Expl. Prof. Waibel

Automatische Lokalisierung der Lippenregion in Videobildern von Gesichtern

Diplomarbeit
von
Dietrich Büsching
31. August 1994

Universität Karlsruhe
Fakultät für Informatik
Institut für Logik, Komplexität und Deduktionssysteme

Lehrstuhl Prof. Dr. A. Waibel

Betreuer: Dr. Paul Duchnowski

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig erstellt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 31.08.1994.

Zusammenfassung

In der vorliegenden Diplomarbeit werden Verfahren zur Detektion der Lippenwinkel in Videobildern von Gesichtern beschrieben und verglichen. Die Verfahren setzen sich aus Groberkennung des Mundbereiches und Feinerkennung der Lippenwinkel zusammen.

Für die Groberkennung wird ein dreischichtiges Perzeptron eingesetzt. Die Eingabeschicht nimmt das verkleinerte und vorverarbeitete Eingangsbild auf. Die Ausgabeschicht besteht aus einem Gitter von Ausgabeeinheiten, die mögliche Positionen des Mundes im Bild repräsentieren. Um eine Detektion für verschiedene Positionen, Kopfgrößen und Personen zu ermöglichen, werden aus einer relativ geringen Anzahl von Ausgangsbildern durch Verschieben und Skalieren des Gesichtes eine große Zahl von Trainingsbildern erzeugt.

In der Feinerkennung werden die Lippenwinkel in einem rechteckigen Ausschnitt um die in der Groberkennung geschätzte Position des Mundes gesucht. Dazu wurden drei Verfahren getestet. Das erste Verfahren benutzt die in der Groberkennung eingesetzte Architektur in leicht abgewandelter Form. Das zweite Verfahren vergleicht kleinere Teile der Zielregion mit Grauwerttemplates der Mundwinkel. Beim Vergleich wird die normalisierte Kreuzkorrelation zwischen Bildausschnitt und Template berechnet. Im dritten Verfahren wird das Eingabefeld eines neuronalen Netzes an verschiedene Positionen des zu durchsuchenden Bildausschnitts plaziert und die Aktivierung der Ausgabeeinheiten als Maß für das Vorhandensein der Mundwinkel an der jeweiligen Position gewertet.

Das erste Verfahren liefert für die personenabhängige Erkennung (Trainingsperson gleich Testperson) sehr gute Ergebnisse und ist schneller als die anderen beiden Verfahren. Die normalisierte Kreuzkorrelation liefert insbesondere bei der personenunabhängigen Erkennung unzuverlässigere Schätzungen als die auf neuronalen Netzen basierenden Methoden. Das dritte Verfahren erzielt bei der personenunabhängigen Erkennung (Trainingspersonen ungleich Testpersonen) bessere Ergebnisse als das erste Verfahren, ist aber langsamer. Abschließend wird gezeigt, wie durch die Verwendung rezeptiver Felder mit gleichen Gewichten das dritte Verfahren erheblich beschleunigt werden kann.

Inhaltsverzeichnis

1 Einleitung und Beschreibung der Aufgabenstellung

1.1 Zweck der Lippenlokalisation: Lippenlesen

Die automatische Spracherkennung ist ein sehr schwieriges Problem. Obwohl durch Techniken wie künstliche neuronale Netze und Hidden Markow Models wesentliche Fortschritte erzielt worden sind, erreichen automatische Verfahren nicht die menschliche Leistungsfähigkeit in diesem Gebiet. Ein Grund für die schlechteren Leistungen automatischer Spracherkenner ist ihre Fehleranfälligkeit gegenüber Hintergrundgeräuschen und dem gleichzeitigen Sprechen mehrerer Personen.

Nach Bregler et al. (93) gibt es zwei wesentliche Arten von Zusatzinformation, die von Menschen bei der Auswertung von akustischen Sprachsignalen geringer Qualität herangezogen werden können. Zum einen sind dies Informationen von höheren Abstraktionsebenen, die Syntax und Semantik der gesprochenen Worte betreffen. In automatischen Spracherkennern wird versucht, diese Zusatzinformation in Form von Grammatiken oder Sprachmodellen zur Verfügung zu stellen. Die andere Möglichkeit besteht darin, schon auf der Ebene der Erkennung von Phonemen Zusatzinformationen zu nutzen. Der Mensch tut dies beispielsweise, indem er die Lippenbewegungen des Sprechers beobachtet. Einige Menschen ohne Hörvermögen sind sogar in der Lage, Sprache allein mit dieser visuellen Information und zusätzlicher Kontextinformation zu verstehen (Lippenlesen).

Die folgenden Informationen zum Lippenlesen stammen aus Garcia et al. (92). Die visuelle Information des Lippenlesens ist besonders hilfreich bei der Erkennung von Konsonanten. Insgesamt ist aber der Informationsgehalt der Lippenbewegungen geringer als die akustische Information. Dies ist auch an der Anzahl der Viseme, elementaren Lippenbewegungen, die den akustischen Phonemen entsprechen, zu erkennen. Einige Phoneme sind visuell nicht zu unterscheiden (z.B. /p/, /b/ und /m/). Daher ist beispielsweise im Englischen die Anzahl der Viseme um ein Drittel geringer als die Anzahl der Phoneme. Durch Heranziehung von Kontextinformation sind einige taube Menschen dennoch in der Lage, ihren Gesprächspartner nur anhand der Lippenbewegungen zu verstehen. Zum Erlernen des Lippenlesens ist allerdings im Gegensatz zum Lernen akustischer Sprache Wissen über die Regeln der zu erkennenden Sprache unabdingbar. Dies könnte am geringen Informationsgehalt der Lippenbewegungen gegenüber den akustischen Sprachsignalen liegen. Ein interessantes Phänomen, das die Bedeutung des Lippenlesens auch für nicht hörgeschädigte Menschen unterstreicht, ist der McGurk-Effekt. Wenn einer Versuchsperson der Laut /bi/ akustisch und der Laut /gi/ visuell synchron präsentiert werden, nimmt sie ein /di/ wahr. Dies läßt sich dadurch erklären, daß das /gi/ hinten, das /bi/ vorn und das /di/ in der Mitte des Sprachtraktes erzeugt wird. Die Signale /gi/ und /bi/ werden zu einem Mittelwert verschmolzen, der zur Erkennung des /di/ führt.

Beim Entwurf eines Systems zum maschinellen Lippenlesen taucht die Frage nach den Merkmalen auf, die am besten die visuelle Information der Lippenbewegungen

wiedergeben und als Eingabedaten für das System dienen können. In den existierenden Systemen zum Lippenlesen wurden verschiedene Eingabemerkmale verwendet.

Petajan et al. benutzten binäre Bilder des Mundes, die durch die Festlegung eines Schwellwertes aus Grauwertbildern erzeugt wurden. Pentland und Mase bestimmten mit Techniken des optischen Flusses 4 Geschwindigkeitswerte (obere und untere Lippe und die beiden Mundwinkel). Stork et al. (92) plazierten 10 Markierungen auf dem Gesicht eines Sprechers. Diese Markierungen konnten mit einem kommerziell erhältlichen Bildverarbeitungssystem mit hoher Frequenz (60 Hz) verfolgt werden. Stork et al. berechneten daraus fünf Maßzahlen als Eingabe für die visuelle Spracherkennung. Der an der Universität Karlsruhe entwickelte Ansatz (Bregler et al. 93, Duchnowski et al. 94) benutzt als Eingabewerte die nur wenig vorverarbeiteten Grauwerte innerhalb eines rechteckigen Rahmens, der die Mundregion umfaßt. Da zur Spracherkennung eine konnektionistische Architektur benutzt wird, die mit dem Backpropagationalgorithmus trainiert wird, besteht die Möglichkeit, den Lernvorgang bis auf die Ebene der Erkennung von Merkmalen der räumlichen und zeitlichen Veränderung von Grauwerten auszudehnen. Man erwartet dabei, daß durch den Lernalgorithmus Merkmale gefunden werden, die die Information beeinhalten, die für die Spracherkennung wichtig ist. Demgegenüber ist man bei der Vorgabe von Merkmalen durch einen menschlichen Systementwickler, von dessen korrektem Wissen über gute Merkmale und der nicht immer trivialen Konstruktion von zuverlässig arbeitenden Merkmalsdetektoren abhängig.

1.2 Objekterkennung für ein realistisches Szenario des Lippenlesens

Das in Karlsruhe entwickelte System zum automatischen Lippenlesen erfordert, daß sich der Mund des Sprechers innerhalb des Eingabefensters befindet. Zur Zeit wird dies dadurch erreicht, daß während der Aufnahme von Trainings- oder Testsequenzen von Mundbildern das gesamte von der Kamera aufgenommene Videobild auf einem Monitor dargestellt wird. Ein eingeblendeter Rahmen zeigt den Ausschnitt des Bildes an, der für das Lippenlesen verwendet wird. Der Sprecher muß bei der Aufnahme darauf achten, daß sein Mund sich stets innerhalb dieses Fensters befindet. Bei einem praktischen Einsatz eines Systems zum Lippenlesen kann man natürlich kaum solche Anforderungen an den Sprecher stellen. Realistischer wäre es, anzunehmen, daß sich das Gesicht des Sprechers irgendwo im Blickbereich der Kamera befindet. Zudem sollte der Sprecher nicht gezwungen sein, seine Position beizubehalten. Diese Szenario erfordert visuelle Objekterkennungsmechanismen, die das Gesicht und insbesondere die Lippen eines Sprechers lokalisieren und verfolgen.

Zur Erreichung dieses Zieles bietet sich ein zweistufiges Vorgehen an. In einem ersten Schritt wird das Gesicht des Sprechers lokalisiert, im zweiten Schritt wird dann im Bereich des Gesichtes nach den Lippen gesucht. Die erste Teilaufgabe wurde von Martin Hunke im Rahmen seiner Diplomarbeit (Hunke 94) an der Carnegie Mellon University gelöst. Die zweite Teilaufgabe ist Gegenstand dieser Diplomarbeit. Durch sie wird die Lücke zwischen grober Gesichtserkennung und Lippenle-

sen geschlossen. Die Aufgaben dieses Zwischenstückes lassen sich folgendermaßen genauer definieren: Der Gesichtserkenner von Hunke (eine kurze Beschreibung des eingesetzten Verfahrens findet sich in Kapitel 2) liefert als Ausgabe die Koordinaten der Ecken zweier rechteckiger Fenster im Videobild. Das Suchfenster ist der Bereich, in dem der Gesichtserkenner in einem neuen Suchschritt das Gesicht suchen wird. Es ist so groß gewählt, daß sich das verfolgte Gesicht nur bei extrem schnellen Bewegungen des Kopfes im nächsten Suchschritt (nach ca. 0.2 s) nicht in diesem Fenster befindet. Folglich kann man auch für vom Lippenerkenner evtl. in der Zeit zwischen zwei Ausgaben des Gesichtserkenners aufgenommenen Bildern davon ausgehen, daß das Gesicht sich innerhalb des letzten vom Gesichtserkenner gelieferten Suchfensters befindet. Das zweite vom Gesichtserkenner berechnete Fenster umfaßt die ungefähre Position des lokalisierten Gesichtes. Mit Hilfe dieser Informationen soll das Lippendetektionssystem die Position des Mundes aus aufgenommenen Videobildern bestimmen. Da zum Lippenlesen eine möglichst genaue Positionsangabe erforderlich ist, wurde die Aufgabe so präzisiert, daß das Lippenerkennungssystem die Position der Mundwinkel ausgeben soll. Abbildung 1 gibt einen Überblick über die Struktur des Gesamtsystems.

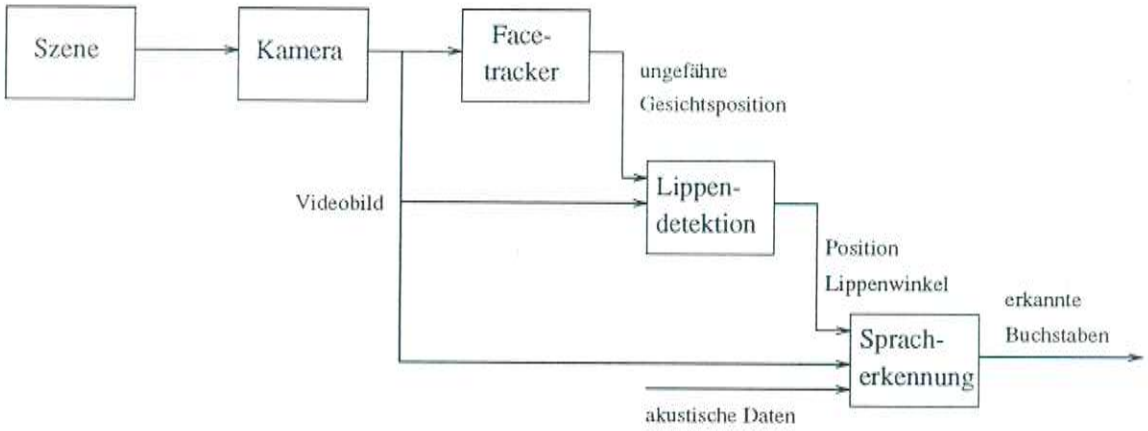

Abbildung 1: Gesamtsystem Lippenlesen

Ähnliche Probleme wurden auch in den in Kapitel 3.3 kurz zusammengefaßten Arbeiten gelöst. Im vorliegenden Fall kommen zwei die Aufgabe erschwerende Zusatzanforderungen hinzu.

- Schwach kontrollierte Beleuchtung und Objekte im Hintergrund

  Die in Kapitel 3.3 zusammengefaßten Arbeiten verwendeten fast ausschließlich Bilder von künstlich frontal beleuchteten Personen mit untexturiertem Hintergrund (z.B. ein Wand). Dies schränkt die Möglichkeit von falschen Positionshypothesen praktisch auf den Bereich des im Bild sichtbaren Körpers der Person ein. Die gezielte Beleuchtung führt zu kontrastreichen Bildern. Um dem oben erwähnten Szenario möglichst nahe zu kommen, wurde keine zusätzliche

Lichtquelle benutzt und als Hintergrund die normale Arbeitsumgebung im Datensammelraum der Spracherkennungsgruppe in Karlsruhe gewählt. Es wurde lediglich darauf geachtet, daß keine Schlagschatten Teile des Gesichtes völlig abdunkelten.

- Erkennung eines in sehr verschiedenen Ausprägungen auftretenden Objektes

  Der Mund kann in sehr unterschiedlichen Öffnungsgraden in den zu verarbeitenden Bildern auftreten. Bei zusammengepressten Lippen ist nur ein dünner dunkler Strich zu erkennen. Bei einem weit offenem Mund fällt vor allem ein großes dunkles Loch auf. Es ist allerdings auch möglich, daß durch die Zähne die Region zwischen den Lippen heller als die Umgebung ist.

2 Grundlegende Bildverarbeitungstechniken: Korrelation und Kantendetektion

In diesem Kapitel möchte ich zwei Techniken aus der Bildverarbeitung beschreiben, die zum Verständnis dieser Arbeit besonders wichtig sind. Die Korrelation von Templates mit Bildern, in denen nach einem Zielobjekt gesucht wird, ist ein sehr gebräuchliches Verfahren bei der Lokalisierung von Gesichtern und Gesichtsteilen. Im nächsten Kapitel werden entsprechende Arbeiten kurz beschrieben. Kantendetektion ist traditionell eines der Grundprobleme der Bildauswertung. Die Detektion von gerichteten Kanten ist ein Vorverarbeitungsverfahren, das die Erkennung des Mundes wesentlich erleichtert. Es wird in dieser Diplomarbeit sehr häufig verwendet.

2.1 Korrelation mit Templates

In korrelationsbasierten Verfahren (template matching) wird ein Bild des zu suchenden Objektes mit den Grauwerten an verschiedenen Positionen des zu durchsuchenden Bildes verglichen. Dieser Vergleich läßt sich nach Duda und Hart (73) folgendermaßen formalisieren:

g(i,j) sei das Bild, in dem gesucht wird. t(i,j) sei das Bild des zusuchenden Objektes - das Template. D sei der Bereich, in dem für das Template definierte Werte vorliegen.

Dann läßt sich das Maß der Übereinstimmung an der Stelle (m,n) als

$$M(m,n) = \sum_{(i-m,j-n)\in D} |g(i,j) - t(i-m, j-n)|$$

berechnen. Häufiger wird aber die Kreuzkorrelation benutzt. Sie ist definiert als:

$$R_{gt}(m,n) = \sum_{(i-m,j-n)\in D} g(i,j)t(i-m, j-n)$$

Diese beiden Maßzahlen geben ungefähr die gleichen Ergebnisse, wenn die Energie des Bildes $g^2(i,j)$ über das gesamte Bild annähernd gleich ist. Da dies für Bilder natürlicher Szenen nur selten der Fall ist und die Ergebnisse der reinen Kreuzkorrelation bei Nichteinhaltung dieser Bedingung häufig unbefriedigend sind, wird der Kreuzkorrelationswert meistens bezüglich Helligkeit und Kontrast normalisiert. Die normalisierte Kreuzkorrelation ist:

$$C_N(m,n) = \frac{((I_T(m,n)\cdot T)/ \text{ Größe von T }) - E(I_T(m,n))\cdot E(T)}{\sigma(I_T(m,n))\cdot\sigma(T)}$$

$I_T(m,n)$ ist dabei der Bereich des Bildes, der mit dem Template multipliziert wird. T ist das Template. Das Skalarprodukt $I_T(m,n)\cdot T$ entspricht der einfachen Kreuzkorrelation zwischen Bildauschnitt und Template. Durch Division durch die Templategröße wird der Korrelationswert bezüglich der Templategröße normalisiert.

Durch die Subtraktion des Produktes der mittleren Helligkeiten von Bildausschnitt und Template $E(I_T(m,n)) \cdot E(T)$ wird der Einfluß von Unterschieden in der Gesamthelligkeit von Bildausschnitt und Template beseitigt. Schließlich wird durch Division durch die Standardabweichungen $\sigma(I_T(m,n))$ und $\sigma(T)$ des Grauwertes der Korrelationswert bezüglich des Kontrastes normalisiert. Diese Definition der normalisierten Kruzkorrelation lehnt sich an die Definition des Korrelationskoeffizienten zwischen zwei Zufallsvariablen aus der Statistik an. Genau wie der Korrelationskoeffizient kann auch die normalisierte Kreuzkorrelation Werte zwischen -1,0 und 1,0 annehmen.Falls $I_T(m,n) = T$ gilt, ergibt sich:

$$C_N(m,n) = \frac{E(t^2(i,j)) - E^2(t(i,j))}{Var(t(i,j))} = 1.0 \qquad ((i,j) \in D)$$

Korrelationsbasierte Techniken haben den Nachteil, daß sie sehr rechenaufwendig sind. Durch ein hierarisches Suchverfahren kann der Aufwand verringert werden. Dabei wird zunächst eine stark verkleinerte Kopie des Templates in einer ebenso verkleinerten Kopie des Bildes gesucht. Nur Positionen, an denen der Korrelationswert eine bestimmte Stärke erreicht, werden zur Suche in höheren Auflösungen verwendet. Dieses Vorgehen hat auch den Vorteil, daß bei geringerer Auflösung kleinere Abweichungen des Aussehens eines Objektes in Template und dem zu verarbeitenden Bild den Korrelationswert nicht so stark beinflußen.

2.2 Detektion gerichteter Kanten

Ein häufig verwendetes und recht einfaches Verfahren zur Detektion von Kanten ist die Faltung mit dem Sobeloperator. Dieser besteht aus zwei 3×3 Masken, mit denen eine Approximation des Grauwertgradienten in x- bzw. y-Richtung ermittelt werden kann (siehe Abbildung 2). Die Richtung ϕ des maximalen Gradienten läßt sich damit nach der Formel $\tan(\phi) = \frac{g_y}{g_x}$ für $g_x \neq 0$ berechnen. Für $g_x = 0$ gilt: $g_y < 0 \Rightarrow \phi = -90^o$ und $g_y > 0 \Rightarrow \phi = 90^o$. Siehe dazu Bild 3. Mit Hilfe der berechnete Richtung des maximalen Gradienten ist es möglich, die Bildpixel anhand ihrer Kantenrichtung zu klassifizieren. In dieser Arbeit werden vertikale und vor allem horizontale Kantenfelder verwendet. Ein Bildpixel wird dabei z.B. dem horizontalen Kantenfeld zugeordnet, wenn die Kantenrichtung in einer gewissen Bandbreite (z.B. $\pm 45^o$) um die Horizontale herum liegt.

