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


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.

[Signature]

Inhaltsverzeichnis

1 Einführung .................................................. 1
   1.1 Motivation und Aufgabenstellung ........................................ 2
   1.2 Ansatz und Kapitelüberblick ....................................... 2
   1.3 Danksagung .................................................................... 3

2 Verwandte Arbeiten .................................................................. 5

3 Merkmale zur Gesichtslokalisierung ...................................... 7
   3.1 Farbe ........................................................................ 7
   3.1.1 Chromatikdiagramm ................................................ 8
   3.1.2 Farbverteilungen ..................................................... 8
   3.1.3 Farbverschiebungen .................................................. 9
   3.1.4 Gesichtsfarbenklassifikation (FCC) ................................ 11
   3.2 Bewegung .................................................................... 16
   3.3 Zusammenhängende Objekte .......................................... 17
   3.4 Verknüpfung der Merkmale und Objekterkennung ............. 18

4 Aufbau des Gesamtsystems ................................................... 20
   4.1 Lokalisieren eines Gesichtes ........................................... 20
   4.1.1 Interessante Bereiche ................................................ 21
   4.1.2 Auswahl eines Bereiches .......................................... 21
   4.2 Nachführung der Kamera .............................................. 22
   4.2.1 Virtuelle Kamera ...................................................... 22
   4.2.2 Kalibrierung der Kamera .......................................... 23
   4.2.3 Wiederfinden eines Gesichtes .................................... 24
   4.2.4 Anpassung des Farbenklassifikators ......................... 24
   4.3 Das Gesamtsystem ....................................................... 24
   4.4 Umschaltung zwischen den Phasen ................................... 25
   4.5 Beispiel einer Kameranachführung .................................. 26

5 Künstliche neuronale Netze .................................................. 31
   5.1 Aufgabe und Training der Netze ..................................... 31
   5.1.1 Repräsentation der Eingabe ....................................... 32
   5.1.2 Repräsentation der Ausgabe ...................................... 33
   5.1.3 Das Back-Propagation Verfahren ............................... 34
   5.2 Neuronale Netze ohne FCC ........................................... 36
   5.3 Neuronale Netze mit FCC .............................................. 37
6 Künstliche Bilder und Filmsequenzen
   6.1 Trainieren von neuronalen Netzen mit künstlichen Bildern 40
      6.1.1 Anforderungen an die Trainingsmenge 40
      6.1.2 Zuordnung von Netzein- und Ausgabe 41
   6.2 Filmsequenzen 42
   6.3 Generierung künstlicher Bilder 42
      6.3.1 Einrichtung der Datenbasen 42
      6.3.2 Das Blue-Screen Verfahren 43
      6.3.3 Berechnung eines Bildes 44

7 Auswertung 47
   7.1 Testsequenzen 47
      7.1.1 Aufnahme von Testsequenzen 47
      7.1.2 Markierung der Testsequenzen 48
      7.1.3 Auswahl der Testsequenzen 48
      7.1.4 Evaluation einer Testsequenz 49
   7.2 Ergebnisse 50
      7.2.1 Testsequenz 1 51
      7.2.2 Testsequenz 2 52

8 Zusammenfassung 58

9 Ausblick 60
   9.1 Erweiterungen des Systems 60
      9.1.1 Stereosehen 60
      9.1.2 Lokalisieren von Lippen und Augen 60
      9.1.3 Lokalisieren mehrerer Gesichter 60
   9.2 Anwendungen des Systems 60
      9.2.1 Bildtelefon 60
      9.2.2 Gesichtsidentifizierung 61
      9.2.3 Auswertung der Sprecherposition 61
Zusammenfassung

Kapitel 1

Einführung


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].

1.1 Motivation und Aufgabenstellung


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 lokализierten 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 gehört wurde. Umgekehrt kann die Information über Gesichtspositionen dazu verwertet werden, in bestimmte Richtungen hineinzu hören 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


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 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.


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


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 unnaturalen 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


1. Farbe
2. Bewegung
3. zusammenhängendes Objekt

einzel- 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 Digitalisierung 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.


