA, 1989.

[10] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, New York, 1991.

[11] R. A. Hutchinson and W. J. Welch. Comparison of Neural Networks and Conventional Techniques for Feature Location in Facial Images. In *First IEE Int. Conf. on Artificial Neural Networks*, pages 201–205, London, U.K., 1989.

[12] T. Kohonen, P. Lehtio, E. Oja, A. Kortekangas, and K. Makisara. Demonstration of Patter Processing Properties of the Optimal Associative Mappings. In *Proceedings International Confernece on Cybernetics and Society*, Washington, D.C., 1977.

[13] D. Pomerleau. *Neural Network Perception for Mobile Robot Guidance*. PhD thesis, School of Computer Sience, CMU, February 1992.

[14] W. K. Pratt. *Digital Image Processing*. Wiley–Interscience, New York, 1991.

[15] P. W. Rander. Real–Time Image–Based Face Tracking. Master's thesis, Carnegie Mellon University, Pittsburgh, U.S.A., 1993.

[16] T. Sakai, M. Nagao, and T. Kanade. Computer Analysis and Classification of Photographs of Human Faces. In *First USA–Japan computer conference*, Yokyo, Japan, 1972.

[17] M. Stone. Cross–validation: A review. *Math. Operationsforsch. Statist., Ser. Statistics*, 9(1), 1978.

[18] M. A. Turk and A. P. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, March 1991.

[19] S. Uras, F. Girosi, A. Verri, and V. Torre. A Computational Approach to Motion Perception. *Biological Cybernetics*, 60(2):79–87, 1988.

[20] A. Waibel, M. T. Voe, P. Duchnowski, and S. Manke. Multimodal Interfaces. *Artificial Intelligence Review Journal*, 1994.

[21] G. Wyszecki and W. S. Stiles. *Color Science*. John Wiley & Sons, Inc., New York, 1967.