Nach Kitchen und Malin (89) beträgt der maximale Fehler bei der Berechnung der Richtung des maximalen Gradienten mit dem Sobeloperator für eine beliebig in der Ebene orientierte Stufenfunktion (z.B. eine Halbebene weiß, die andere schwarz) 3,7 Grad. Dieser Fehler entsteht durch die Approximation der Ableitungsoperation durch eine Faltung mit einer Maske mit einer Rasterung endlicher Größe und letztendlich auch schon durch die Aufnahme des Bildes, bei der der tatsächliche Grauwertverlauf aufgrund der begrenzten räumlichen Abtastfrequenz nur annähernd erfaßt werden kann.

-1	0	1
-2	0	2
-1	0	1

1	2	1
0	0	0
-1	2	-1

Abbildung 2: Die Masken des Sobeloperators für den horizontalen und vertikalen Gradienten

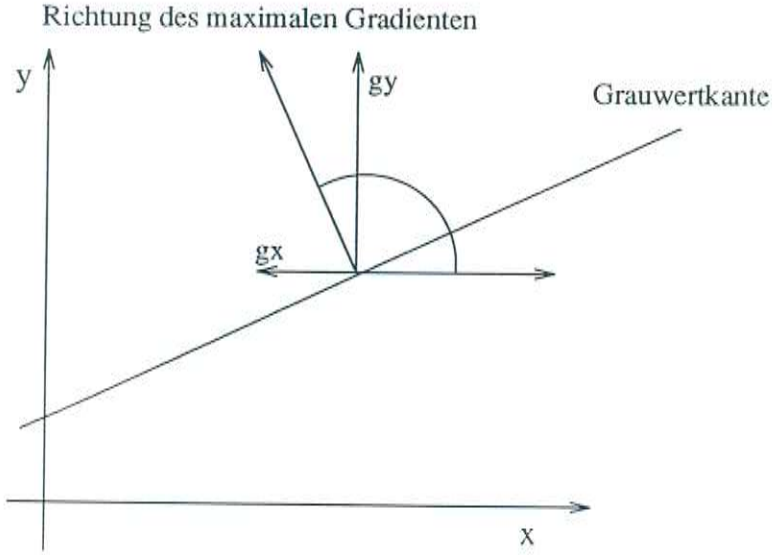

Abbildung 3: Eine Grauwertkante, die Grauwertgradienten g_x und g_y und die Richtung des maximalen Gradienten

3 Andere Arbeiten

Die Lokalisierung von Gesichtsteilen ist nur selten alleiniges Thema von Forschungsarbeiten. Meisten steht sie im Zusammenhang mit der Identifikation von Gesichtern (d.h Wiedererkennung von einzelnen Personen). Verfahren zur Gesichtserkennung benötigen die Position von Gesichtsteilen wie Augen und Mund zur Berechnung von geometrischen Merkmalen (z.B. Breite Augenbrauen), zur Wiedererkennung des Gesichtes durch Wiedererkennung von Gesichtsteilen oder zur Normalisierung der Gesichtsbilder (z.B. anhand der Positionen der Augen).

Die Lokalisierung von Gesichtsteilen kann aber auch zur Bewältigung anderer Aufgaben als die Personenidentifikation verwendet werden. Craw et al. (92) berichten von einem Projekt bei dem die Frequenz des Augenblinzelns zur Überprüfung der Aufmerksamkeit eines Kraftfahrzeugfahrers herangezogen werden soll. Um Augenbewegungen von sonstigen Bewegungen unterscheiden zu können, muß die Position der Augen bestimmt werden. Ballard und Stockman (92) extrahieren die Position der beiden Augen und der Nase aus Gesichtsbildern, um die Blickrichtung des Gesichtes zu bestimmen und damit den Mauszeiger eines Computers zu steuern. Zur schnellen Erkennung der Augen suchen sie die Glanzlichter auf der Augenhornhaut.

Generell lassen sich die in der Literatur beschriebenen Verfahren zur Lokalisierung von Gesichtern und Gesichtsteilen in drei Klassen unterscheiden: korrelationsbasierte Verfahren, modellbasierte Verfahren, die "klassische" Bildverarbeitungsmerkmale wie z.B. Kanten verwenden und andere Verfahren, die auf speziellen Merkmalen beruhen.

3.1 Korrelationsbasierte Verfahren

Eine relativ frühe Arbeit zur Lokalisierung und Erkennung von Gesichtern und Gesichtsteilen mit Grauwerttemplates stammt von Baron (81). Er benutzte Templates von markanten Gesichtspartien um individuelle Gesichter wiederzuerkennen. Über diese Experimente hinaus interessiert er sich auch für mögliche Verarbeitungsmechanismen zur Gesichts- und Objekterkennung im Gehirn.

Turk und Pentland (91) haben die Hauptkomponentenanalyse (principal components analysis) für die komprimierte Repräsentation von Gesichtern genutzt. Bei der Hauptkomponentenanalyse werden die Pixel eines Bildes als N-dimensionaler Vektor betrachtet. Die Hauptkomponenten sind die Vektoren, in deren Richtung das Ensemble aller betrachteten Bilder am stärksten schwankt. Sie entsprechen den Eigenvektoren der größten Eigenwerte der $N \times N$ Kovarianzmatrix:

$$C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T$$

(M = Anzahl Bilder, Φ_n = Bild(vektor))

Durch Projektion von Bildern auf die Hauptkomponenten läßt sich ein Bild durch wenige (in Turk und Pentlands Experimenten 40) Koeffizienten repräsentieren und falls nötig, mit geringen Fehlern wieder rekonstruieren. Voraussetzung ist allerdings, daß das Bild aus dem Ensemble, aus dem die Hauptkomponenten berechnet wurden, stammt, oder Bildern des Ensembles ähnlich ist. Diese Eigenschaft läßt sich für ein Detektionsverfahren für Objekte des Ensembles (z.B. Gesichter) nutzen. Man betrachtet dabei den Rekonstruktionsfehler bei der o.g. Datenreduktion an verschiedenen Bildpositionen. Ein geringer Fehler deutet aud das Vorhandensein des gesuchten Objektes hin. Dieses Verfahren läßt sich den templatebasierten Verfahren zuordnen, da die Projektion der verschiedenen Bildausschnitte auf die Hauptkomponentenvektoren Korrelation mit verschiedenen Templates entspricht.

Brunelli und Poggio (93) vergleichen zwei Verfahren zur Identifikation von Gesichtern. Das eine Verfahren benutzt geometrische Maßzahlen eines Gesichtes, das andere beruht auf der Korrelation von Gesichtsausschnitten wie den Augen mit entsprechenden Templates dieser Ausschnitte der zu erkennenden Personen. Beide Verfahren setzen eine Lokalisierung von Gesichtsmerkmalen voraus. Dabei werden für unterschiedliche Gesichtsteile verschiedene Verfahren eingesetzt. Zunächst werden die Augen mit normalisierter Kreuzkorrelation gesucht. Bei der Berechnung des Korrelationswertes einer Maske an einer Bildposition wird dabei für jeden Punkt der Maske in einer dem Maskenpunkt entsprechenden kleinen Umgebung des Bildes der am besten passende Bildpunkt gesucht. Dies ermöglicht ein besseres Auffinden von leicht veränderten Merkmalen. Der Mund und die Nase werden dann mit Hilfe von Projektionen gerichteter Kanten gesucht. Vertikale Kanten werden detektiert, wenn die vertikale Komponente des Gradienten größer als seine horizontale Komponente ist und der Betrag des Gradienten einen Grenzwert überschreitet. Entsprechendes gilt für horizontale Kanten. Projektionen sind die Summen der Intensitäts- oder Gradientenwerte in vertikaler oder horizontaler Richtung. Der Mund wird bei Brunelli und Poggio beispielsweise innerhalb einer durch die gefundene Position der Augen eingegrenzten Umgebung mit Hilfe der horizontalen Projektion der horizontal ausgerichteten Kanten und der horizontalen Projektion des Intensitätswertes gesucht. Ohne das dies ausdrücklich erwähnt wird, scheinen in der von Brunelli und Poggio verwendeten Datenbasis nur Gesichter mit neutralem Gesichtsausdruck und geschlossenem Mund enthalten zu sein.

Bichsel und Pentland (94) nutzen normalisierte Kreuzkorrelation zur Lokalisierung und Verfolgung von Gesichtern. Sie sind dabei besonders an der Analyse der Leistungsfähigkeit des Verfahrens für die Erkennung von im Raum rotierten Objekten (Köpfen) interessiert. Ein implementiertes System verfolgt Köpfe durch Korrelation mit tiefpassgefilterten Templates eines Durchschnittskopfes in verschiedenen Auflösungen und mit verschiedenen Blickrichtungen.

Beymer (93) verwendet normalisierte Kreuzkorrelation für die Detektion von Gesichtsteilen. Ähnlich wie bei Brunelli und Poggion wird für die Berechnung des Korrelationswertes ein modifiziertes Verfahren angewandt, bei dem leichte Abweichungen des Bildes vom Template nicht zur Reduktion des Korrelationswertes führen.

Dabei wird mit einem Algorithmus zur Berechnung des optischen Flusses eine Korrespondenz zwischen Template und Bildpunkten hergestellt. Anschließend wird das Bild entsprechend des berechneten Feldes von Verschiebungsvektoren transformiert und dann der Korrelationswert mit dem Template berechnet.

Vincent et al. (91) nutzen Multilayerperzeptronen und Backpropagation für die Konstruktion von Detektoren für Gesichtsteile. Die Eingabeeinheiten des Netzes sind in einem Fenster angeordnet, das an alle möglichen Positionen des Eingabebildes plaziert wird. Es werden nur zwei hidden units verwendet. Positionen an denen der Ausgabewert der Ausgabeeinheit des Netzes einen Grenzwert überschreitet, werden als Suchbereiche für eine Suche in einer höheren Auflösung markiert, die nach den gleichen Prinzipien wie die grobe Suche abläuft. In der Suche mit höherer Auflösung wird nach räumlich begrenzten Merkmalen wie den Mundwinkeln gesucht. Die Ausgaben der Feinsuche werden in einem weiteren Schritt auf ihre Plausibilität hin untersucht. Dabei werden statistische Daten über die relative Position der gesuchten Merkmale verwendet. Das gesamte Verfahren hat starke Ähnlichkeit mit korrelationsbasierten Verfahren, da sich die Gewichte zwischen Eingabeschicht und hidden units als Templates verstehen lassen, die mit dem Bild an verschiedenen Positionen korreliert werden.

Hutchinson und Welsh (89) vergleichen die Leistungen von einfacher Korrelation und künstlichen neuronalen Netzen bei der Lokalisierung von Augen. Die Aufgabe wurde dadurch erleichtert, daß nur in einem Fenster mit der doppelten Größe des Auges um das Auge herum gesucht wurde. Bei nicht normalisierten Daten schnitten neuronale Netze besser ab. Bei normalisierten Bildern (gleiche mittlere Helligkeit, gleiche Standardabweichung der Helligkeit) brachte dagegen Korrelation mit einem durchschnittlichen Template bessere Ergebnisse. Ein solches Template wird durch Durchschnittsbildung der Grauwerte aus einer Reihe von Bildern erzeugt.

3.2 Modellbasierte Verfahren

Craw et al. (92) haben ein iteratives Verfahren zur Anpassung einer Kopfschablone an die Umrisse des Kopfes im Bild und an die einzelnen Gesichtsmerkmale wie Augen und Mund entwickelt. Zur Erkennung des Umrisses wird die Schablone zufälligen Deformationen unterworfen und immer wieder mit dem Bild verglichen. Die Art der Deformationen ist so eingeschränkt, daß sie immer wieder zu tatsächlich möglichen Kopfformen führen. Die Einschränkungen werden durch statistisch ermittelte Maßzahlen vorgegeben. Die Suche nach der bestmöglichen Übereinstimmung von Template und Bild erfolgt durch simulated annealing (d.h. Deformationen mit zeitlich abnehmender Stärke, Übergang von einem Deformationszustand zu einem neuen mit einer Wahrscheinlichkeit, die von der Differenz der Qualität der Anpassung von Schablone und Bild abhängt). In einer zweiten Phase wird dann mit einer blackboardartigen Kontrollstruktur und lokalen Suchmodulen für bestimmte Merkmale eine genaue Anpassung des Modells an das Bild vorgenommen.

Yuille, Hallinan und Cohen (92) benutzen verformbare Templates zum Lokalisieren und zur genaueren Beschreibung von Augen und Mund. Die verformbaren Templates bestehen aus einfachen Grundelementen (Kreis, Parabelabschnitt), deren genaue Ausprägung von Parametern abhängt. Beispielsweise gibt es im Template für das Auge 11 freie Parameter. Diesen Grundelementen der Templates sind Kräfte oder Potentiale zugeordnet, durch die sie von bestimmten Bildmerkmalen angezogen werden. Als Bildelemente oder -merkmale werden die unverarbeiteten Grauwerte und Felder verwendet, die die Anwesenheit von Kanten, Grauwertminima und Grauwertmaxima anzeigen. Die Anpassung der verformbaren Templates an das Bild erfogt durch iterative Veränderung der Parameter des Templates mit einem Gradientenabstiegsverfahren. Die zu minimierende Energie ergibt sich als gewichtete Summe der Potentiale der Grundelemente des Templates. Die Gewichte sind dabei für verschiedene Phasen des Anpassungsvorganges unterschiedlich.

Cootes et al. (93) haben ein allgemeines Verfahren zur Bestimmung von Lageparametern und internen Parametern (im Gesicht z.B. Abstand Auge - Augenbrauen) von Objekten in Bildern entwickelt. Sie verwenden dabei einen modellbasierten Ansatz. Das Modell kann aus einer Trainingsmenge von Beispielexemplaren gewonnen werden. Es umfaßt Positionsinformationen von Objektpunkten und den typischen Grauwertverlauf in der jeweiligen Umgebung der Punkte. Zur iterativen Berechnung einer neuen Ausprägung des flexiblen Modells wird für jeden modellierten Objektpunkt ein Translationsvektor berechnet, der die geschätzte Position des Punktes in die Richtung anpaßt, in der Grauwertinformation in Bild und Modell möglichst gut übereinstimmen. Anhand dieser Vektoren werden dann die globalen Lage- und Formparameter neu berechnet.

3.3 Andere Merkmale

Reisfeld und Yeshurun (92) berichten über einen Operator, der für einen gegebenen Bildpunkt das Maß an Punktsymmetrie an dieser Stelle berechnet. Bei einer Anwendung dieses Operators auf Bilder von Gesichtern treten hohe Symmetriewerte für die Gesichtsteile Augen und Mund auf.

Das bereits in Kapitel 1 erwähnte Gesichtserkennungs- und verfolgungssystem von Hunke (94) basiert im wesentlichen auf einer Farbklassifikation der Bildpixel. Dazu werden die pro Bildpunkt vorliegenden drei Intensitätswerte (Rotanteil, Grünanteil und Blauanteil) bezüglich der Gesamthelligkeit normiert und damit auf einen zweidimensionalen Farbraum abgebildet. Pixel, deren Farbe typisch für Gesichter ist, werden als mögliche Gesichtspixel, andere Pixel als Hintergrund markiert. Zusätzlich wird zur Klassifizierung der Pixel auch Information über Intensitätsänderungen zwischen zwei aufeinanderfolgenden Bildern (diese Änderungen werden häufig durch Bewegungen der Person verursacht sein) berücksichtigt. Aus der Menge der als evtl. zum Gesicht gehörig klassifizierten Pixel wird alternativ durch Bestimmung des größten zusammenhängenden Bereichs solcher Pixel oder mittels eines neuronalen

Netzes der Bereich des Gesichtes erkannt. Zur Erhöhung der Zuverlässigkeit des Systems wird beim ersten Auffinden des Gesichtes ein relativ breiter auf mehrere Personen passender weiter Bereich von Gesichtsfarben verwendet. Im Verlauf der weiteren erfolgreichen Erkennung des Gesichtes wird dieser Bereich auf die tatsächlich im verfolgten Gesicht vorkommenden Farben eingeengt.

4 Erkennung mit Projektionen von gerichteten Kanten

Zu Beginn der Durchführung dieser Diplomarbeit stellte sich nach Durchsicht der Literatur zum Thema Lokalisation von Gesichtern und Gesichtsteilen (siehe Kapitel 2) heraus, daß nur in einer Arbeit ein Verfahren zur Erkennung des Mundes beschrieben wird (Brunelli und Poggio, 1993). Die Autoren benutzen dabei, wie bereits erwähnt, Projektionen der vertikalen und horizontalen Kanten. Diese Projektionen lassen sich wie weiter unten beschrieben recht einfach und schnell berechnen. Ich entschied mich daher, den Nutzen der Projektionen zur Erkennung des Mundes in einem Versuch zu untersuchen.

Zur Erkennung des Mundes wurden horizontal ausgerichtete Kanten (mit $|g_y| > |g_x|$) ausgewählt, da diese bei geschlossenem oder leicht geöffnetem Mund eindeutig gegenüber anderen Ausrichtungen überwiegen. Ein Bildpunkt wurde als horizontale Kante markiert, wenn die Richtung des maximalen Gradienten innerhalb einer gewissen Schwankungsbreite ($\pm 25°$) um die Vorzugsrichtung lag und der Betrag des Gradienten einen Schwellwert überstieg. Dieser Schwellwert wurde automatisch so bestimmt, daß ein fester Anteil der Punkte (z.B. 5%) als horizontale Kanten markiert wurden. Die automatische Bestimmung des Schwellwertes macht die Kantendetektion unempfindlich gegenüber Variationen im Kontrast der zu verarbeitenden Bilder.

Abbildung 4: Grauwertbild

Für die Untersuchung des Wertes von Projektionen gerichteter Kanten für die Erkennung wurden Bilder einer Person mit unterschiedlichem Öffnungsgrad des Mundes vor einem neutral weißem Hintergrund aufgenommen. Dadurch wurden

Beeinflußungen durch Hintergrundkanten ausgeschlossen. Anschließend wurden zu diesen Bildern die horizontalen Projektionen (also Summen in horizontaler Richtung) der horizontal ausgerichteten Kanten berechnet. Es zeigte sich, daß es im oberen Teil des Gesichtes stets zwei Maxima gab, die mit der Spitze des Kopfes und der Höhe der Augen zusammenfiel. Im unteren Teil des Bildes wurde das nächstkleinere Maximum fast immer durch die Kanten des Mundes erzeugt (siehe Bild 6). Aufgrund dieser Beobachtung ließ sich ein erstes primitives Programm zur Detektion des Mundes konstruieren, das einfach nach diesem Maximum im Bereich unter den beiden stärksten Maxima der horizontalen Projektion der horizontalen Kanten im Bild suchte. Nach der Bestimmung der vertikalen Position des Mundes ließ sich die horizontale Position des Mundes einfach durch eine vertikale Projektion der horizontalen Kanten im Bereich der ausgewählten vertikalen Position bestimmen.

Dieses einfache Programm war für die gegebenen einfachen Bildfolgen relativ robust. Es hatte Probleme bei weit offenem Mund und bei zusammengepressten Lippen. Im ersten Fall tauchen horizontale Kanten nur in der Mitte von Ober- bzw. Unterlippe auf. Dadurch erscheinen in der horizontalen Projektion der Kanten zwei mäßig hohe Maxima mit einem recht deutlichen vertikalen Abstand (siehe Abbildung 7). während bei nur leicht geöffnetem Mund die Kanten beider Lippen einen zusammenhängenden "Peak" bilden. Im zweiten Fall tauchen nur relativ schwache horizontale Kanten an den Lippen auf. Die Projektionen der Nasenspitze oder des Kinnes ergeben dann möglicherweise höhere Werte als die des Mundes. Obwohl die einfache Maximumbestimmung in diesen Fällen versagte, wäre das Verfahren durch genauere Analyse der Projektionen und die Einführung von Fallunterscheidungen (z.B. zwei kleine Peaks oder ein großer) sicherlich noch verbesserungsfähig gewesen. Ein anderes Problem führte aber dazu, daß dieses einfache Verfahren nicht weiter verfolgt wurde.

Die Forderung nach einem neutralen Hintergrund erschien als zu drastisch. Falls diese Forderung jedoch nicht eingehalten wird, verlieren die Projektionen der Bildkanten rapide an Informationsgehalt. Das Auftauchen eines Hintergrundobjektes mit starken horizontalen Kanten in der unteren Bildhälfte machte die Erkennung mit dem o.g. einfachen Verfahren bereits unmöglich, da es aufgrund seiner stärkeren Projektion vom Verfahren als Mund erkannt wurde.