### 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 entspricht einer Normierung der Farbanteile bzgl. der absoluten Summe der RGB-Werte und führt theoretisch zu einer gegenüber Helligkeit invarianter 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 Farbintensität erhalten. Die Farbnormierung bildet die RGB-Werte auf das Chromatikdiagramm gemäß \(f : \mathbb{R}^3 \to \mathbb{R}^2\) mit

\[
q = (r, g) = f(R, G, B)
\]

und

\[
r = \frac{R}{R + G + B}, \quad g = \frac{G}{R + G + B}
\]

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 \(N_{\alpha, \beta}\) des Auftretens jeder der möglichen \(\frac{256^3}{2}\) Farben \(q = (r, g)\) des Chromatikdiagramms ermittelt. Die normierten Häufigkeiten \(\overline{N}_{\alpha, \beta}\) werden als Farbverteilung definiert:

\[
\overline{N}_{\alpha, \beta} = \frac{N_{\alpha, \beta}}{\max_{i+j \leq 255} N_{i, j}}
\]
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.


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


Unter verschiedenen Beleuchtungsverhältnissen wurde dazu desselbe Gesicht mehrfach aufgenommen und jeweils die häufigste Farbe \( \bar{F} = (7,7) \) in der Farbverteilung gemessen. In Tabelle 3.1 sind Durchschnitt und Varianz der gemessenen Werte angegeben:

<table>
<thead>
<tr>
<th>Einstellung der Kamera</th>
<th>Beleuchtungsverhältnisse</th>
<th>Mittelwert</th>
<th>Varianz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatik</td>
<td>Innenaufnahmen</td>
<td>(113,90)</td>
<td>(3,7,2,7)</td>
</tr>
<tr>
<td>Automatik</td>
<td>Außenaufnahmen</td>
<td>(103,78)</td>
<td>(2,9,0,4)</td>
</tr>
<tr>
<td>Automatik</td>
<td>Innen- und Außenaufnahmen</td>
<td>(108,84)</td>
<td>(5,2,5,8)</td>
</tr>
<tr>
<td>Innenaufnahmen</td>
<td>Innenaufnahmen</td>
<td>(107,97)</td>
<td>(1,1,0,7)</td>
</tr>
<tr>
<td>Innenaufnahmen</td>
<td>Außenaufnahmen</td>
<td>(80,80)</td>
<td>(3,1,3,1)</td>
</tr>
<tr>
<td>Außenaufnahmen</td>
<td>Innenaufnahmen</td>
<td>(160,88)</td>
<td>(8,9,4,0)</td>
</tr>
<tr>
<td>Außenaufnahmen</td>
<td>Außenaufnahmen</td>
<td>(112,77)</td>
<td>(0,1,0,5)</td>
</tr>
<tr>
<td>entsprechend Beleuchtung</td>
<td>Innen und Außenaufnahmen</td>
<td>(109,89)</td>
<td>(2,6,9,9)</td>
</tr>
</tbody>
</table>

**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 Individual Face Color Classifier)
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 \( N_{r,g} \) 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 ausgebildet 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 \( N_{r,g} \) mit \((r, g) \in \text{Mundfarbe möglich.}

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


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 fehklassifiziert. 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 Ausschnitts 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 Gesichtausschnittes (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: Vergrößerter Bildausschnitt](image)

![Abbildung 3.7: IFCC mit wenigen Pixeln](image)

(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 vergrößerten 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.

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

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

Genereller FCC


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 klasifizieren. 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


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

Abbildung 3.13: Berücksichtigung des Hintergrundes
(a) FCC mit Subtraktion der Hintergrund-Farbenverteilung, (b) Anwendung des FCC, (c) Ausblendung der Nicht-Hauptfarben

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-Farbenverteilung von der Farbverteilung des Gesichtes durchgeführt.

Deduktion des generellen Klassifikators


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.
3.3 Zusammenhängende Objekte

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 Gesichtsbeereiche jedoch genugend strukturiert.


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 zusammenzufassen zu können, werden die Pixel auf Nachbarschaftsbeziehungen untersucht. Im folgenden bezeichnen $x(p)$ und $y(p)$ die Koordinaten eines Pixels $p$.

\textbf{Def. 1:} Zwei Pixel $p_1$ und $p_2$ heißen \textit{benachbart}, wenn ihre Koordinaten der Bedingung $\|x(p_1) - x(p_2)\| \leq 1 \land \|y(p_1) - y(p_2)\| \leq 1$ genügen.