Trotz dieser Unzulänglichkeiten wurden bei den durchgeführten Experimenten jedoch zwei für die weitere Arbeit wesentliche Erkenntnisse gewonnen:

1. Horizontale Kanten sind prinzipiell geeignet, um den Mund zu identifzieren.
2. Um eine robuste Erkennung des Mundes zu gewährleisten, sollte zunächst eine grobe Schätzung der Position des Gesichts vorliegen. Wenn diese Schätzung so gut ist, daß der Suchbereich auf das Gesicht der Person eingeschränkt ist, ist das Suchproblem wesentlich einfacher. Der Mund muß jetzt nur von bestimmten anderen Gesichtsteilen (Nase, Kinn) unterschieden werden und nicht von irgendwelchen unvorhersehbaren Hintergrundobjekten.

Abbildung 5: Vertikale Kantenelemente im Grauwertbild

Abbildung 6: Horizontale Kanten mit horizontaler Projektion (rechts)

Abbildung 7: Horizontale Projektion bei offenem Mund

5 Lippenerkennung mit gelernter Translationsinvarianz

5.1 Motivation für den Einsatz neuronaler Netze

Aufgrund der im vorigen Kapitel beschriebenen Schwierigkeiten bei der manuellen Konstruktion eines Detektors für des Mund wurde der Einsatz von Lernverfahren bei der Konstruktion eines solchen Detektors in Erwägung gezogen. Für den Einsatz eines Lernverfahrens spricht auch die ebenfalls im vorigen Kapitel begründete Wichtigkeit der Umgebung des Mundes für dessen Erkennung. Es erscheint schwierig, diese Zusatzinformation systematisch in einem konventionellen Objekterkennungsverfahren zu berücksichtigen. In einer in Kapitel 2 kurz angesprochenen Arbeit (Craw et al. 92) wurde dazu eine blackboardartige Kontrollarchitektur eingesetzt. Dieser Ansatz erscheint recht kompliziert zu sein. Außerdem wurde dort in einem wichtigen Erkennungsschritt (Erkennung des Umrißes des Kopfes) das sehr rechenaufwendige Verfahren des "simulated annealing" eingesetzt. Da der Lippenfinder auf Bildfolgen angewandt werden soll, sind Verfahren, die mehrere Minuten zur Analyse eines Bildes benötigen, sehr unpraktisch. Craw et al. geben zwar keine konkreten Rechenzeiten an, simulated annealing ist aber für seine langsame Konvergenz bekannt.

Eine andere naheliegende Vorgehensweise zur Integration von globaleren Informationen in die Lippensuche ist eine Suche nach einem Gesicht mittels Korrelation, wie sie z.B. Bichsel (94) durchführt. Durch die Verwendung von Templates verschiedener Grössen konnte Bichsel die dafür erforderliche Rechenzeit auf eine Sekunde senken. Dies wäre für eine Lippenfindung in einer abgespeicherten Bildfolge wohl tragbar. Allerdings funktioniert das beschriebene System nur für sechs bestimmte Personen. Diese Einschränkung wird sicherlich durch Unterschiede im Aussehen der Menschen begründet sein. Faktoren, die zu starken Differenzen im Grauwertverlauf der Gesichtsregion führen, könnten beispielsweise Unterschiede in Haarfarbe, Frisur und Kopfform sein. Bei einem Verfahren, das auf einem Vergleich von Grauwerten beruht, ist außerdem nicht klar, wie es auf Änderungen in der relativen Helligkeit des Hintergrundes (z.B. heller als Gesicht oder dunkler als Gesicht) reagiert. Das in diesem Kapitel beschriebene auf einem künstlichen neuronalen Netz beruhende Verfahren zur groben Erkennung ist im Vergleich zu den beiden oben genannten Alternativen schnell und funktioniert für sehr unterschiedlich aussehende Personen und Bildhintergründe.

Der Einsatz eines mit Backpropagation trainierten neuronalen Netzes für die gestellte Aufgabe wurde auch durch den Erfolg eines solchen Ansatzes bei der autonomen Steuerung eines Kraftfahrzeuges (Pomerleau, 92) motiviert. Mit einer einfachen Netzarchitektur (30×32 Eingabeeinheiten, 4 versteckten Einheiten und 30 Ausgabeeinheiten) ist das von Pomerleau entwickelte System ALVINN (Autonomous Land Vehicle in a Neural Network) in der Lage, ein Fahrzeug über Straßen unterschiedlicher Qualität zu steuern. Es erreicht dabei an die Leistungsfähigkeit eines

wesentlich komplexeren mit Spezialhardware arbeitenden Systems (Dickmann und Gräfe, 88) heran. Beim Entwurf des in diesem Kapitel beschriebenen Systems gab das ALVINN-System wichtige Anregungen, besonders bei der Wahl der Ausgaberepräsentation und der Verwendung künstlich erzeugter Trainingsdaten.

5.2 Die implementierte Netzarchitektur

Die in diesem Abschitt beschriebene Netzarchitektur ist diejenige, die im Vergleich zu vielen anderen getesteten Möglichkeiten die besten Ergebnisse liefert. In den folgenden Abschnitten wird der Einfluß von einigen möglichen Änderungen (z.B. in der Vorverarbeitung oder Ausgaberepräsentation) auf die Leistungsfähigkeit des Systems diskutiert.

Die Suche nach den Mundwinkeln einer im Bild zu sehenden Person wird in zwei Stufen jeweils durch künstliche neuronale Netze durchgeführt. Das erste Netz liefert eine grobe Schätzung der Position des Mittelpunktes des Mundes. Das zweite Netz sucht in einem rechteckingen Fenster um diese Position nach den Mundwinkeln. Wie im ALVINN-System wird auch hier jeweils eine dreischichtige Architektur mit einer versteckten Schicht verwendet.

Groberkennung

Die Größe des zu verarbeitenden Bildes wird zunächst von 256×256 Pixel auf 34×34 Pixel reduziert. Mit dem Sobeloperator werden dann die Grauwertgradienten in x- und y-Richtung bestimmt. Aufgrund der Maskengröße des Sobeloperators können für die Zeilen und Spalten am Bildrand keine Gradientenwerte errechnet werden. Die Gradientenbilder haben daher nur die Größe 32×32 Pixel. Wie in Kapitel 3 beschrieben, werden nun ein vertikales und ein horizontales Kantenbild extrahiert. Diese Kantenbilder haben von 0 verschiedene Werte an den Bildpositionen, an denen die Richtung des detektierten Kantenelementes höchstens um $\pm 25°$ von der horizontalen oder vertikalen Vorzugsrichtung abweicht. Für beide Kantenbilder wird unabhängig voneinander ein Amplitudengrenzwert so bestimmt, daß die Anzahl der Bildpunkte mit größerer Amplitude und korrekter Richtung 10% der Bildpixel ausmacht. Bildpunkten mit korrekter Richtung, die diesen Grenzwert übertreffen, wird der Eingabewert 1 anderen Bildpunkten mit korrekter Richtung und Amplitude A der Wert $\frac{A}{Grenzamplitude}$ zugeordnet.

Die beiden gerichteten Kantenbilder der Größe 32×32 Pixel bilden die Eingabeschicht des Netzes. Um den Rechenaufwand für Training und Anwendung des Netzes zu begrenzen, ist nicht jede Eingabeeinheit mit jeder versteckten Einheit verbunden. Vielmehr sind die Eingabeeinheiten in $4 \times 4 = 16$ Felder mit jeweils $8 \times 8 \times 2 = 128$ Eingabeeinheiten eingeteilt. Jedem Eingabefeld sind 8 versteckte Einheiten zugeordnet, die vollständig mit allen Eingabeeinheiten des Feldes verbunden

sind. Insgesamt existieren also $16 \times 8 = 128$ versteckte Einheiten. Die Ausgabeeinheiten sind in einem Gitter angeordnet, das den Bereich der möglichen Positionen des Mundes überdeckt. Es wird vorausgesetzt, daß sich das zu lokalisierende Gesicht vollständig im Bild befindet und der Abstand von Stirn und Kinn zwischen 45% und 80% der Bildhöhe ausmacht. Dies schränkt den Bereich der möglichen Positionen des Mundes erheblich ein.

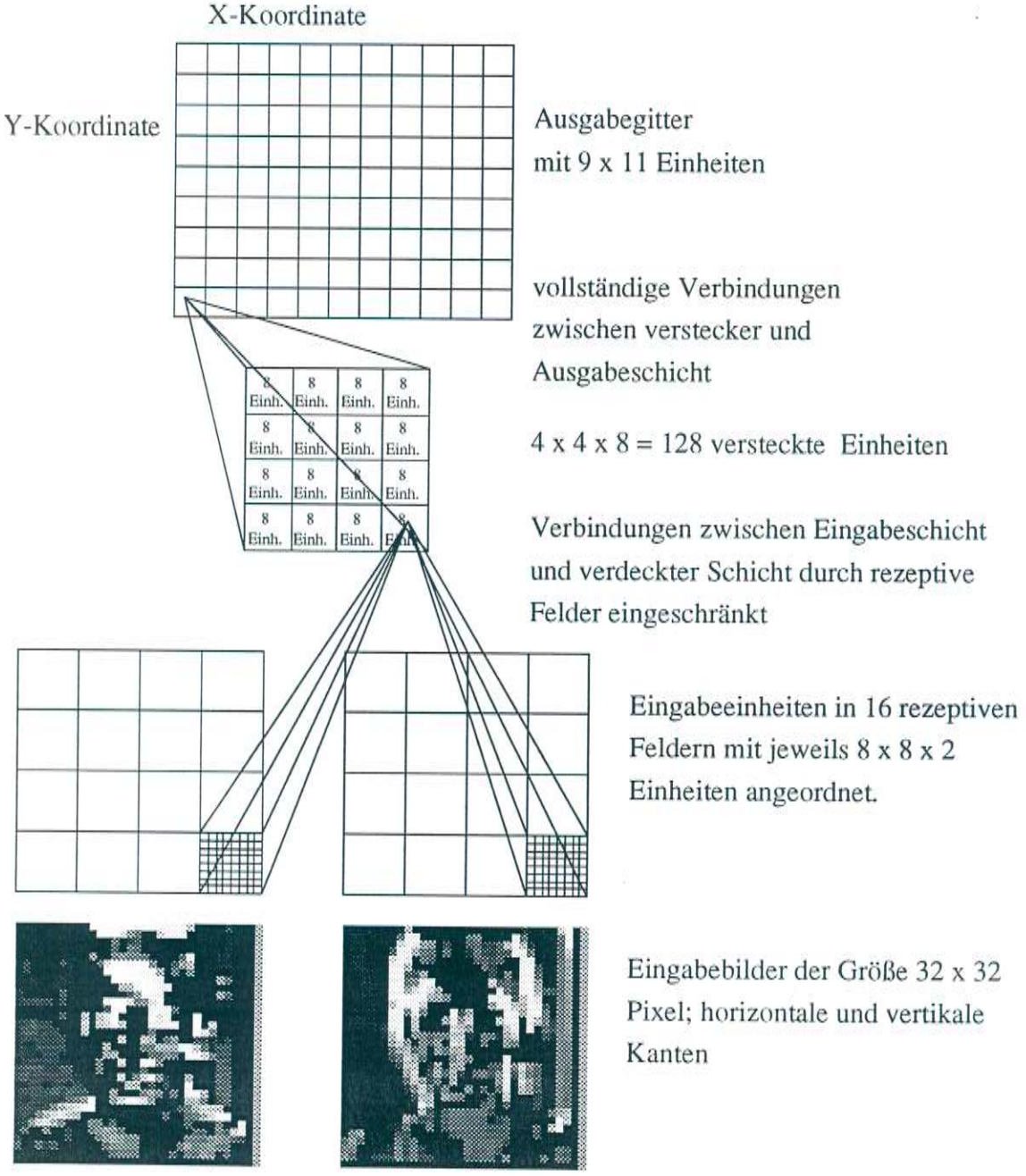

Abbildung 8: Netzarchitektur für die Groberkennung

Die Anzahl der Verbindungen zwischen verdeckter Schicht und Ausgangsschicht

und damit auch der Rechenaufwand wird dadurch weiter reduziert, daß nur jeder zweiten Zeile und Spalte eine Ausgabeeinheit zugeordnet ist. Insgesamt reichen daher $9 \times 11 = 99$ Ausgabeeinheiten aus. Dabei ist jede Ausgabeeinheit mit allen versteckten Einheiten verbunden. Abbildung 8 gibt einen Überblick über die Architektur des Netzes.

Die Ausgabeeinheiten werden so trainiert, daß sie einen hohen Ausgabewert erreichen, wenn der Mund sich nahe der Position im Bild befindet, die sie repräsentieren und andernfalls die Aktivierung 0 annehmen. Bei der Verwendung des Netzes wird dann diejenige Bildposition als vermutliche Position des Mundes angenommen, der die Ausgabeeinheit mit der höchsten Aktivierung zugeordnet ist.

Die Aktivierungen der versteckten Einheiten und der Ausgabeeinheiten werden in diesem Netz und im Netz zur Feinerkennung mit der Aktivierungsfunktion $\frac{1}{1+e^{-h}}$ (h = gewichtete Summe der Eingänge) berechnet (siehe z.B. Hertz et al., 91). Diese Aktivierungsfunktion nimmt Werte zwischen 0 und 1 an.

Einzelheiten des Trainings

In den Bildern, auf deren Basis das Netz trainiert wird, werden die linke obere Ecke des Gesichtes (in Stirnhöhe), die rechte untere Ecke (in Höhe des Kinnes) und die Mitte des Mundes markiert. Aus jedem gegebenen Bild werden 8 Hilfsbilder gewonnen, in denen die Höhe des Gesichtes zwischen 45% und 80% der Bildhöhe von 32 Pixeln ausmacht. Diese Hilfsbilder ermöglichen eine effizientere Erzeugung von Trainingsbildern, die das Gesicht in verschiedenen Größen an verschiedenen Bildpositionen zeigen. Sie erlauben es, rechenintensive Vorverarbeitungsschritte schon vor Beginn des Trainings vorzunehmen, ohne daß alle präsentierten Trainingsbilder einzeln abgespeichert werden müßten. Das Gesicht befindet sich in der Mitte der Hilfsbilder, die dem o.g. Vorverarbeitungsschritt der Extraktion von zwei gerichteten Kantenbildern unterworfen werden. Obwohl also zwei Bilder vorliegen, spreche ich zur Vereinfachung weiter von einem Bild. Um das 32×32 Bild des Gesichtes und seiner näheren Umgebung herum werden die Hilfsbilder so erweitert, daß ein 32×32 großes Fenster, das auf das Hilfsbild gelegt wird, gerade noch in jedem Fall das Gesicht vollständig enthält. Diese Bedingung führt dazu, daß Hilfsbilder von kleineren Gesichtern größer als die von großen Gesichtern sind. Bei diesen Erweiterungen der Hilfsbilder wird so weit wie möglich das Ausgangsbild verwendet. Häufig geht aber der Bereich des Hilfsbildes über die Grenzen des Ausgangsbildes hinaus. Entsprechende Positionen der Hilfsbilder werden auf den Wert 0 (keine Kante) gesetzt. Die Abbildungen 9 und 10 zeigen ein Grauwertbild einer Person und zwei daraus erzeugte Hilfsbilder. Bei diesen Hilfsbildern bestand die Vorverarbeitung in einer Kantendektektion ohne Berücksichtigung der Kantenrichtung. Hilfsbilder zu anderen Vorverarbeitungen sind im Kapitel 6 in den Abbildungen 12 bis 14 zu sehen.

Abbildung 9: Ausgangsbild

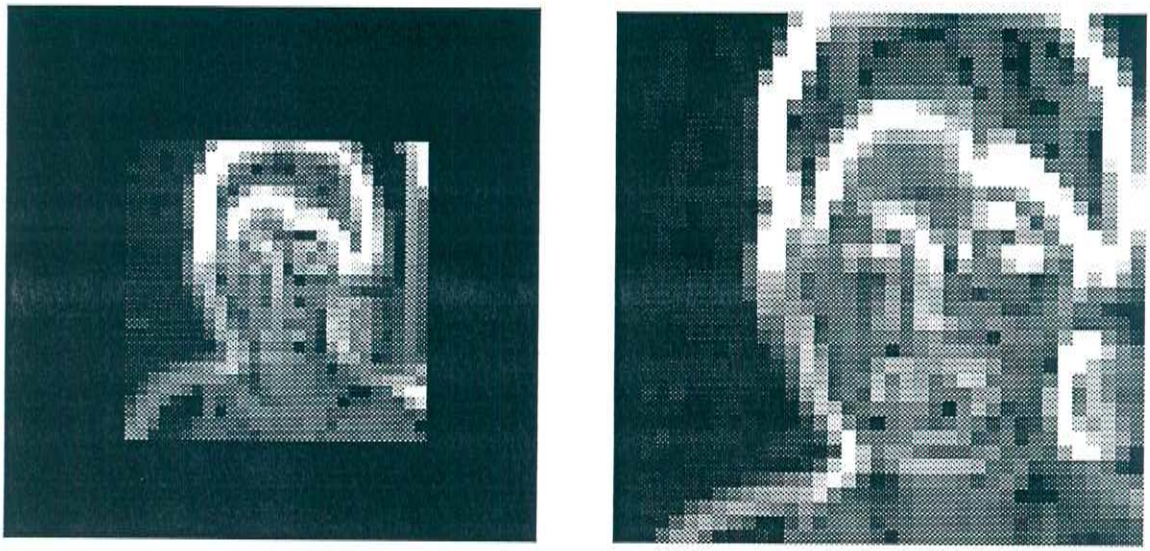

Abbildung 10: Hilfsbild zur Groberkennung: ungerichtete Kanten, maximale und minimale Verkleinerung

Beim Training wird ein 32×32 Pixel großes Fenster über alle möglichen Positionen eines gegebenen Hilfsbildes geschoben und die in diesem Fenster enthaltenen Bildwerte als Eingabe des Netzes verwendet. Der Aufwand zur Erzeugung der Trainingsdaten ist daher zu dieser Zeit auf ein reines Umkopieren beschränkt. Alle Hilfsbilder

werden der Reihe nach so lange benutzt, bis die vorgegebene Zahl von Trainingsschritten erreicht ist. Da aus einem Hilfsbild mehr als hundert Trainingsbilder erzeugt werden können und ein zu starker Einfluß der zu Beginn des Trainings präsentierten Bilder verhindert werden soll, wird eine gegebene Position eines Hilfsbildes nur mit einer gewissen Wahrscheinlichkeit zum Training benutzt.

Bei der Vorgabe der Aktivierungen der Ausgabeeinheiten hat es sich als vorteilhaft erwiesen, nicht nur der am besten passenden Ausgabeeinheit eine von 0 verschiedene Aktivierung vorzugeben (Pomerleau 92). Einer Ausgabeeinheit, die genau auf die Position des Mundes fällt, wird eine Aktivierung von 1 vorgegeben. Für andere Ausgabeeinheiten fällt dieser Wert normalverteilt entsprechend ihrer Distanz zum Mund ab. In den Experimenten wurde eine Normalverteilung mit der Standardabweichung 2,5 verwendet. Für das Training wurde einfaches Backpropagation (Rumelhart et al. 86) mit der Lernrate 0,05 und ohne Momentum angewandt. Nach ca. 40.000 Trainingsschritten verbesserten sich die gelernten Netze nur noch wenig. In dieser Arbeit bedeuten die Begriffe "Trainingsschritt" oder "Präsentation eines Bildes" einen Vorwärtspaß durch das Perzeptron mit anschließendem Backpropagation des an den Ausgabeeinheiten aufgetretenen Fehlers.

Feinerkennung

Das Netz zur Feinerkennung der Lippenwinkel wurde in Anlehnung an das erfolgreiche Netz zur Groberkennung des Gesichtes entworfen. Bei der Feinerkennung wird ein rechteckiges Fenster um die vom ersten Netz grob geschätzte Mundposition ausgeschnitten. Da auch zur Feinerkennung eine Verarbeitung in der Auflösung von 256×256 Pixeln (mittlere Mundbreite 42 Pixel) ein zu großes Netz erfordern würde, wird die Größe des Bildes jedoch zuerst um den Faktor von 1,875 reduziert. Der dann ausgeschnittene Bereich umfaßt nach Anwendung des Sobel-Operators 36 Zeilen und 60 Spalten in niedriger Auflösung. Bei der Feinerkennung werden nur horizontal ausgerichtete Kanten berücksichtigt. Der Bereich der akzeptierten Richtungen liegt bei $\pm 45°$ um die Vorzugsrichtung. Das Verfahren zur Normalisierung unterscheidet sich von dem bei der Groberkennung angewandten nur im Anteil der Eingabewerte mit der Größe 1.0, der hier bei 8% liegt. Die Eingabeeinheiten sind in $6 \times 6 = 36$ Eingabefelder der Größe 6×10 aufgeteilt. Jedem Eingabefeld sind 10 versteckte Einheiten zugeordnet.