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

1. $\mathcal{M} = \{p\}$ 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 $q$.

\textbf{Def. 3:} Ein Pixel $p$ hat das Merkmal

- \textit{Gesichtsfarbe bzgl. eines FCC's}, wenn $\overline{N}_{r,g} \geq \alpha$ gilt, mit $p = (r, g)$ und $N$ der Farbverteilung des FCC's.

- \textit{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 einer 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 $\alpha$ und $\beta$ 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._


### 3.4 Verknüpfung der Merkmale und Objekterkennung

Mit den Definitionen 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 GFCN 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 Granuwert 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.

Abbildung 3.14: Lokalisation 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 Kameraposition und Objektivbrennweite werden laufend angepaßt.


4.1 Lokalisieren eines Gesichtes


4.1.1 Interessante Bereiche


4.1.2 Auswahl eines Bereiches


Abbildung 4.1: Lokalisation

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 Objekttes des vorhergehenden Bildes gefunden, kann die Gesichtssuche erfolgreich abgeschlossen werden und die Phase der Kameranachführung beginnt.
4.2 Nachführung der Kamera


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 Verarbeitungs geschwindigkeit 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.


Um Bewegungen zu erkennen, muß das zeitlich vorhergehende Bild bekannt sein. Da jedoch nicht das Gesamtimage, sondern nur die virtuelle Kamera eingelesen wird, erneut bei einer Verschiebung derselben eine Informationsstrecke. Solange die physikalische Kamera in ihren Einstellungen nicht verändert wird, wird der Hintergrund als konstant angenommen und ein einmaliges Einlesen des Gesamtimages 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 Gesamtimage verändert hat, wird Bewegung nicht nur bei der Person, sondern durch die Veränderung auch im Hintergrund detektiert. Eine sehr hoher Prozentsatz veränderter Pixel deuten daher auf eine Inkonsistenz im gespeicherten Gesamtimage und veranlassen 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 Gesamtimage 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 Brennweiteeinstellung 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.


4.2.3 Wiederfinden eines Gesichtes


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 Belichtungsverhältnissen anzupassen.

4.3 Das Gesamtsystem

4.4 Umschaltung zwischen den Phasen


![Diagramm der Umschaltung zwischen den Phasen](image)

**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, gefunden 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


Abbildung 4.4: Beispielsequenz, Teil 1
Abbildung 4.5: Beispielsequenz, Teil 2
Abbildung 4.6: Beispielsequenz, Teil 3
Abbildung 4.7: Lokalisierte Gesichter
Kapitel 5

Künstliche neuronale Netze


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 Erkennungsarten 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 Änderung 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 projiziert 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 zwischen den Phasen **Lokalisierung** und **Kameranachführung** des Gesamtsystems entscheidend ist.

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.
5.1. Aufgabe und Training der Netze

- **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


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 konnten 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 schlechterere 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

In allen folgenden Netzen wurde daher die normalverteilte Darstellung gewählt.

5.1.3 Das Back-Propagation Verfahren


Ein unabhängiges Testset zeigt nach jedem Durchlauf der Trainingsmenge die Generalisierungsähigkeit des Netzes an. Das Training wird abgebrochen, wenn das Netz beginnt, die gezeigten Trainingsbeispiele anwendend zu lernen. Die Leistungsfähigkeit des Netzes wird anschließend von einem weiteren unabhängigen Validitätstest 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 \( \Theta_j \) 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 Assoziations \( d_j \) verglichen und ein der Differenz entsprechender Fehler, hier der Mean Square Error, berechnet:

\[
E = \frac{1}{2} \sum_j (o_j - d_j)^2
\]

(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:

\[
\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial o_j} \cdot \frac{\partial o_j}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{kj}}
\]

(5.5)

Die einzelnen Gradienten lassen sich leicht berechnen:

\[
\frac{\partial E}{\partial o_j} = o_j - d_j
\]

(5.6)

\[
\frac{\partial o_j}{\partial a_j} = f'(a_j) = o_j \cdot (1 - o_j)
\]

(5.7)

\[
\frac{\partial a_j}{\partial w_{kj}} = o_k
\]

(5.8)

Damit ergibt sich:

\[
\frac{\partial E}{\partial w_{kj}} = (o_j - d_j) \cdot o_j \cdot (1 - o_j) \cdot o_k = \delta_j \cdot o_k
\]