Das Netz wird so trainiert, daß es einen Mund in jeder Position erkennen kann, in der beide Mundwinkel im Bild sichtbar sind. Anders als bei der Groberkennung kann daher der Bereich der sinnvollen Ausgabepositionen kaum eingeschränkt werden. Nur ein Auftreten des linken Mundwinkels am rechten Bildrand oder des rechten Mundwinkels am linken Bildrand ist unter dieser Bedingung unmöglich. Die Ausgabeeinheiten sind wiederum auf jede zweite Zeile und Spalte verteilt. Insgesamt sind die Ausgabeeinheiten für jeden Mundwinkel in 19 Zeilen und 23 Spalten angeordnet. Es gibt also insgesamt 874 Ausgabeeinheiten. Wie bei der Groberkennung sind alle versteckten Einheiten mit allen Ausgabeeinheiten verbunden. Als

geschätzte Position der Mundwinkel werden wiederum die Bildpositionen angenommen, die den Ausgabeeinheiten des linken oder des rechten Mundwinkels mit den höchsten Aktivierungen entsprechen.

Auch das Training verläuft ähnlich wie bei der Groberkennung. Es werden 7 Hilfsbilder für jedes vorgegebene Bild erzeugt. Die Normalisierung der Größe orientiert sich am Abstand der Augen der Person. Der Augenabstand wird dabei zwischen 21 und 32 Pixeln variert. Analog zur Groberkennung befindet sich der Mund in der Mitte der Hilfsbilder. Um Einflüsse von starken horizontalen Kanten des an der linken und rechten Seite des Hilfsbildes sichtbaren Hintergrundes auf die Normalisierung auszuschließen wird nur der mittlere Bereich des Hilfsbildes bei der Errechnung des schon häufiger erwähnten für die Normalisierung wichtigen Grenzwertes herangezogen. Aus den vor Beginn des Trainings erzeuten Hilfsbildern lassen sich dann wieder einfach durch Umkopieren Trainingsbilder erzeugen. Auch bei der Feinerkennung wird zur Auswahl der zum Training benutzten Bilder ein Zufallsgenerator eingesetzt. Die Berechnung der Aktivierungen der Ausgabeeinheiten erfolgt mit der gleichen Normalverteilung wie bei der Groberkennung. In der Abbildung 11 sind zwei Hilfsbilder zur Feinerkennung zu sehen, die aus Bild 9 erzeugt wurden. Hilfsbilder, bei denen andere Vorverarbeitungen verwendet wurden, sind in Kapitel 6 in den Abbildungen 16 bis 20 zu sehen.

Abbildung 11: Hilfsbild zur Feinerkennung: ungerichtete Kanten, maximale und minimale Verkleinerung

6 Ergebnisse zur Netzarchitektur mit gelernter Translationsinvarianz

In diesem Kapitel werden Experimente beschrieben, in denen die Leistungsfähigkeit der im vorigen Kapitel beschriebenen Netzarchitektur getestet und verschiedene Variationen des Erkenners ausprobiert wurden. Zunächst wird in Abschnitt 6.1. beschrieben, wie die in den Experimenten verwendeten Bilder aufgenommen wurden. In Abschnitt 6.2. werden verschiedene Vorverarbeiungen miteinander verglichen. In Abschnitt 6.3. werden Variationen der internen Netzstruktur, insbesondere der Anzahl der versteckten Einheiten untersucht. Der Einfluß von unterschiedlichen Ausgaberepräsentationen auf die Erkennungsgenauigkeit wird in Abschnitt 6.4. analysiert. In Abschnitt 6.5. wird gezeigt, daß eine Interpolation der Netzausgaben auf eine größere Genauigkeit als durch den Abstand der Ausgabeeinheiten vorgegeben ist, keine Verbesserung der Erkennungsgenauigkeit bringt. Der Einfluß von Änderungen der Gesamtgröße der Netze auf die Erkennungsgenauigkeit wird in Abschnitt 6.6. diskutiert. Während in den vorhergehenden Abschnitten Grob- und Feinerkennung getrennt getestet wurden, wird in Abschnitt 6.7. die Leistung des Gesamtsystems beschrieben. In Abschnitt 6.8. wird schließlich über Experimente berichtet, bei denen nur auf Bilder einer Person trainiert und getestet wurde (personenabhängige Erkennung).

6.1 Aufnahme der verwendeten Bilder

Bei der Aufnahme der Bilder wurde das von Martin Hunke (94) entwickelte Facetrackingsystem benutzt. Diese Programm verwendet nur Bilder der Größe 100×150 Pixel. Der verwendete Framegrabber wird dementsprechend eingestellt und liefert dann bis zur nächsten Initialisierung nur noch Bilder dieser Auflösung, die aber für die weitere Verarbeitung (Lippenerkennung, Lippenlesen) zu gering ist. Daher wurden die Bilder mit einem zweiten Framegrabber auf einem anderen Rechner aufgenommen. Dies führt zu dem Problem, daß von Facetracker und Bildaufnahmeprogramm unterschiedliche Bilder (Frames) genutzt werden.

Die zur Aufnahme ihres Gesichtes bereiten Personen wurden gebeten, sich während der Aufnahme im Raum hin und her zu bewegen. Dieser Aufforderung folgten auch einige der Personen. Die schlechte Synchronisation zwischen Facetracker und Aufnahmeprogramm, die wohl auch durch das benutzte relativ langsame Aufnahmeprogramm bedingt war, führte bei schnellen Bewegungen zu erheblichen Differenzen zwischen der realen Position des Gesichtes im aufgenommenen Bild und der vom Facetracker gelieferten Position. Die Folgen dieser Differenzen wurden dadurch vergrößert, daß vom Aufnahmeprogramm nur ein Bildausschnitt um die vom Facetracker gelieferte Position des Mittelpunktes des Gesichtes verwendet wurde. Die Größe dieses Ausschnittes richtet sich nach der Länge der kleineren Seite des virtuellen Suchfensters.

Mit dem beschriebenen System wurden 20 Sequenzen von verschiedenen Personen mit insgesamt 590 Bildern aufgenommen. Davon war in 73 Bildern der Kopf nicht zu sehen oder erheblich durch den Bildrand abgeschnitten. Da daher nicht alle Bilder weiterverwendet werden konnten, wurden nur diejenigen Bilder ausgewählt, in denen die vom Facetracker geschätzte Gesichtsposition in etwa mit der tatsächlichen Position übereinstimmt. Es wurden 424 Bilder weiter verwendet. Diese Auswahl (aber auch die Verwendung des Facetrackers generell) führen dazu, daß sich die Gesichter zumindest horizontal fast ausschließlich in der Bildmitte befinden. Die vertikale Position des Gesichtes weicht in den aufgenommenen Bildern stärker von der Bildmitte ab, da in der verwendeten Version des Facetrackers der sichtbare Teil des Halses fast immer dem Gesichtsbereich zugeordnet wird. Rückblickend wäre es wohl sinnvoller gewesen, die Auswahl der unbrauchbaren Bilder nach anderen Kriterien (z.B. Mindestabstand von Mund und Augen zum Bildrand) zu treffen. Insgesamt ist festzustellen, daß die folgenden Testergebnisse für Bilder, in denen sich ein Gesicht nahe eines Bildrandes befindet, bei der Groberkennung nicht unbedingt repräsentativ sind. Um dieses Manko auszugleichen, sind in Abschnitt 6.7 die Ergebnisse eines Tests der Groberkennung mit 93 Bildern zu sehen, die das Gesicht relativ vollständig enthalten, aber bei den anderen Experimenten nicht verwendet wurden. In Abbildung 28 ist ein Bild aus jeder Sequenz zu sehen.

6.2 Vorverarbeitung

Groberkennung

Der Groberkennung auf der Grundlage von Grauwertkanten liegt die Vermutung zugrunde, daß der Umriß des Kopfes ein wichtiges Erkennungsmerkmal bei der groben Lokalisierung des Mundes ist. Dieses Merkmal ist schon bei niedriger Auflösung mit hoher Robustheit gegenüber dem Aussehen des Hintergrundes erkennbar. Bei ersten Experimenten mit Grauwertkanten wurde nur die Amplitude des Grauwertgradienten (in normalisierter Form) als Eingabe in ein neuronales Netz verwendet. Weitere Versuche zeigten jedoch, daß die Aufteilung in zwei Kantenbilder mit horizontalen und vertikalen Kanten die Erkennungssicherheit steigert. Zum Training wurden jeweils 3 Bilder von 10 Personen ausgewählt und in verschiedenen Größen und mit verschiedenen Translationen dem Netz präsentiert. Zum Test wurden 212 Bilder von 10 anderen Personen verwendet. In Abbildung 15 wird der Erkennungsfehler für die Eingabe von normalisierten Grauwerten (siehe Feinerkennung), ungerichteten Sobelkanten und zwei gerichteten Kantenbildern verglichen. Die Ergebnisse mit zwei Kantenfeldern erscheinen im Hinblick auf die Auflösung der verwendeten Bilder als bemerkenswert gut (ein Eingabepixel entspricht 8 realen Bildpixeln und die Positionen der Ausgabeeinheiten sind sogar jeweils 16 Pixel voneinander entfernt).

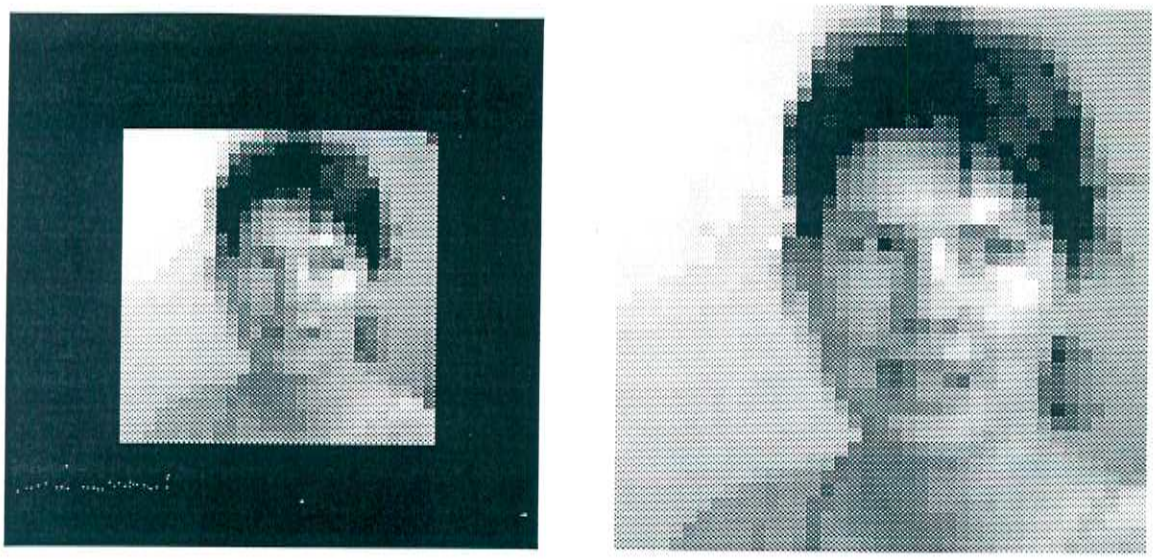

Abbildung 12: Hilfsbild zur Groberkennung: normalisierte Grauwerte, minimale und maximale Verkleinerung

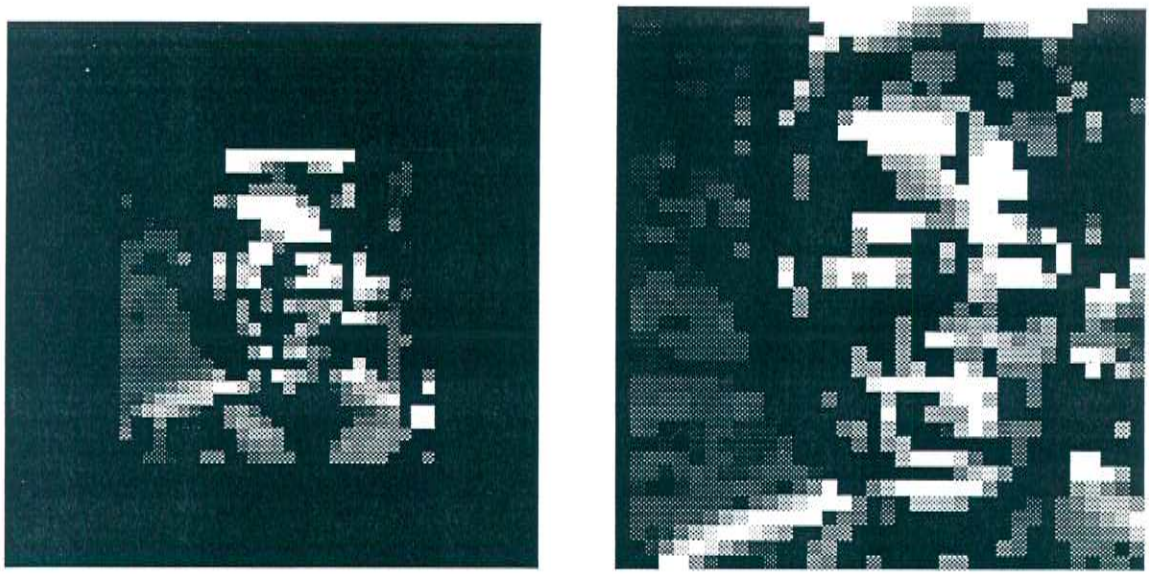

Abbildung 13: Hilfsbild zur Groberkennung: horizontale Kanten, maximale und minimale Verkleinerung

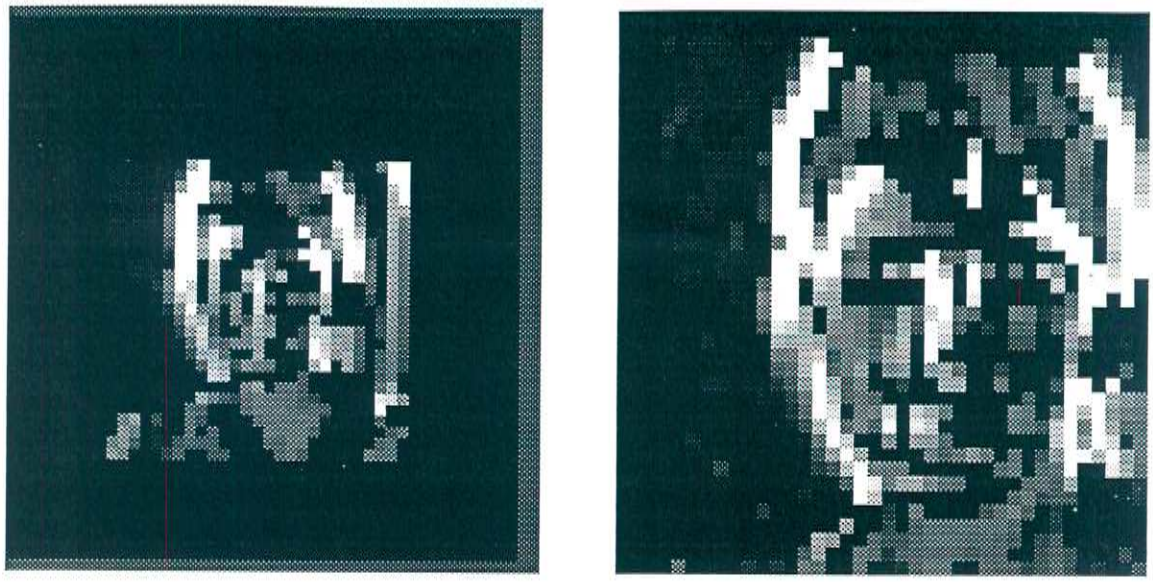

Abbildung 14: Hilfsbild zur Groberkennung: vertikale Kanten, maximale und minimale Verkleinerung

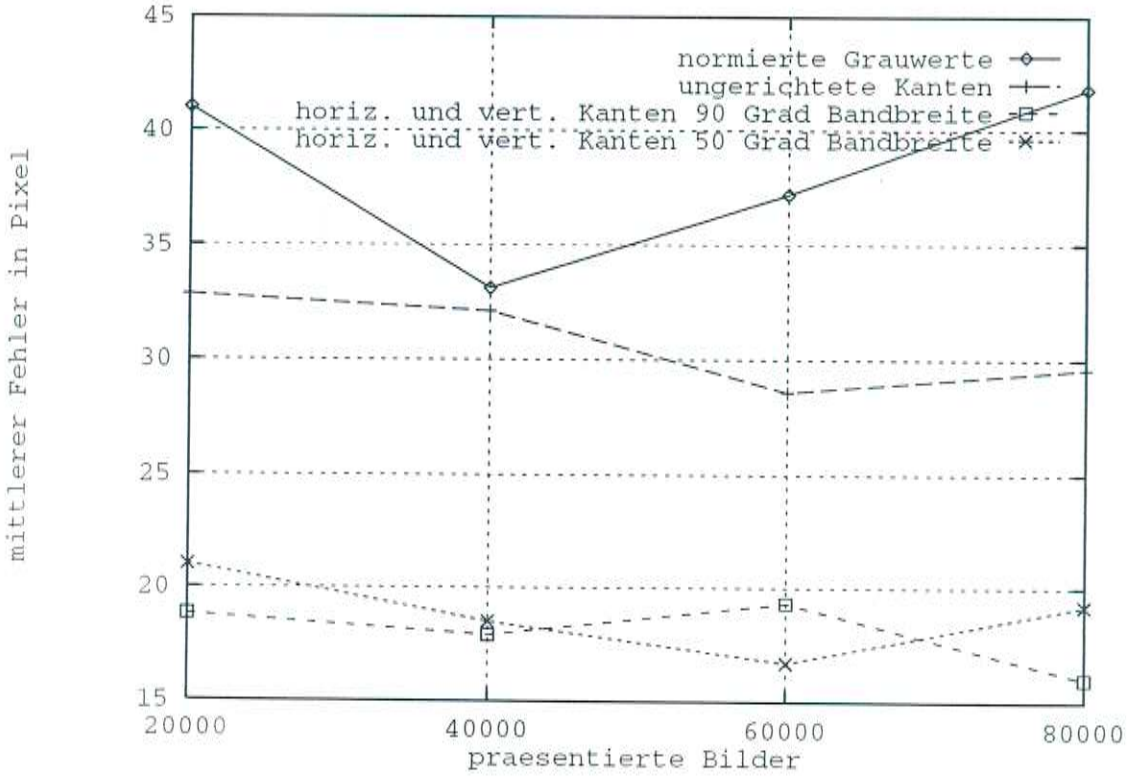

Abbildung 15: Abhängigkeit der Erkennungsleistung von der Vorverarbeitung

Feinerkennung

Hier wurden eine größere Anzahl von Methoden miteinander verglichen. Da das weiter oben beschriebene Netz sehr groß ist und ein Trainingslauf auf einer Alpha-Workstation (DEC AXP 600) ca. 2 Tage dauert, erfolgte dieser Vergleich mit Netzen geringerer Größe. Das Eingabegitter ist hier nur 24×40 Pixel groß. Es gibt $4 \times 4 = 16$ Eingabefelder, die Zahl der versteckten Einheiten pro Eingabefeld ist wie beim größeren Netz 10 und für jeden Mundwinkel existieren 13×14 Ausgabeeinheiten.

Getestete Eingabefelder waren:

- normalisierter Grauwert
- ungerichtete Sobelkanten
- horizontale Kantenfelder mit verschiedenen Richtungstoleranzen
- eine Kombination von horizontalen und vertikalen Kanten
- eine Kombination von ungerichteten Kanten und einem Maß der Rötlichkeit der Pixel
- das Maß der Rötlichkeit allein

Das Grauwertbild wurde normalisiert, indem ein minimaler und ein maximaler Grenzwert so bestimmt wurden, daß 5% der Bildpixel heller als der maximale Grauwert und 5% dunkler als der minimale Grauwert waren. Die diesen Pixeln entsprechenden Eingabeeinheiten erhielten die Aktivierungen 1 bzw. 0. Alle anderen Helligkeitswerte wurden linear zwischen 1 und 0 abgebildet.

Zur Berechnung eines Maßes der Rötlichkeit eines Bildpixels wurde der Intensitätswert des Rotanteils durch die Gesamthelligkeit des des Bildpixels geteilt:

$$Rotanteil = \frac{R}{R + G + B}$$

Diese Rötlichkeitswerte wurden dann wie oben beschrieben so normalisiert, daß 5% der Bildpixel den Wert 0.0 und 5% den Wert 1.0 erhalten. Die jeweils am stärksten und am schwächsten verkleinerten Hilfsbilder (siehe dazu Kapitel 4) zu jeder Vorverarbeitung sind in den Abbildungen 16 bis 19 zu sehen. Sie sind aus dem Farbbild erzeugt worden, daß in Abbildung 9 als Grauwertbild zu sehen ist.

Abbildung 16: Hilfsbild zur Feinerkennung: normalisierte Grauwerte, maximale und minimale Verkleinerung

Abbildung 17: Hilfsbild zur Feinerkennung: vertikale Kanten, maximale und minimale Verkleinerung

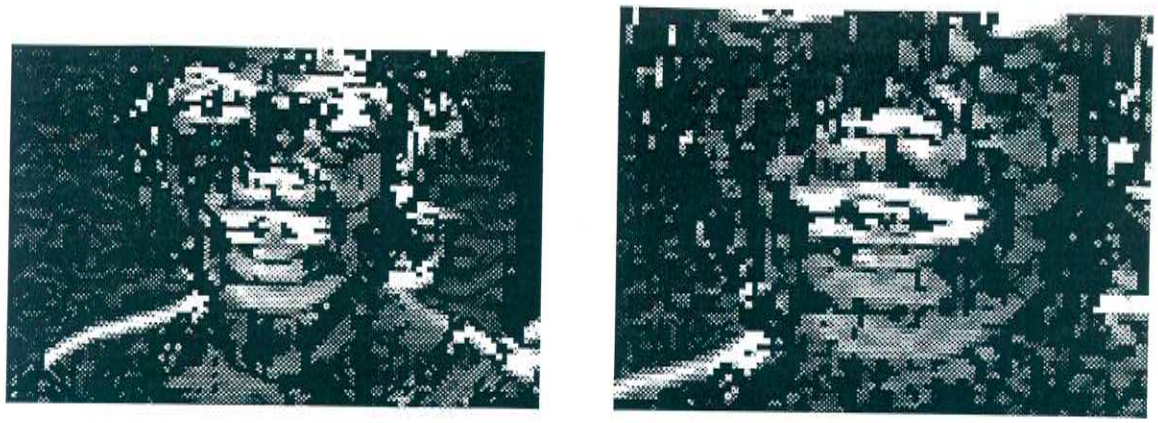

Abbildung 18: Hilfsbild zur Feinerkennung: horizontale Kanten, maximale und minimale Verkleinerung

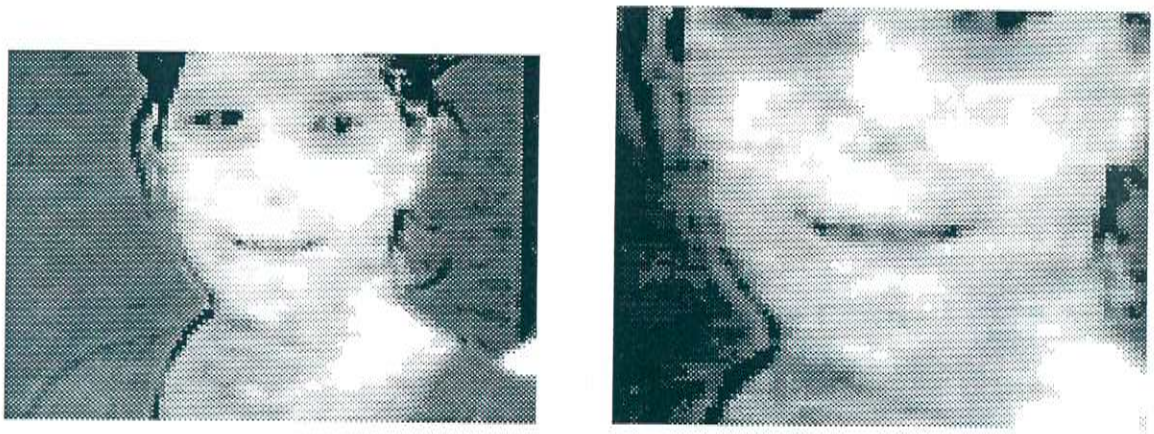

Abbildung 19: Hilfsbild zur Feinerkennung: normalisierter Rotanteil, maximale und minimale Verkleinerung

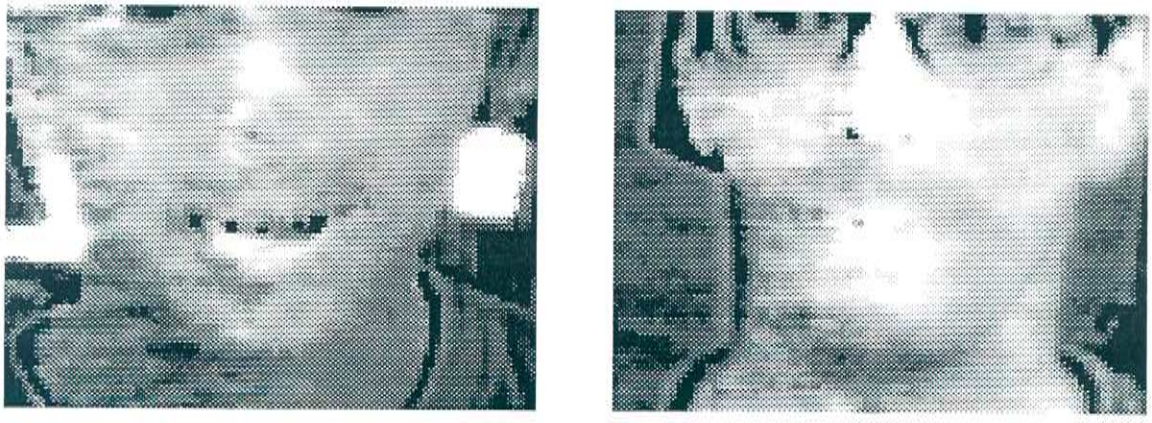

Abbildung 20: Beispiele für den Rötlichkeitswert von Gesichtern, links ist die Unterlippe gut zu erkennen

Zum Training mit den verschiedenen Vorverarbeitungen wurden jeweils 3 Bilder von 10 verschiedenen Personen verwendet, die mit den oben beschriebenen Translationen und Skalierungen dem Netz präsentiert wurden. Zum Test wurden 212 Bilder von 10 anderen Personen genutzt. In jedem dieser Bilder wurden 10 Bildausschnitte der Größe des Eingabefensters des Netzes zufällig ausgewählt. In diesen Bildausschnitten waren beide Mundwinkel mit einem Rand von mindestens 2 Pixeln in jede Richtung sichtbar. Durch dieses Testverfahren wird verhinderd, daß sich Fehler bei der Groberkennung störend auf die Bewertung der Feinerkennung auswirken. Ist nach völlig falscher Groberkennung der Mund nicht im Eingabefenster des Netzes zur Feinerkennung, gibt der Abstand von tatsächlicher und geschätzter Mundwinkelposition keine Auskunft mehr über die Qualität des Netzes zur Feinerkennung, sondern stellt für die Bewertung nur noch Rauschen dar. In Abbildung 21 ist der mittlere euklidische Abstand der geschätzten Positionen von den per Hand markierten Positionen gegenüber der Anzahl der präsentierten Testbilder aufgetragen. Dabei sind die mittleren Abstände für den linken und den rechten Mundwinkel einfach addiert worden. Der Fehler bei der Eingabe des normalisierten Rötlichkeitswertes liegt stets oberhalb von 45 Pixeln.Der Übersichlichkeit halber sind diese Ergebnisse daher nicht in die Grafik übernommen worden.

Das relativ schlechte Abschneiden des Farbwertes liegt meiner Einschätzung nach an den starken Schwankungen in der Rötlichkeit der Lippen zwischen verschiedenen Personen. In manchen Bildern des normalisierten Rötlichkeitswertes treten die Lippen klar hervor, bei anderen Bildern wie dem hier gezeigten sind die Lippen nicht nennenswert röter als die umgebende Haut. In Abbildung 20 sind zwei weitere Beispiele zu sehen. Dieses Abweichen von der intuitiven Einschätzung von Lippen als rötlich, könnte durch die speziellen Eigenschaften der in Videokameras verwendeten Rezeptoren verursacht sein. Pomerleau (92) berichtet, daß in dunklen Bildpartien der bläuliche Farbanteil relativ zu Rot- und Grünanteil stärker ist als in hellen Bildausschnitten.

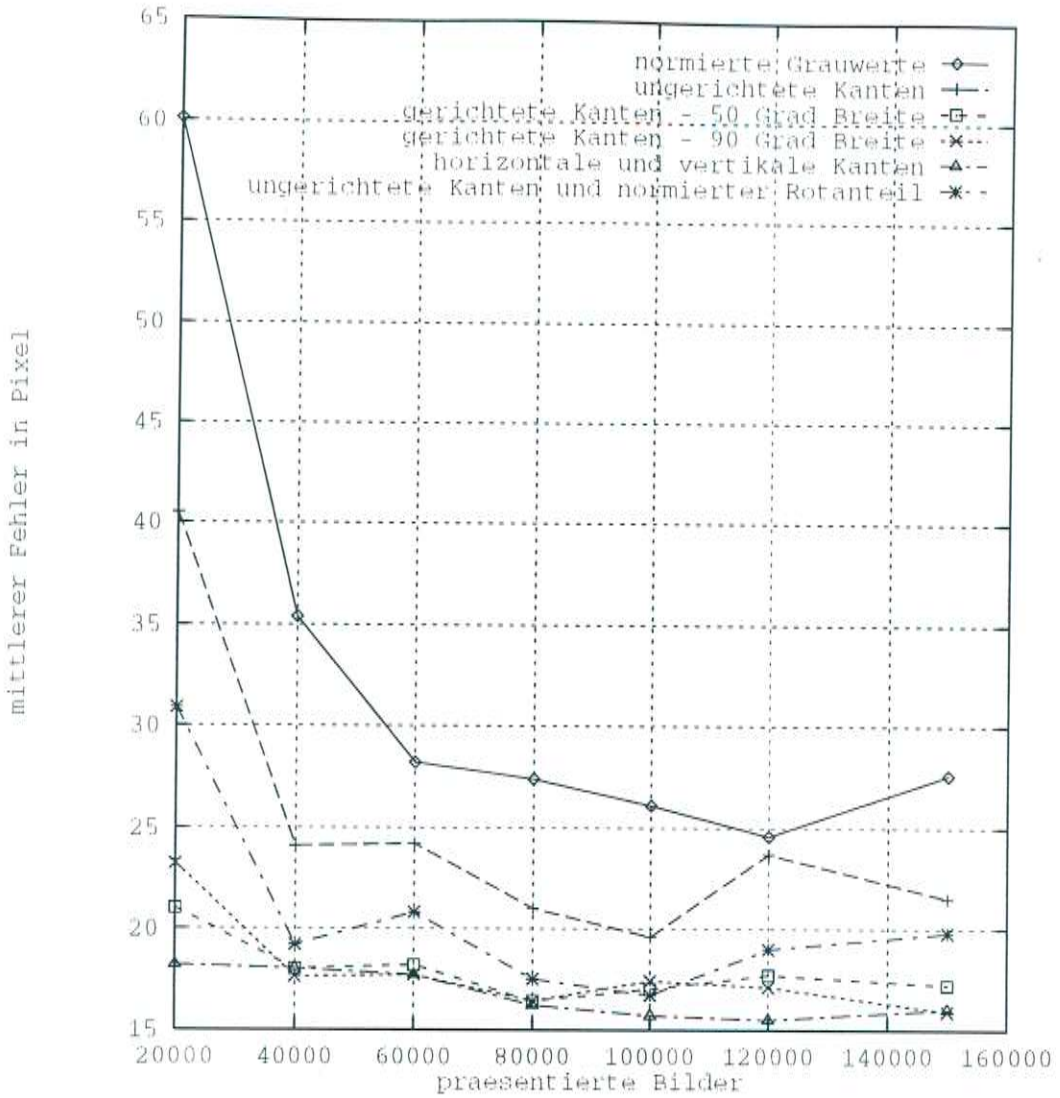

Abbildung 21: Abhängigkeit der Erkennungsleistung von der Vorverarbeitung

In den Tabellen 1 bis 6 sind genauere Verteilungen der Fehler für den linken Mundwinkel zu sehen. Dabei wurde für jede Vorverarbeitung die Trainingsdauer mit den besten Ergebnissen gewählt.

6.3 Interne Netzstruktur

Pomerleau (92) berichtet in seiner Dissertation von Experimenten mit verschiedenen Formen von versteckten Netzschichten. Dabei wurden verschiedene Anzahlen von versteckten Schichten, verschiedene Anzahlen von versteckten Einheiten und verschiedene Verbindungsmuster zwischen versteckter Schicht und Eingabeschicht (rezeptive Felder oder vollständige Verbindungen) getestet. Das Ergebnis dieser Experimente ist, daß die meisten getesteten Architekturvarianten vernünftige Ergeb-

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	28,0	22,4	7,4
2-3	28,0	23,6	20,5
4-5	19,7	16,0	21,8
6-9	11,7	16,3	21,9
10-13	2,2	5,8	6,8
14-18	1,0	2,6	2,9
> 18	9,4	13,3	18,7

Tabelle 1: Fehlerverteilung für Eingabe von normalisierten Grauwerten

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	23,6	26,2	6,5
2-3	24,6	26,8	19,2
4-5	19,1	18,7	23,9
6-9	18,6	13,5	25,5
10-13	6,0	2,5	8,3
14-18	2,4	2,8	3,1
> 18	5,7	9,5	13,5

Tabelle 2: Fehlerverteilung für Eingabe von Sobelamplituden ohne Richtungsbeschränkung

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	27,9	22,6	7,7
2-3	28,4	25,1	20,9
4-5	18,9	20,2	23,7
6-9	16,4	19,2	28,3
10-13	4,1	5,4	8,6
14-18	1,5	2,8	4,6
> 18	2,8	4,7	6,2

Tabelle 3: Fehlerverteilung für Eingabe von Sobelamplituden von Kanten mit maximal 25° Abweichung von der Horizontalen

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	24,9	21,8	7,2
2-3	29,1	25,7	21,1
4-5	20,7	21,1	24,2
6-9	16,5	20,8	28,1
10-13	4,9	3,9	9,9
14-18	1,4	2,8	3,2
> 18	2,5	4,8	6,3

Tabelle 4: Fehlerverteilung für Eingabe von Sobelamplituden von Kanten mit maximal 45° Abweichung von der Horizontalen

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	28,2	27,2	9,2
2-3	30,7	29,0	26,7
4-5	19,2	18,4	24,9
6-9	15,1	15,5	23,6
10-13	2,7	3,4	6,8
14-18	0,9	2,0	2,6
> 18	3,2	4,5	6,2

Tabelle 5: Fehlerverteilung für Eingabe von Sobelamplituden mit vertikalem und horizontalem Eingabefeld

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	23,5	23,2	6,7
2-3	28,2	26,7	21,3
4-5	20,1	19,5	22,7
6-9	17,4	18,2	28,1
10-13	5,3	4,1	9,5
14-18	1,4	1,0	2,2
> 18	4,1	7,3	9,5

Tabelle 6: Fehlerverteilung für Eingabe von Sobelamplituden ohne Richtungsbeschränkung und normalisiertem Rotanteil

nisse erbrachten, ohne daß eine Variante die o.g. Implementierung mit vier versteckten Einheiten, die vollständig mit der Eingabeschicht verbunden sind, wesentlich übertrifft.

Diese Ergebnisse ließen die Erwartung entstehen, daß auch für das Problem der Lippenlokalisierung die Anzahl der versteckten Einheiten nicht von entscheidender Bedeutung ist. Diese Erwartung wurde durch Experimente mit Netzen mit unterschiedlichen Zahlen versteckter Einheiten pro rezeptivem Feld bestätigt. Die Bilder 22 und 23 zeigen Ergebnisse für Netze zur Groberkennung bzw. Feinerkennung.

In den oben beschriebenen Versuchen wurden die Netze zur Feinerkennung meist mit 150.000 oder 180.000 Präsentationen von Bildern trainiert. In Abbildung 24 kann man sehen, daß noch längeres Trainieren keine sichtbaren Verbesserungen bringt.

Es wurden auch einige Experimente betreffend der Aufteilung des Eingabefeldes in rezeptive Felder durchgeführt. Die meisten dazu vorliegenden Ergebnisse sind im Abschnitt zu Variationen in der Auflösung des Eingabefeldes zu sehen. Dort werden Architekturen mit 16 bzw. 36 rezeptiven Feldern verglichen. In einem weiteren Versuch wurde ein Netz zur Groberkennung mit nur 4 rezeptiven Feldern bei einer Größe des Eingabefeldes von 32×32 Pixeln getestet. Dieses Netz hatte die gleiche Gesamtanzahl an versteckten Einheiten wie das Netz mit 16 rezeptiven Feldern, dessen Ergebnisse in Abbildung 22 zu sehen sind. Jedem rezeptiven Feld waren daher nicht 8 sondern 32 versteckte Einheiten zugeordnet. Alle anderen Parameter (Vorverarbeitung, Ausgaberepräsentation etc.) blieben gleich. Der durchschnittliche Fehler bei einem Trainingslauf von 100.000 präsentierten Bildern sank nicht unter 25 Pixel, was schlechter als das Ergebnis des Netzes mit 16 rezeptiven Feldern ist.

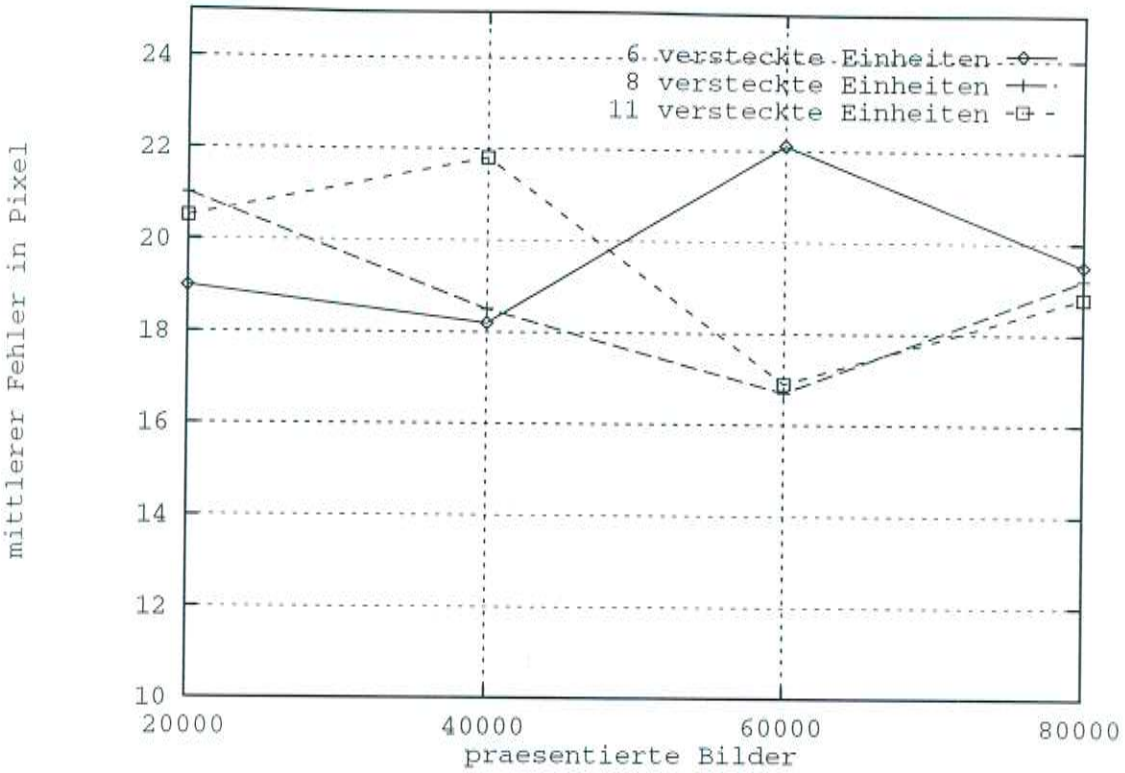

Abbildung 22: Abhängigkeit der Erkennungsleistung von der Anzahl der versteckten Einheiten pro rezeptivem Feld (Groberkennung)

6.4 Ausgaberepräsentation

Auch in diesem Abschnitt möchte ich mich noch einmal auf die Arbeit von Pomerleau (92) beziehen. Neben Untersuchungen zur Steuerung von Fahrzeugen hat er auch Experimente zur Steuerung eines Roboterarms durchgeführt. An der Spitze des Armes ist eine Schraube angebracht, die in Löcher eingeführt werden muß. Die nicht sehr große Präzision des Roboterarmes erfordert eine sensorgeführte Feineinstellung, um die Schraube direkt über den Löchern zu positionieren. Dazu ist eine Kamera an der Spitze des Armes befestigt. Zur Steuerung des Armes muß die Position des Loches in der Bildebene bestimmt werden. Pomerleau verwendet dabei Netze mit Ausgabeeinheiten für X- und Y-Koordinate. Beide Gruppen von Ausgabeeinheiten werden wie die Ausgabeeinheiten bei der Fahrzeugsteuerung mit einer von der besten Ausgabeeinheit abfallenden Aktivierungsstärke trainiert.