(5.9)

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

\[
\delta_j = \frac{\partial E}{\partial a_j}
\]

(5.10)

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

Um den Fehlergradienten \( \delta_j \) auch für Neuronen einer tiefer liegenden Schicht bestimmen zu können, wird der Teiltterm 5.6 aus den Fehlern der Neuronen der höheren Schicht berechnet:

\[
\frac{\partial E}{\partial o_j} = \sum_h \frac{\partial E}{\partial a_h} \cdot \frac{\partial a_h}{\partial o_j} = \sum_h \delta_h w_{jh}
\]

(5.11)

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

\[
\Delta w_{kj}(t + 1) = -\eta \frac{\partial E}{\partial w_{kj}(t)}
\]

(5.12)
Im Allgemeinen wird zusätzlich ein Momentum $\alpha$ verwendet, das zu jedem Trainingsschritt einen Teil der letzten Gewichtsänderung aufaddiert.

$$\Delta w_{kj}(t+1) = \alpha \cdot \Delta w_{kj}(t) - \eta \frac{\partial E}{\partial w_{kj}(t)}$$ (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 $\eta$ im Verlauf des Trainings verringert und das Momentum $\alpha$ 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 hochstenerzielen Erkennungsraten für die Netzstruktur aus Bild 5.1 aufgelistet. Die Erkennungsraten ergibt sich aus dem prozentualen Anteil der richtig klassifizierten Bilder aus dem Validitätssatz, 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 geschätzte 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 Retinafläche trainiert. Einige Beispiele für Grauwertbilder in dieser Auflösung sind in Bild 6.6 (a) abgebildet. Folgende Netze wurden mit unterschiedlich 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.

<table>
<thead>
<tr>
<th>Netz</th>
<th>Retina</th>
<th>Ausgabe</th>
<th>Anzahl Gewichte</th>
<th>Erkennungsraten</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32x32</td>
<td>X, Y, Größe</td>
<td>40000</td>
<td>32%</td>
</tr>
<tr>
<td>2</td>
<td>24x24</td>
<td>Größe</td>
<td>23000</td>
<td>65%</td>
</tr>
<tr>
<td>3</td>
<td>2x24x24</td>
<td>Größe</td>
<td>46000</td>
<td>68%</td>
</tr>
</tbody>
</table>

*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

5.3 Neuronale Netze mit FCC


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.

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:

<table>
<thead>
<tr>
<th>Netz</th>
<th>Retina</th>
<th>Ausgabe</th>
<th>Anzahl Gewichte</th>
<th>Erkennungsrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16x16</td>
<td>X, Y</td>
<td>21000</td>
<td>93.6%</td>
</tr>
<tr>
<td>2</td>
<td>12x12</td>
<td>Größe</td>
<td>12000</td>
<td>95.2%</td>
</tr>
</tbody>
</table>

*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


6.1.1 Anforderungen an die Trainingsmenge

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


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


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.


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.


6.3 Generierung künstlicher Bilder


6.3.1 Einrichtung der Datenbasen

6.3. Generierung künstlicher Bilder

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


Abbildung 6.2: Beispielbilder aus der Datenbasis

Eine weitere Datenbasis wurde mit 10 Hintergrundbildern der Auflösung 600·460 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
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.


Abbildung 6.3: Farbverteilung des Hintergrundes

Abbildung 6.4: Ausschneiden eines Gesichtes

6.3.3 Berechnung eines Bildes

Abbildung 6.5: Berechnung eines künstlichen Bildes

(a) zufälliges Auswählen eines Hintergründes 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 · 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


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 verändernde 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


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:


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


### 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 Justierung 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 Leistungswerte 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 Gespräches 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 \( y_{\text{max}} = 1.88 \) zeigt, daß die maximale Bewegung eines Gesichtes in dieser Testsequenz in vertikaler Richtung das 1.88 fache der Gesichtshöhe beträgt.