Für die Erkennung der Mundwinkel bietet sich diese Form der Ausgaberepräsentation ebenfalls an. Experimente zeigen jedoch, daß eine Anordnung der Ausgabeeinheiten in einem Gitter bessere Erkennungsergebnisse erzielt. Bevor ich auf diese Experimente eingehe, möchte ich anschaulich zeigen, welche Nachteile eine Trennung der Berechnung von X- und Y-Koordinate haben kann. Dazu möchte ich zunächst auf die von Pomerleau getesteten Ausgaberepräsentationen zur Fahrzeugsteuerung

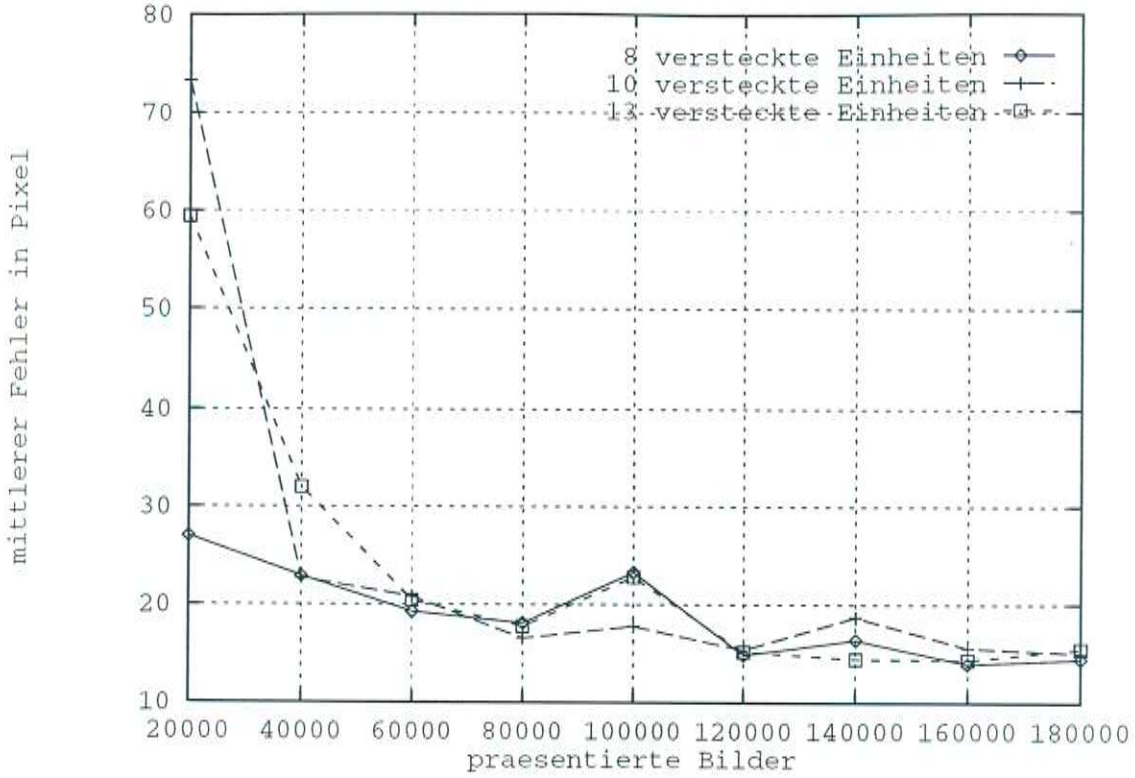

Abbildung 23: Abhängigkeit der Erkennungsleistung von der Anzahl der versteckten Einheiten pro rezeptivem Feld; Feinerkennung mit 36×60 Eingabefeld

eingehen. Dies sind:

- Eine Zeile von Ausgabeeinheiten, wobei die Aktivierung beim Training normalverteilt um die vorgegebene Steuerungsrichtung abfällt.
- Eine Zeile von Ausgabeeinheiten, bei der der Einheit, die der vorgegebenen Steuerungsrichtung am nächsten ist, die maximale Aktivierung und allen anderen Einheiten eine minimale Aktivierung vorgegeben wird.
- Eine einzelne Ausgabeeinheit, deren Aktivierung kontinuierlich von den möglichen Steuerungsrichtungen abhängt (z.B. minimale Aktivierung $\Rightarrow$ nach links; maximale Aktivierung $\Rightarrow$ nach rechts).

Die Repräsentation durch eine einzelne Ausgabeeinheit schnitt bei Pomerleaus Experimenten am schlechtesten ab. Er begründet dies unter anderem mit folgender Überlegung. In einem Eingabefeld einer Straße sind Merkmale vorhanden, die auf eine Biegung der Straße nach links und für eine Biegung nach rechts sprechen. Bei Verwendung einer einzelnen Ausgabeeinheit werden diese Hinweise häufig zur Steuerungsrichtung "geradeaus" zusammengefaßt. Diese Richtung ist aber wahrscheinlich falsch, da für sie keine Hinweise vorliegen. Ein Netz, das eine Zeile von

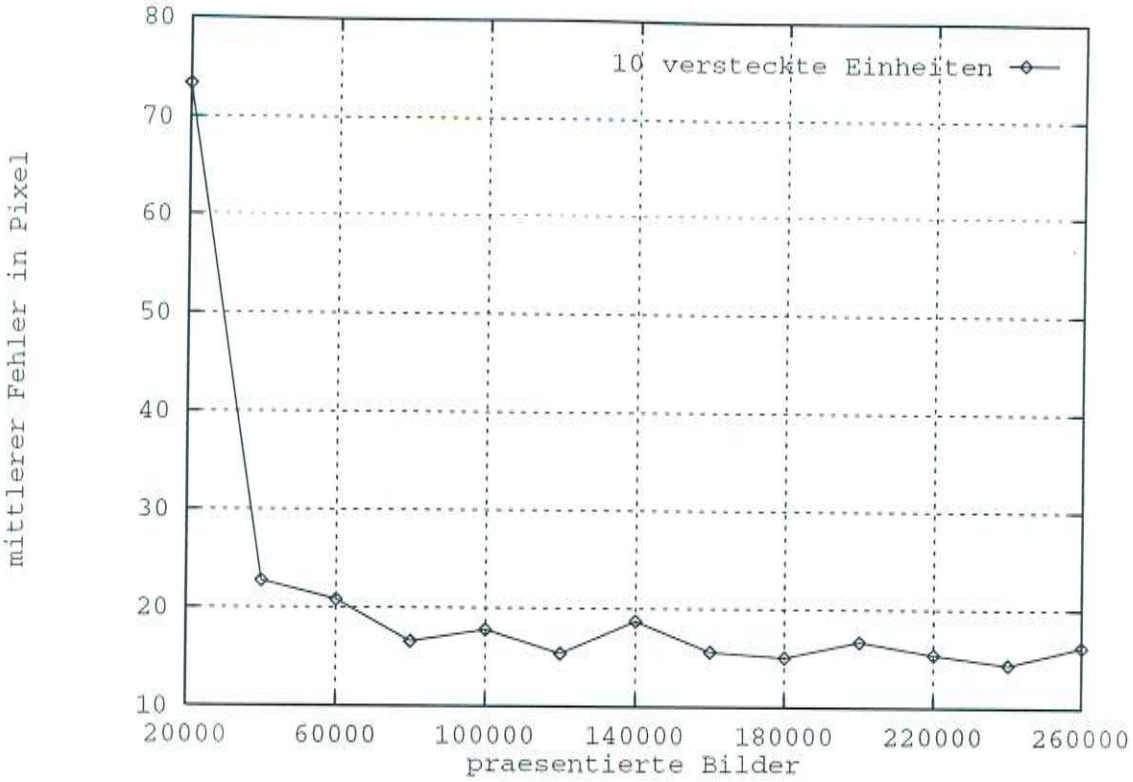

Abbildung 24: Fortsetzung des Trainingsverlaufes für das Netz mit 10 versteckten Einheiten pro Feld aus der vorhergehenden Abbildung

Ausgabeeinheiten hat, ordnet die Merkmale, die für ein Steuern nach links bzw. rechts sprechen, den zugehörigen Ausgabeeinheiten zu. Es wird dann die Richtung gewählt, für die die besten Hinweise vorliegen.

Ein ähnliches Problem kann bei der getrennten Ermittlung von X- und Y-Koordinate bei der Lokalisierung der Mundwinkel auftreten. Es ist möglich, daß an in der x-Koordinate verschiedenen Stellen im Kinnbereich in gleicher Bildhöhe einige schwächere Hinweise z.B. für den linken Mundwinkel vorliegen, die aber jeweils allein schwächer sind als die von der tatsächlichen Position des Mundwinkels ausgehenden Hinweise. Bei separater Bestimmung der Y-Koordinate werden diese schwächeren Hinweise jedoch auf eine Ausgabeeinheit summiert und übertreffen dann möglicherweise die Aktivierung durch die Hinweise von der realen Position des Mundwinkels. Bei einer Anordnung der Ausgabeeinheiten in einem Gitter sind solche Probleme nicht zu erwarten. Dort werden die erwähnten schwächeren Hinweise aus der Kinnregion auf verschiedene Ausgabeeinheiten, die sich in der X-Koordinate unterscheiden, verteilt.

Bild 25 zeigt die Erkennungsgenauigkeit in der Feinerkennung bei der Anordnung der Ausgabeeinheiten in einem Gitter, bei getrennten Ausgabeeinheiten für X- und Y-Koordinate mit gleicher bzw. getrennter versteckten Schicht für X- und Y-

Koordinate. Diese Trennung wurde von Pomerleau (94) und Baluja und Pomerleau (94) als vorteilhaft beschrieben.

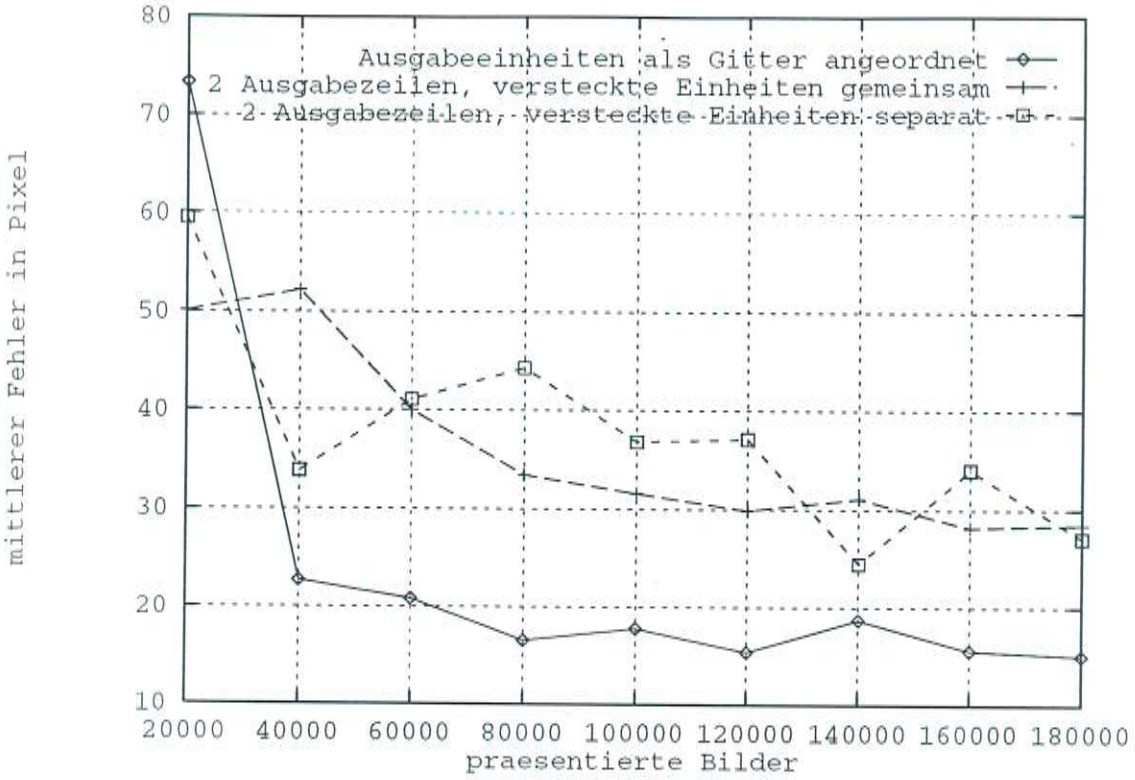

Abbildung 25: Abhängigkeit der Erkennungsleistung von der Ausgaberepräsentation

Eine genauere Analyse des auftretenden Fehlers bei getrennter Berechnung von X- und Y-Koordinate (siehe Tabelle 7) läßt die o.g. Begründung der Fehler plausibel erscheinen. Fehler treten hauptsächlich in der Y-Koordinate auf. Die den Mundwinkeln am ähnlichsten erscheinenden Objekte sind aber in der Nasen- und Kinnpartie zu finden. Offenbar führt die Summation der in diesen Bereichen auftretenden Hinweise zu verstärkten Fehlerkennungen. Auch bei der Groberkennung wurde eine Ausgaberepräsentation mit zwei Ausgabezeilen für X- und Y-Koordinate getestet. Hier unterscheiden sich die Ergebnisse für die verschiedenen Ausgaberepräsentationen nicht so stark. Das Netz mit getrennter Repräsentation von X- und Y-Koordinate macht jedoch möglicherweise leichter sehr große Fehler als das sonst benutzte Netz in dem die Ausgabeeinheiten in einem Gitter angeordnet sind (vgl. Tabellen 8 und 9).

6.5 Interpolation der Netzausgaben (Anpassung einer Normalverteilung)

Um den Rechenaufwand bei Training, Test und Anwendung zu begrenzen, sind in den implementierten Netzen nur zu jeder zweiten Eingabezeile und -spalte Ausgabeeinheiten vorhanden. Im Training wird die Aktivierung der Ausgabeeinheiten

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	21,8	24,2	5,5
2-3	24,7	27,6	17,4
4-5	16,1	15,8	18,0
6-9	21,1	12,3	24,3
10-13	6,9	2,6	9,4
14-18	4,0	1,7	4,9
> 18	5,4	15,8	20,5

Tabelle 7: Fehlerverteilung für getrennte versteckte Schicht und getrennte Ausgabezeilen für X- und Y-Koordinate (140000 Iterationen)

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-4	33,0	19,3	5,2
5-9	27,8	20,3	17,5
10-14	21,2	17,5	18,9
15-19	10,8	17,5	18,4
20-29	3,8	17,8	28,8
30-39	2,0	6,1	8,5
40-49	0,0	0,5	0,5
50-59	0,0	0,5	0,5
> 59	1,4	0,5	1,9

Tabelle 8: Fehlerverteilung Groberkennung mit getrennter Ausgabe X- und Y-Koordinate - bestes Ergebnis des Trainingslaufs - mittlerer Fehler in Pixel dx: 9,3; dy: 13,9; insgesamt 18,5

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-4	37,3	22,6	8,5
5-9	32,5	19,3	18,9
10-14	17,0	16,5	19,8
15-19	8,0	17,9	17,5
20-29	3,3	17,9	25,9
30-39	1,9	5,2	8,5
40-49	0,0	0,0	0,5
50-59	0,0	0,5	0,5
> 59	0,0	0,0	0,0

Tabelle 9: Fehlerverteilung Groberkennung ohne Interpolation; mittlerer Fehler in Pixel dx : 7,5; dy : 13,3; insgesamt: 16,7

jedoch entsprechend ihrem Abstand zur Position des Mundes im Eingabefeld berechnet. Daher kommt es vor, daß keiner Ausgabeeinheit die volle Aktivierung von 1 vorgegeben wird, sondern zwei oder vier Ausgabeeinheiten die gleiche maximale Aktivierung < 1 erhalten. Eine Anpassung einer Normalverteilung an die vorgegebenen Ausgabewerte würde eine Rekonstruktion der Mundposition in der Auflösung des Eingabefeldes möglich machen. Es wurde nun getestet, ob eine Anpassung an eine Normalverteilung auf die tatsächlichen Ausgabewerte mit Interpolation zwischen den Ausgabeeinheiten die Genauigkeit des Netzes verbessert. Bei dieser Interpolation wurden auch die Positionen des Eingabefeldes zwischen den Ausgabeeinheiten getestet. Die Anpassung wurde als Faltung mit einer Normalverteilung in einer Umgebung um die Ausgabeeinheit mit der maximalen Aktivierung vorgenommen. Die Ergebnisse dieser Interpolation wurde mit den Ergebnissen der "einfachen" Strategie verglichen, bei der die Position der Ausgabeeinheit mit maximaler Aktivierung als Position des Mundes angenommen wird. Die Ergebnisse sind in den Tabellen 9 und 10 zu sehen. Offensichtlich bringt eine Interpolation keine Vorteile. Diese Experiment wurde nur für die Groberkennung durchgeführt. Da die Trainingsmethode bei der Feinerkennung jedoch die gleiche ist, wird wohl auch dort Interpolation zu keinen Verbesserungen führen.

6.6 Änderungen der Auflösung der Netze

In diesem Abschnitt möchte ich Experimente beschreiben, bei denen die Auflösung des Netzes gesteigert wurde. Es wurden mehr Eingabeeinheiten, mehr rezeptive Felder und mehr Ausgabeeinheiten verwendet. Trainingsdaten, Vorverarbeitung und Größe des Eingabefensters sowie Art der Ausgaberepräsentation wurden jedoch konstant gehalten. Die größeren Netze in Grob- und Feinerkennung unter-

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-4	42,5	22,6	6,1
5-9	35,8	14,6	18,9
10-14	13,7	12,3	17,9
15-19	3,8	19,3	20,3
20-29	0,5	22,2	26,5
30-39	0,5	6,6	7,5
40-49	0,0	1,9	2,4
50-59	0,0	0,5	0,5
> 59	0,0	0,0	0,0

Tabelle 10: Fehlerverteilung Groberkennung mit Interpolation; mittlerer Fehler in Pixel dx: 6,6; dy: 14,7; insgesamt: 17,6

scheiden sich von den entsprechenden kleineren Netzen durch eine um den Faktor 1.5 bessere Auflösung. Dies führt dazu, daß sie die $1,5^2 = 2,25$ -fache Anzahl von Eingabeeinheiten, versteckten Einheiten und Ausgabeeinheiten haben. Da die Ausgabeeinheiten vollständig mit allen versteckten Einheiten verbunden sind, steigt die Anzahl der Gewichte für diese Verbindungen sogar um das $2,25^2 \simeq 5$ -fache. Die Anzahl der Gewichte zwischen Eingabeschicht und versteckter Schicht steigt nicht so schnell. Sie ist aber ohnehin im Vergleich zur Anzahl der Gewichte zwischen versteckter Schicht und Ausgabeschicht gering. Insgesamt steigt also die Anzahl der Gewichte und damit auch die Anzahl aller Rechenoperationen für Training, Test und Anwendung um das ca. 5-fache.

Die Abbildungen 26 und 27 zeigen die Ergebnisse des Vergleichs der Leistungsfähigkeiten von kleineren und größeren Netzen bei der Grob- und Feinerkennung. Die Netze der Groberkennung hatten Eingabefelder der Größe 32×32 bzw. 48×48. Bei der Feinerkennung waren die Eingabefelder 24×40 bzw. 36×60 Einheiten groß. In beiden Fällen hatte das kleinere Netz 4×4 und das größere Netz 6×6 rezeptive Felder. Bei der Groberkennung wirkt sich eine Aufteilung in eine größere Anzahl von Eingabefenstern trotz der gesteigerten Auflösung eher negativ aus. Bei der Feinerkennung führt die Steigerung der Auflösung zu einer etwas besseren Genauigkeit. Der Durchschnittsfehler für die Lokalisierung des linken Mundwinkels sank von 7,8 auf 6,9 Pixel. Die Tabellen 11 und 12 zeigen, daß diese Steigerung der Erkennungsgenauigkeit durch die ungefähr gleich gebliebene Anzahl der groben Fehler begrenzt wird (Daten jeweils für das beste Ergebnis des linken Mundwinkels).

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	25,7	22,8	6,9
2-3	28,0	26,2	21,5
4-5	19,8	19,7	22,6
6-9	16,9	16,9	28,7
10-13	5,0	4,9	10,7
14-18	2,6	4,9	3,7
> 18	2,0	4,6	5,9

Tabelle 11: Fehlerverteilung für Netz mit 24×40 Eingabefeld

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	30,8	30,3	11,6
2-3	32,5	30,5	30,2
4-5	18,3	18,3	24,9
6-9	12,3	12,5	20,9
10-13	3,0	1,3	4,2
14-18	1,0	1,5	1,3
> 18	2,1	5,6	6,9

Tabelle 12: Fehlerverteilung für Netz mit 36×60 Eingabefeld

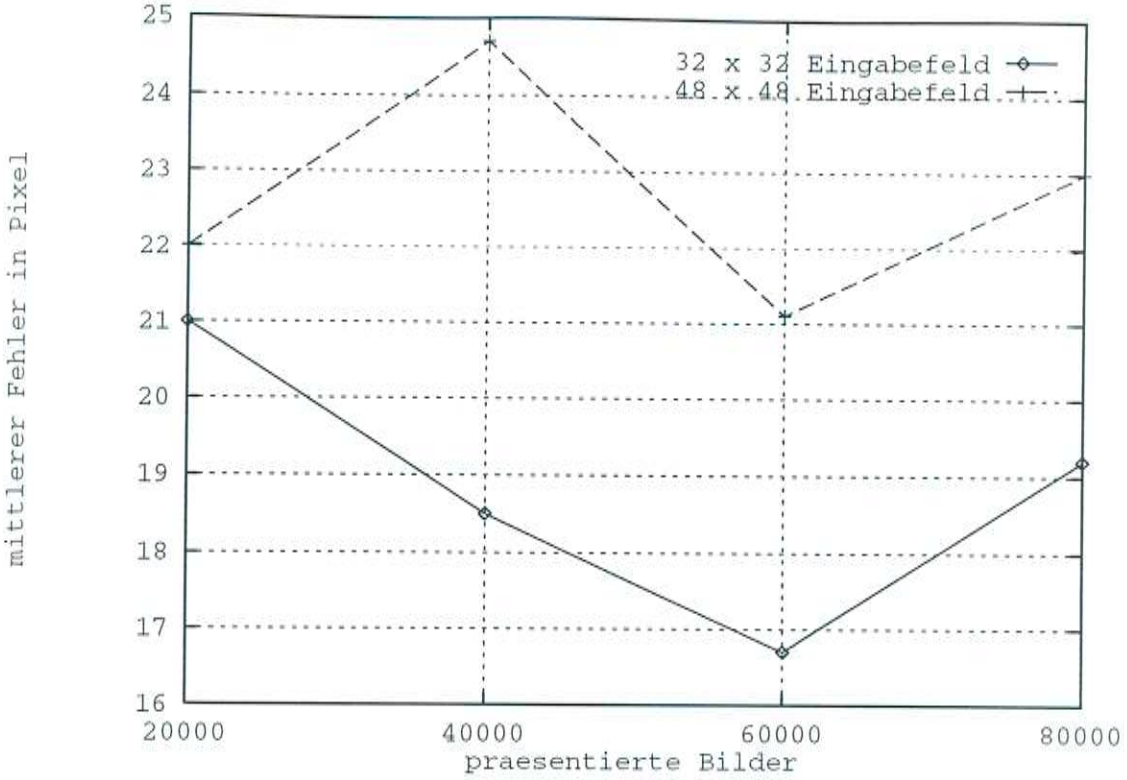

Abbildung 26: Zwei Netze unterschiedlicher Auflösung zur Groberkennung

6.7 Ergebnisse für das Gesamtsystem

Das beste bei der Groberkennung erzielte Ergebnis wurde mit einem Netz erzielt, das als Eingabe horizontale und vertikale Kanten jeweils mit einer Bandbreite von 90° erhält. Der durchschnittliche Fehler der geschätzten Position des Mundes beträgt dabei 16,0 Pixel (siehe Abbildung 15). Um die in Abschnitt 6.1. erwähnten Schwächen der Testbilder etwas ausgleichen, wurde die Groberkennung auch auf 93 weiteren Bildern von insgesamt 20 Personen getestet, bei denen das Gesicht stellenweise sehr am Rand des Bildes lag. In jedem Bild waren jedoch Haaransatz (zumindest teilweise), Kinn und die beiden Augenbrauen vollständig zu sehen. Für diese Testbilder wurde ein durchschnittlicher Fehler von 24,7 Pixel erreicht.

Das beste Ergebnis bei der Feinerkennung basierte auf einer Vorverarbeitung mit horizontalen Kanten einer Bandbreite von 90°. Mit der oben beschriebenen Testmethode wurde ein mittlerer Fehler von 7,4 Pixeln für den linken Mundwinkel und ein mittlerer Fehler von 7,1 Pixeln für den rechten Mundwinkel erzielt (siehe auch Abbildung 21).

Der mittlere Fehler des Gesamtsystems, das durch die Zusammenarbeit dieser beiden Netze gebildet wird, beträgt für die auch sonst verwendeten 212 Testbilder beim linken Mundwinkel 7,6 und beim rechten Mundwinkel 7,4 Pixel. Die Fehlerverteilung

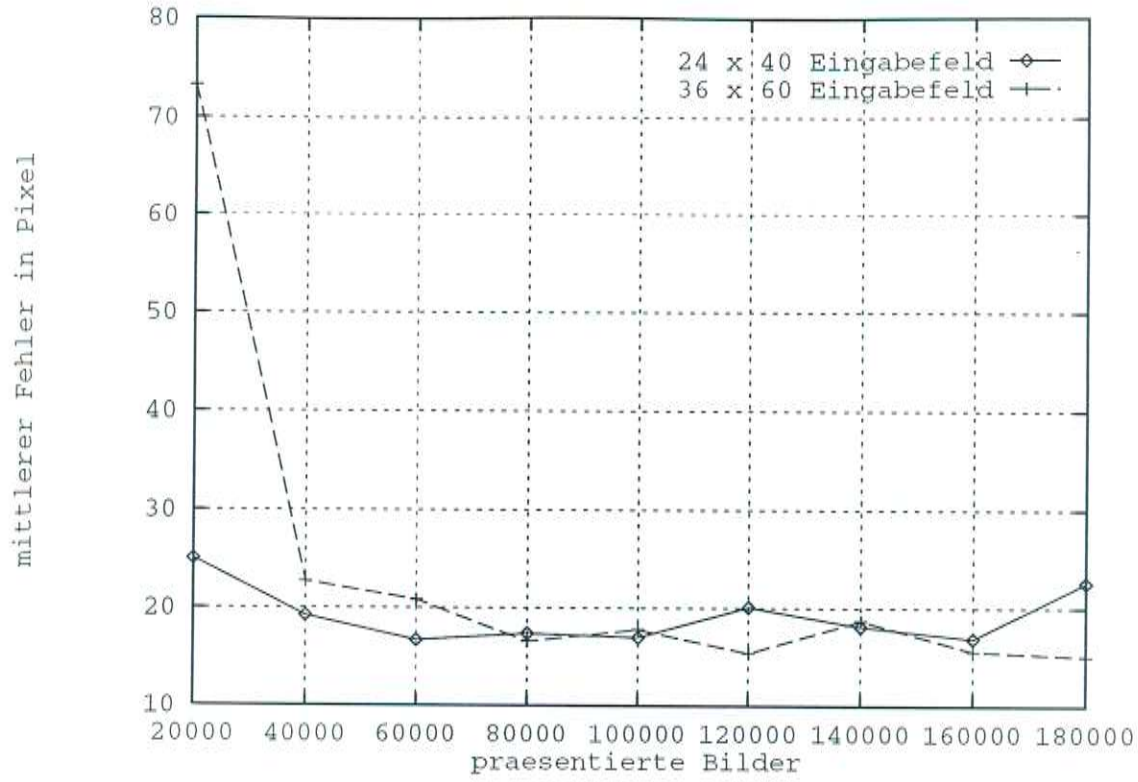

Abbildung 27: Zwei Netze unterschiedlicher Auflösung zur Feinerkennung

für den linken Mundwinkel ist in Tabelle 13 zu sehen. Der mittlere Abstand zwischen den Mundwinkeln in den Testbildern war 42 Pixel. In Abbildung 28 ist das Ergebnis bei der Anwendung des Lippenfinders auf jeweils das siebte Bild aller 20 benutzten Sequenzen zu sehen. Das Quadrat zeigt das Ergebnis der Groberkennung und die beiden Kreuze die geschätzte Position der Mundwinkel.

6.8 Personenabhängige Detektion der Lippenwinkel

In den bisher beschriebenen Experimenten wurde die Fähigkeit zur personenunabhngigen Detektion der Lippen getestet. Dazu wurden neuronale Netze mit Bildern einer Gruppe von Personen trainiert und mit Bildern von anderen Personen getestet. Im Rahmen des Projektes zum Lippenlesen wird bis jetzt jedoch nur eine sprecherabhängige Spracherkennung durchgeführt. Dabei werden nur Bilder einer einzigen Person benutzt. Um dafür optimale Ergebnisse zu erzielen, wurde auch ein Netz nur mit Bildern dieser Person trainiert und mit anderen unter gleichen Bedingungen aufgenommenen Bildern getestet. Beim Test wurden 401 Bilder aus über 100 allerdings recht ähnlichen Bildsequenzen verwendet. In den Abbildungen 29 und 30 sind die Ergebnisse der Groberkennung und Feinerkennung zu sehen. Beim Test der Feinerkennung wurde das in Abschnitt 6.1. angesprochene Verfahren der zufälligen Plazierung der Eingabemaske angewandt. Zur Begrenzung des Rechenaufwandes

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	24,5	25,5	6,1
2-3	37,3	28,3	30,2
4-5	18,9	20,2	27,4
6-9	11,3	14,6	21,2
10-13	3,3	3,3	4,3
14-18	0,9	1,0	2,3
> 18	3,8	7,1	8,5

Tabelle 13: Fehlerverteilung für das Gesamtsystem bei der personenunabhängigen Lippenlokalisierung (linker Mundwinkel)

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	65,6	67,1	45,6
2-3	31,4	27,4	46,1
4-5	2,2	2,8	5,5
6-9	0,0	1,5	1,5
10-13	0,0	0,3	0,3
14-18	0,3	0,0	0,0
> 18	0,5	1,0	1,0

Tabelle 14: Fehlerverteilung des Gesamtsystems bei identischer Trainings- und Testperson

wurden allerdings nur 2 zufällige Positionen pro Bild ausgewählt. Es ist wiederum die Summe des mittleren Fehlers für beide Mundwinkel angegeben. Beim Test des Gesamtsystems wurden die jeweils besten Netze zur Grob- und Feinerkennung verwendet. Der mittlere Fehler für den linken und rechten Mundwinkel betrug 2,4 bzw. 2.6 Pixel. Tabelle 14 zeigt die Fehlerverteilung des Gesamtsystems für den linken Lippenwinkel. Der mittlere Abstand zwischen den Mundwinkeln in den Testbildern betrug 26,3 Pixel.

Abbildung 28: Beispielbilder aus den benutzten Sequenzen; die ersten 10 Sequenzen (von links oben nach rechts unten) Grundlage des Trainings; die letzten 10 Sequenzen: Testbilder (Das Rechteck zeigt das Ergebnis der Groberkennung an.)

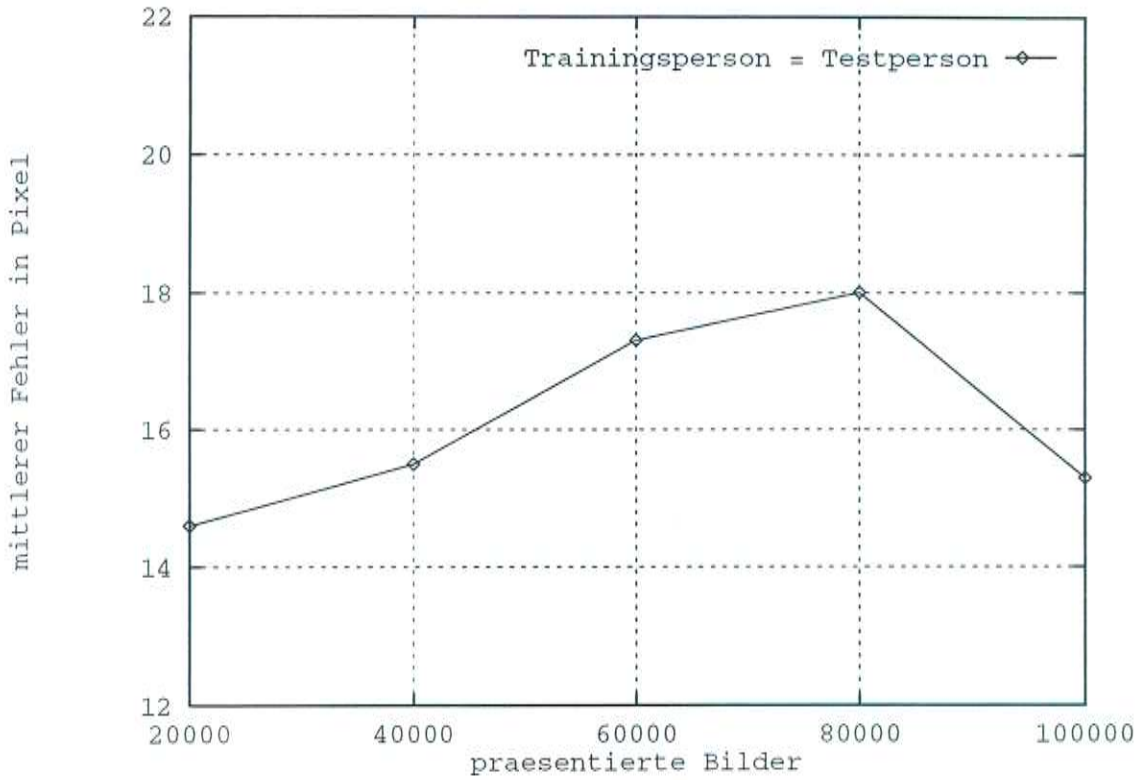

Abbildung 29: mittlerer Fehler bei der Groberkennung einer bestimmten Person

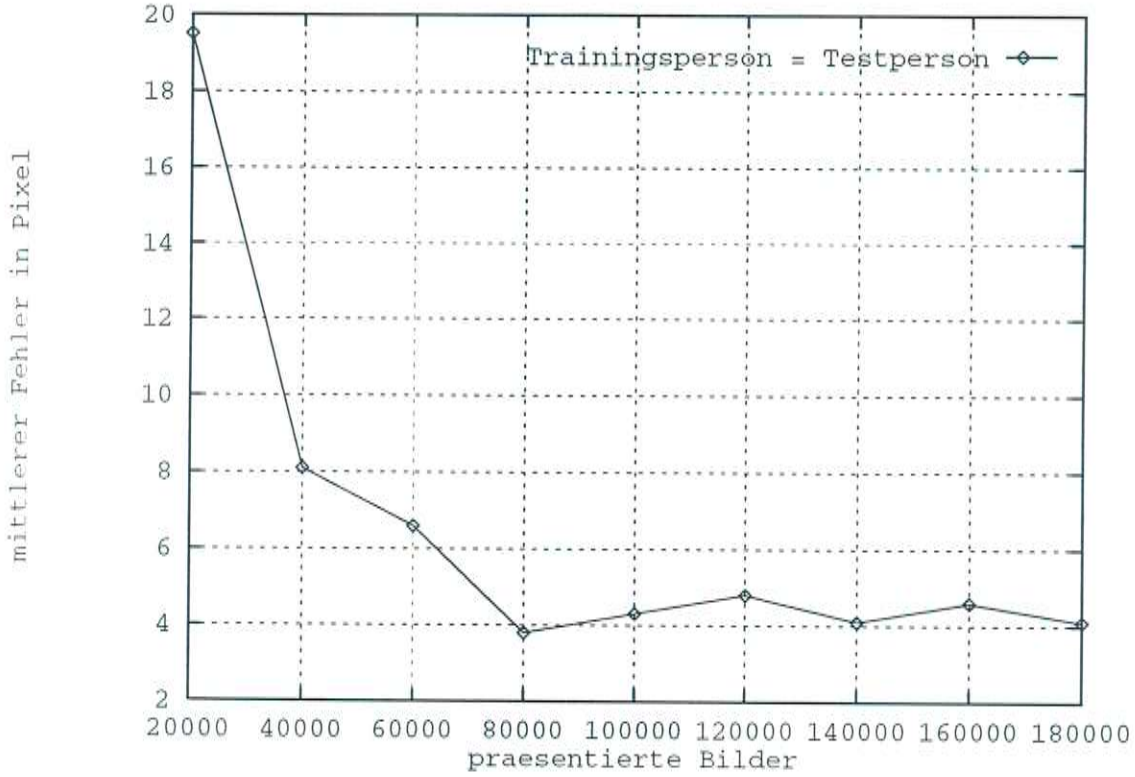

Abbildung 30: mittlerer Fehler bei der Feinerkennung einer bestimmten Person

7 Translationsinvariante Erkennung durch verschiebbare Templates

Anhand des Überblicks über bisherige Arbeiten im Bereich Lokalisierung von Gesichtern und Gesichtsteilen in Kapitel 3.3, läßt sich erkennen, daß in sehr vielen Arbeiten Korrelation oder ähnliche Techniken (z.B. Plazierung der Eingabeeinheiten eines mehrschichtigen Perzeptrons an verschiedenen Bildpositionen) angewandt werden. In diesem Kapitel soll die Leistungsfähigkeit solcher Verfahren mit den Ergebnissen für die in Kapitel 5 und 6 beschriebene Architektur verglichen werden. Insbesondere soll die Frage untersucht werden, ob durch die Verwendung einer templatebasierten Erkennung die Genauigkeit bei der Feinerkennung der Lippenwinkel gesteigert werden kann. Dieses Kapitel ist in mehrere Abschnitte aufgeteilt. In Abschnitt 7.1 werden Experimente zur Feinerkennung mit normalisierter Kreuzkorrelation beschrieben. In Abschnitt 7.2 wird ein Verfahren beschrieben, bei dem zur Erkennung ein mehrschichtiges Perzeptron an verschiedene Bildpositionen plaziert wird. Analog zu Kapitel 6 werden verschiedene Vorverarbeitungen, Netzgrößen und Eingabefeldgrößen getestet. In Abschnitt 7.3 wird gezeigt, wie durch die Verwendung von rezeptiven Feldern der Rechenaufwand für das Verfahren aus 7.2 ohne Reduzierung der Erkennungsgenauigkeit erheblich reduziert werden kann.

7.1 Feinerkennung mit normalisierter Kreuzkorrelation

Wie bereits in Kapitel 3.3 erwähnt, werden bei der normalisierten Kreuzkorrelation Bildausschnitte und im Bild zu findende Prototypen (Templates) miteinander verglichen. Dazu werden die Intensitätswerte von Bildausschnitt und Template wie zwei Vektoren miteinander multipliziert und das Ergebnis bezüglich Templategröße, Durchschnittshelligkeit und Kontrast so normalisiert, daß das Resultat stets zwischen -1,0 und 1.0 liegt.