<table>
<thead>
<tr>
<th>Testsequenz</th>
<th>Anzahl Bilder</th>
<th>( \bar{x} )</th>
<th>( \bar{y} )</th>
<th>( x_{\text{max}} )</th>
<th>( y_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1860</td>
<td>0.19</td>
<td>0.06</td>
<td>1.52</td>
<td>1.88</td>
</tr>
<tr>
<td>2</td>
<td>460</td>
<td>0.19</td>
<td>0.05</td>
<td>1.26</td>
<td>0.56</td>
</tr>
</tbody>
</table>

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


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 Gesichtes. 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 arithmetische Mittel folgender Werte über alle Bilder einer Testsequenz:

$$s_x = \frac{\sum (x_{output} - x_{target})}{w_{target}}$$
$$s_y = \frac{\sum (y_{output} - y_{target})}{h_{target}}$$
$$s_w = \frac{\sum (w_{output} - w_{target})}{w_{target}}$$
$$s_h = \frac{\sum (h_{output} - h_{target})}{h_{target}}$$

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:

$$\sigma_x = (\frac{\sum (w_{output} - w_{target})}{w_{target}})^2$$
$$\sigma_y = (\frac{\sum (h_{output} - h_{target})}{h_{target}})^2$$

Unter Annahme einer Normalverteilung der Abweichungen der Lokalisierungsdaten kann die Genauigkeit berechnet werden, die für einen vorgegebenen Prozentsatz aller Lokalisierungsdaten 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 $[s^2 - 2\sigma, s^2 + \sigma]$ und in 68.3% aller Fälle ein Faktor von $[s^2 - \sigma, s^2 + \sigma]$ 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 die Nutzung 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.

<table>
<thead>
<tr>
<th>Modus</th>
<th>correct</th>
<th>$s_p$</th>
<th>$s_g$</th>
<th>$s_w$</th>
<th>$s_b$</th>
<th>$\sigma_x$</th>
<th>$\sigma_y$</th>
<th>$\sigma_w$</th>
<th>$\sigma_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89.5%</td>
<td>0.11</td>
<td>0.10</td>
<td>1.31</td>
<td>1.23</td>
<td>0.08</td>
<td>0.07</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>96.1%</td>
<td>0.09</td>
<td>0.08</td>
<td>1.26</td>
<td>1.27</td>
<td>0.06</td>
<td>0.05</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>96.1%</td>
<td>0.07</td>
<td>0.09</td>
<td>1.24</td>
<td>1.29</td>
<td>0.06</td>
<td>0.06</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>68.6%</td>
<td>0.16</td>
<td>0.17</td>
<td>1.15</td>
<td>1.18</td>
<td>0.10</td>
<td>0.10</td>
<td>0.38</td>
<td>0.31</td>
</tr>
</tbody>
</table>

**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.


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 · σ gelten. Z.B. kann für 91.7% der Zeit eine geringere Abweichung der X-Position als 10% der Kopfbreite und eine Schätzung der Kopfbreite um das 1.04 bis 1.48 fache des tatsächlichen Wertes erwartet werden.

<table>
<thead>
<tr>
<th>Modus</th>
<th>Zeit</th>
<th>(x)</th>
<th>(y)</th>
<th>(w)</th>
<th>(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ohne NN</td>
<td>65.4%</td>
<td>&lt;0.13</td>
<td>&lt;0.15</td>
<td>1.13,1.36</td>
<td>1.18,1.41</td>
</tr>
<tr>
<td>ohne NN</td>
<td>91.7%</td>
<td>&lt;0.19</td>
<td>&lt;0.21</td>
<td>1.04,1.48</td>
<td>1.08,1.54</td>
</tr>
<tr>
<td>mit NN</td>
<td>46.7%</td>
<td>&lt;0.26</td>
<td>&lt;0.27</td>
<td>0.88,1.50</td>
<td>0.95,1.46</td>
</tr>
<tr>
<td>mit NN</td>
<td>65.5%</td>
<td>&lt;0.30</td>
<td>&lt;0.37</td>
<td>0.70,1.89</td>
<td>0.77,1.81</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>Bilder/sec</th>
<th>correct</th>
<th>(s_x)</th>
<th>(s_y)</th>
<th>(s_w)</th>
<th>(s_h)</th>
<th>(\sigma_x)</th>
<th>(\sigma_y)</th>
<th>(\sigma_w)</th>
<th>(\sigma_h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>89.5%</td>
<td>0.11</td>
<td>0.09</td>
<td>1.24</td>
<td>1.31</td>
<td>0.07</td>
<td>0.06</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>96.1%</td>
<td>0.07</td>
<td>0.09</td>
<td>1.24</td>
<td>1.29</td>
<td>0.06</td>
<td>0.06</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td>98.8%</td>
<td>0.08</td>
<td>0.1</td>
<td>1.23</td>
<td>1.27</td>
<td>0.05</td>
<td>0.07</td>
<td>0.12</td>
<td>0.14</td>
</tr>
</tbody>
</table>

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 Testsequenz 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.

<table>
<thead>
<tr>
<th>Modus</th>
<th>correct</th>
<th>$s_x$</th>
<th>$s_y$</th>
<th>$s_w$</th>
<th>$s_h$</th>
<th>$\sigma_x$</th>
<th>$\sigma_y$</th>
<th>$\sigma_w$</th>
<th>$\sigma_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81.7%</td>
<td>0.85</td>
<td>0.38</td>
<td>2.79</td>
<td>2.06</td>
<td>0.67</td>
<td>0.32</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>100.0%</td>
<td>0.09</td>
<td>0.1</td>
<td>1.39</td>
<td>2.06</td>
<td>0.05</td>
<td>0.07</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>100.0%</td>
<td>0.08</td>
<td>0.12</td>
<td>1.32</td>
<td>2.01</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>0.14</td>
</tr>
</tbody>
</table>

**Tabelle 7.5: Ergebnisse für Testsequenz 2**


<table>
<thead>
<tr>
<th>Modus</th>
<th>correct</th>
<th>$s_x$</th>
<th>$s_y$</th>
<th>$s_w$</th>
<th>$s_h$</th>
<th>$\sigma_x$</th>
<th>$\sigma_y$</th>
<th>$\sigma_w$</th>
<th>$\sigma_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.8%</td>
<td>0.27</td>
<td>0.31</td>
<td>1.43</td>
<td>1.44</td>
<td>0.18</td>
<td>0.24</td>
<td>0.42</td>
<td>0.34</td>
</tr>
<tr>
<td>2</td>
<td>100.0%</td>
<td>0.13</td>
<td>0.15</td>
<td>1.21</td>
<td>1.31</td>
<td>0.08</td>
<td>0.07</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>100.0%</td>
<td>0.12</td>
<td>0.1</td>
<td>1.13</td>
<td>1.28</td>
<td>0.08</td>
<td>0.07</td>
<td>0.17</td>
<td>0.16</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>Modus</th>
<th>Zeit</th>
<th>$x$</th>
<th>$y$</th>
<th>$w$</th>
<th>$h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ohne NN</td>
<td>86.3%</td>
<td>$&lt;0.14$</td>
<td>$&lt;0.17$</td>
<td>$[1.27,1.52]$</td>
<td>$[1.87,2.27]$</td>
</tr>
<tr>
<td>ohne NN</td>
<td>95.5%</td>
<td>$&lt;0.19$</td>
<td>$&lt;0.24$</td>
<td>$[1.16,1.66]$</td>
<td>$[1.70,2.50]$</td>
</tr>
<tr>
<td>mit NN</td>
<td>86.3%</td>
<td>$&lt;0.21$</td>
<td>$&lt;0.22$</td>
<td>$[1.05,1.40]$</td>
<td>$[1.16,1.48]$</td>
</tr>
<tr>
<td>mit NN</td>
<td>95.5%</td>
<td>$&lt;0.29$</td>
<td>$&lt;0.29$</td>
<td>$[0.90,1.62]$</td>
<td>$[1.02,1.68]$</td>
</tr>
</tbody>
</table>

**Tabelle 7.7**: Genauigkeit bei Testsequenz 2


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 Lokalisation 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


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:


- Das Gesicht befindet sich vor einem Hintergrund, der viele der im Gesicht auftretenden Farben aufweist. Die neuronalen Netze führen durch die Formerkennung zu einer Verbesse rung 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 bedeu-
ten. 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 Stereossehen


9.1.2 Lokalisieren von Lippen und Augen

Die Methoden der Farbklassifizierung erlauben 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

9.2.2 Gesichtsidentifizierung


9.2.3 Auswertung der Sprecherposition

In Verbindung mit einem Mikrofon-Array können akustische Sprachsignale aus der Richtung herausgefiltert 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