Um die Ergebnisse mit den in Kapitel 6 genannten Resultaten vergleichbar zu halten, hat das Rechteck der mit Korrelation getesteten möglichen Positionen des zu suchenden Mundwinkels die gleiche Größe wie das Eingabefenster des in Kapitel 5 beschriebenen Netzes (36 × 60 Pixel). Das zu durchsuchende Bild wurde wie in Kapitel 5 beschrieben zuvor noch um den Faktor 1,875 verkleinert. Der Einfachheit halber wurde nur nach dem linken Mundwinkel gesucht. Die verwendeten Templates hatten die Größe 24 × 30 Pixel. Der gesuchte Mundwinkel befand sich jeweils in der Mitte des Templates. Um ein Template an jeder der 36 × 60 Positionen mit dem Bild vergleichen zu können, hatte der aus dem Bild verwendete Ausschnitt die Größe 60 × 90 Pixel. Zum Test der normalisierten Kreuzkorrelation wurde dieses Eingabefenster so plaziert, daß dessen Mitte mit der Mitte des Mundes zusammenfiel. Für alle 36 × 60 möglichen Positionen des Templates wurde der Korrelationswert berechnet und die Position mit dem maximalen Korrelationswert als geschätzte Mundwinkelposition ausgewählt. In einem Experiment wurde

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	34,0	25,0	9,9
2-3	26,4	21,2	18,8
4-5	14,2	9,4	16,5
6-9	10,4	7,6	11,3
10-13	1,4	3,5	3,8
14-18	0,0	8,3	7,2
> 18	13,7	25,0	32,5

Tabelle 15: Fehlerverteilung linker Mundwinkel; normalisierte Kreuzkorrelation mit einem Template; Durchschnittsfehler dx: 6,5 dy: 8,9 insgesamt: 12,7 Pixel

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	36,3	24,5	8,9
2-3	25,0	21,7	26,4
4-5	10,9	16,0	17,5
6-9	5,2	9,4	12,7
10-13	0,9	11,8	5,7
14-18	0,5	7,2	1,4
> 18	21,2	9,4	27,4

Tabelle 16: Fehlerverteilung linker Mundwinkel; normalisierte Kreuzkorrelation mit drei Templates; Durchschnittsfehler dx: 9,2 dy: 6,6 insgesamt: 12,7 Pixel

ein Template des Mundwinkels eines geschlossenen Mundes, in einem anderen Experiment drei Templates von Mundwinkeln mit verschiedenem Grad der Öffnung des Mundes verwendet. Zum Test wurden die in Kapitel 6 genannten 212 Bilder von 10 verschiedenen Personen verwendet. Die Templates wurden aus Bildern anderer Personen erzeugt. Die Tabellen 15 und 16 zeigen die Resultate.

7.2 Feinerkennung durch templateartiges Verschieben eines neuronalen Netzes

Bei der in Kapitel 5 beschriebenen Architektur wird die Positionserkennung durch die Auswertung eines einzigen Bildausschnittes vorgenommen. Das Netz hat daher eine große Anzahl von Ausgabeeinheiten. Die gesuchte Position ergibt sich aus

der Position der Ausgabeeinheit mit der höchsten Aktivierung. Um dieses Netz zu trainieren, müssen ihm verschiedene Bilder des Mundbereichs präsentiert werden, in denen sich die gesuchten Mundwinkel an allen zu erwartenden Positionen befinden. Bei der in diesem Abschnitt beschriebenen Netzarchitektur wird das Netz nur darauf trainiert, das zu erkennende Objekt (Mundwinkel) an einer Position des Eingabefensters zu detektieren. Dazu ist für jeden Mundwinkel nur eine Ausgabeeinheit erforderlich, deren Aktivierung anzeigt, ob sich das gesuchte Objekt an der festgelegten Position des Eingabefensters befindet. Um die Position des Objektes im Bild festzustellen, wird das Eingabefenster an verschiedene Bildpositionen plaziert und jeweils die Aktivierung der Ausgabeeinheiten berechnet. Als Position der Mundwinkel werden die Plazierungen des Eingabefensters angenommen, die die höchsten Aktivierungen der entsprechenden Ausgabeeinheiten bewirken.

Wie in Kapitel 5 wird ein mehrschichtiges Perzeptron mit einer versteckten Schicht verwendet. Die versteckte Schicht bestand in den durchgeführten Experimenten aus 1 bis 12 Einheiten, die vollständig mit der Eingabeschicht verbunden waren. Eine Skizze der Architektur ist in Abbildung 31 zu sehen. Die meisten Experimente wurden wie in Abschnitt 7.1 mit einem 24×30 Pixel großen Eingabefenster durchgeführt. Auch die Verkleinerung der verwendeten Bilder wurde wie in Abschnitt 7.1 und Kapitel 5 durchgeführt. Beim Training wurde das Eingabefenster auf verschiedene Positionen im Mundbereich gelegt. Dabei wurden der in Kapitel 7.1 erwähnte 60×90 Pixel umfassende Bereich benutzt. Dieser Bildausschnitt wurde vorverarbeitet und die Ergebnisse normiert. Zur Begrenzung des Rechenaufwandes erfolgte die Normierung der Eingabewerte nicht für jede Position des Eingabefensters aufs neue sondern nur einmal für den gesamten Bildausschnitt über den das Eingabefenster bewegt wird. Falls sich der Mundwinkel genau in der Mitte des Eingabefensters befand, wurde der Ausgabeeinheit für diesen Mundwinkel die Aktivierung 1,0 vorgegeben. Bei Abweichungen von dieser optimalen Position sank die vorgegebene Aktivierung entsprechend einer Normalverteilung analog zum in Kapitel 5 beschriebenen Netz.

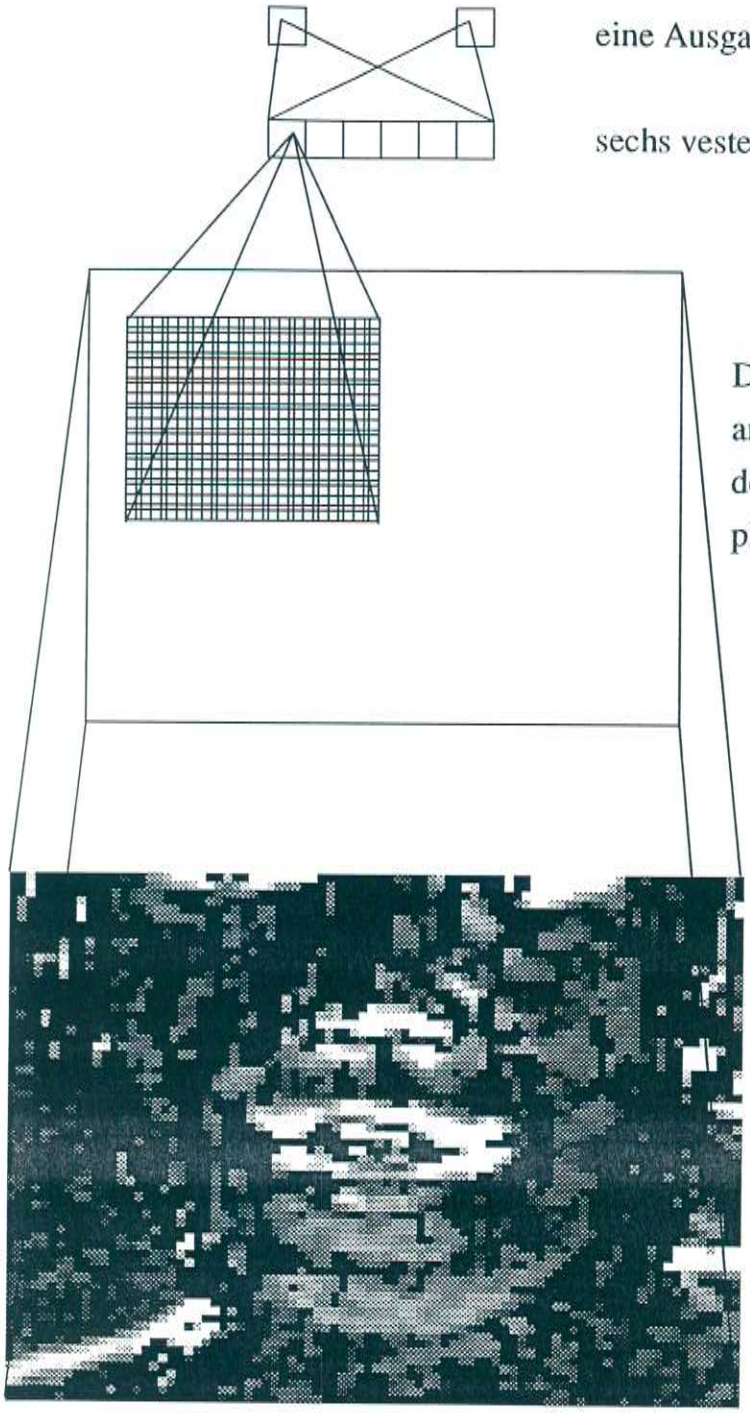

Abbildung 31: Architektur beim Verschieben eines neuronalen Netzes über das vorverarbeitete Ausgangsbild

Es erwies sich als wenig sinnvoll, dem Netz einfach alle möglichen Positionen des Eingabefensters hintereinander zu präsentieren. Da für fast alle dieser Positionen der Ausgabeeinheit ein Wert sehr nahe an 0 vorgegeben wurde, gerieten die trainierten Netze beim verwendeten Gradientenabstiegsverfahren sehr häufig in ein lokales Mini-

mum, bei dem sie für alle Positionen (auch wenn sich der gesuchte Mundwinkel in der Mitte des Eingabefensters befand) einen Wert nahe 0 ausgaben. Wesentlich bessere Ergebnisse wurden erzielt, wenn in jedem Trainingsbild mit Hilfe eines "Zufallsgenerators" eine begrenzte Zahl von "falschen" Positionen ausgewählt und zusammen mit "korrekten" Positionen dem Netz präsentiert wurden. Sowohl mit dem Verhältnis 1:1 als auch mit dem Verhältnis 1:3 von Positionen mit vorgegebenem Ausgabewert größer 0,5 gegenüber Positionen mit Ausgabewert kleiner 0,5 wurden gute Ergebnisse erreicht.

Analog zu den in Kapitel 6 beschriebenen Experimenten wurden verschiedene Vorverarbeitungen, Netzgrößen und Größen des Eingabefeldes getestet. Bei diesen Versuchen wurden die Netze mit 50 Bildern von 10 verschiedenen Personen trainiert und mit 212 Bildern von 10 anderen Personen getestet (Trainings- und Testmenge wie in Kapitel 6).

Beim Test wurde das bereits in Abschnitt 7.1 erwähnte Raster von 36×60 möglichen Positionen verwendet. Es wurde allerdings nur ein Viertel der Positionen (jede zweite Zeile und Spalte) getestet. Daher war die Auflösung bei der Schätzung der Position der Mundwinkel genauso groß wie bei dem in Kapitel 5 beschriebenen Netz.

Vorverarbeitungen

Getestete Vorverarbeitungen waren:

- normalisierter Grauwert
- horizontale Kanten mit 90^o Bandbreite
- horizontale und vertikale Kanten mit jeweils 90^o Bandbreite

Die einzelnen Vorverarbeitungen wurden wie in den vorhergehenden Kapiteln beschrieben durchgeführt. Wie beim Training erfolgte die Normierung über den gesamten Bereich, über den das Eingabefenster des Netzes bewegt wird. Beim Vergleich der Vorverarbeitungen wurde eine Architektur mit 6 versteckten Einheiten und einer Größe des Eingabefeldes von 24×30 Pixeln verwendet. Die Ergebnisse sind in Abbildung 32 zu sehen. Dabei ist wie in den entsprechenden Darstellungen in Kapitel 6 und den anderen graphischen Darstellungen des Erkennungsfehlers in diesem Abschnitt der mittlere Fehler von linkem und rechtem Mundwinkel addiert worden.

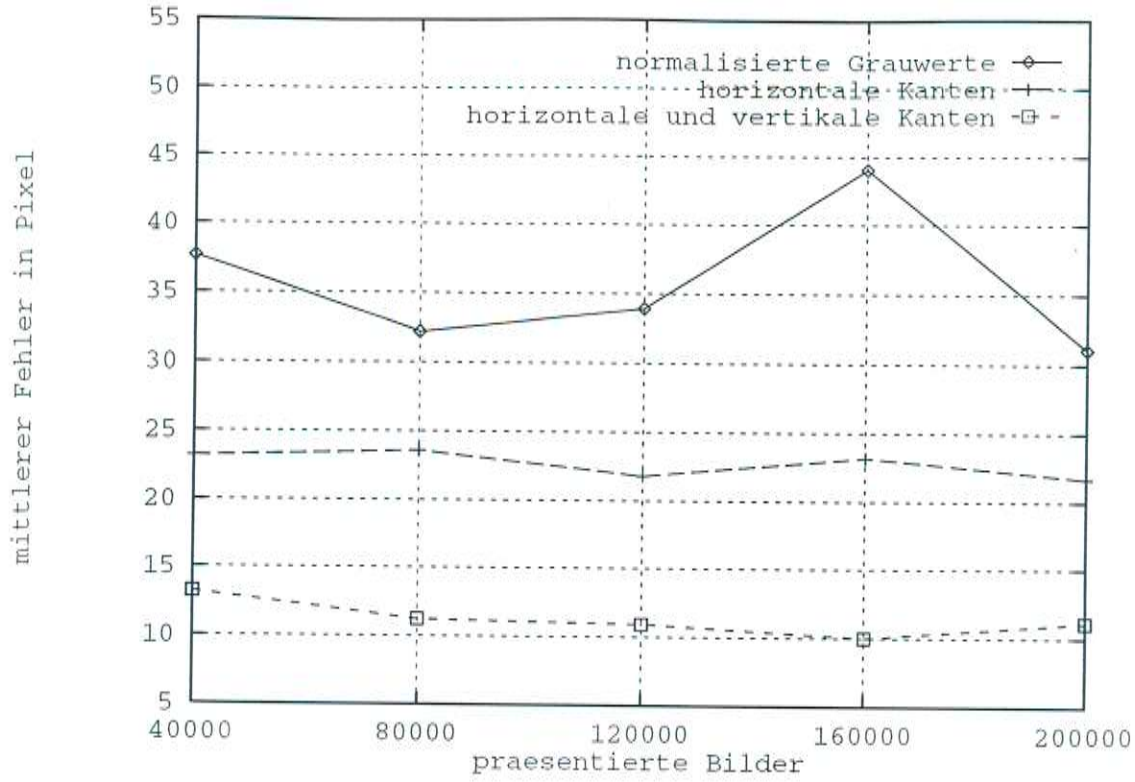

Abbildung 32: Erkennungsgenauigkeit für verschiedene Vorverarbeitungen

Da die Vorverarbeitung mit zwei Kantenrichtungen am Besten abschnitt, wurde sie in weiteren Experimenten ausschließlich eingesetzt.

Netzgröße

Hier wurde die Anzahl der versteckten Einheiten variiert. Die Ergebnisse sind in Abbildung 33 aufgeführt. In einem Experiment, dessen Ergebnisse nicht in der Abbildung zu sehen sind, wurde auch ein Netz mit nur einer versteckten Einheit (also praktisch ohne versteckte Schicht) für die Erkennung des linken Mundwinkels trainiert. Der dabei erreichte mittlere Fehler von 6,1 Pixel ist nicht dramatisch schlechter als der entsprechende mit drei versteckten Einheiten erreichte Wert von 4,7 Pixeln.

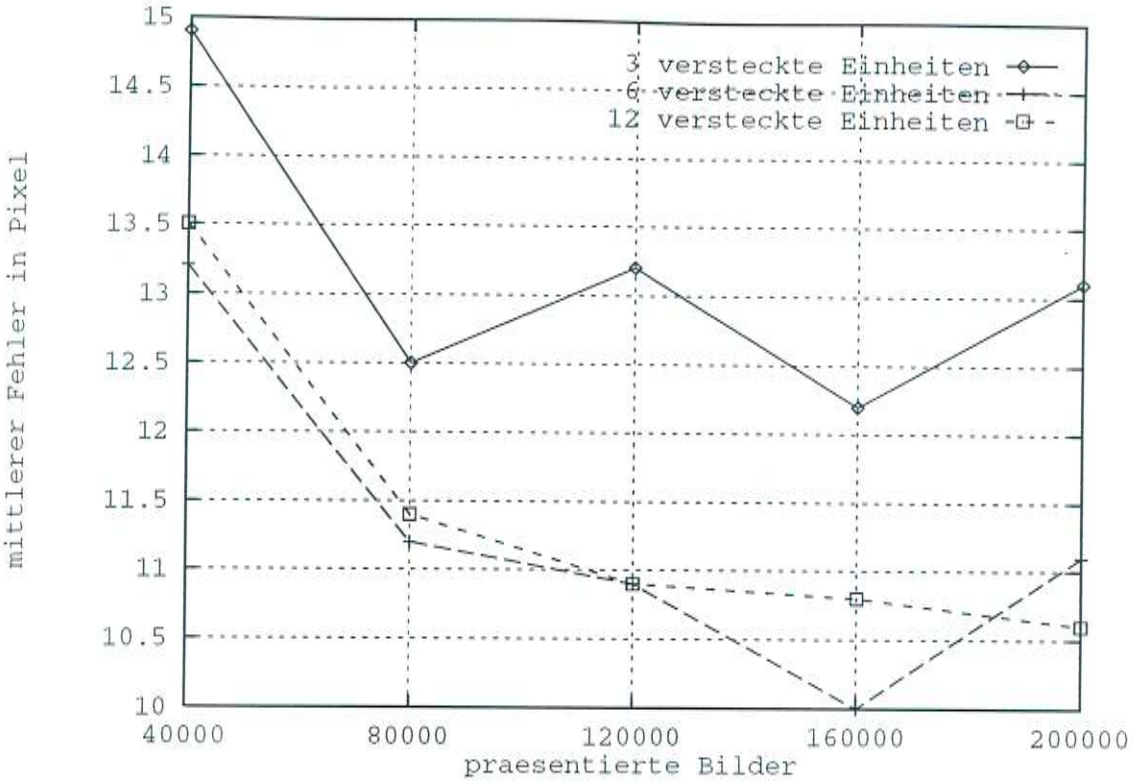

Abbildung 33: Erkennungsgenauigkeit für verschiedene Anzahlen versteckter Einheiten

Größe Eingabefeld

Es wurden drei verschiedene Größen von Eingabefeldern getestet. Die Ergebnisse sind in Abbildung 34 zu sehen.

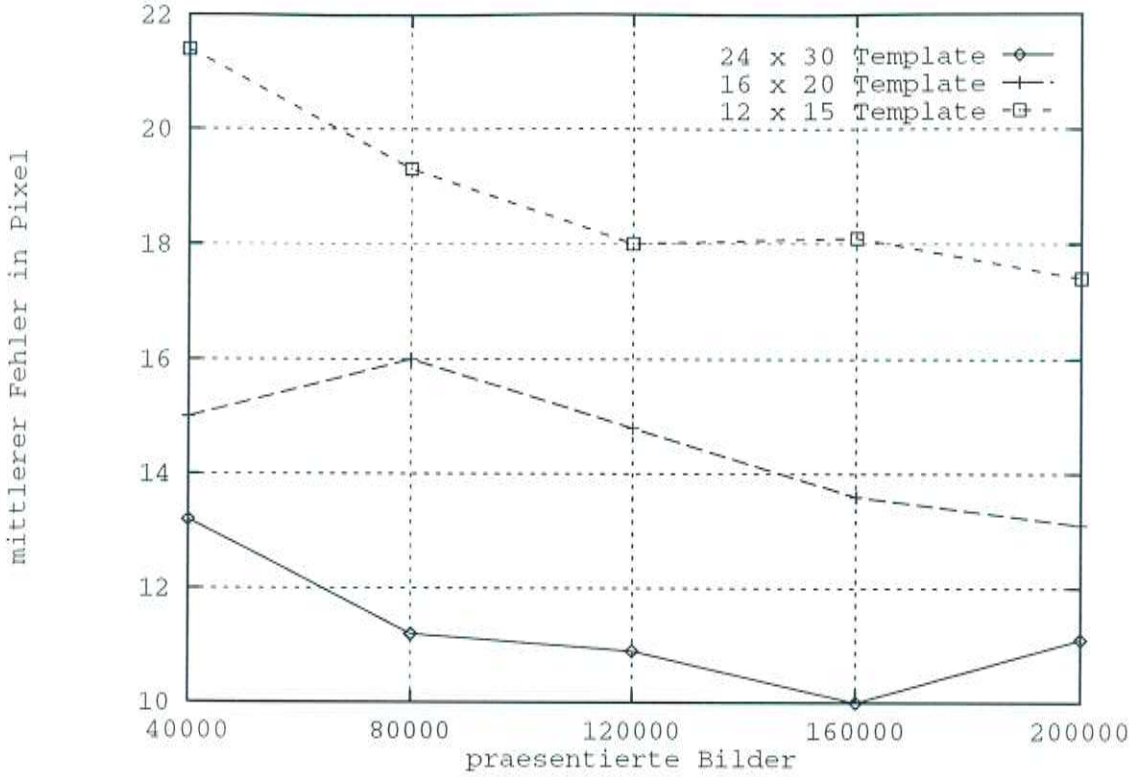

Abbildung 34: Erkennungsgenauigkeit für verschiedene Größen des Eingabefeldes

Für alle in diesem Abschnitt erwähnten Ergebnisse gilt, daß der mittlere Fehler sehr stark durch grobe Fehlerkennungen (outlier) beeinflußt wird. Um dies zu illustrieren, sind in den Tabellen 17 bis 20 die Fehlerverteilungen für den linken und rechten Mundwinkel bei Netzen mit der Vorverarbeitung in zwei gerichtete Kantenfelder, einer Größe des Eingabefeldes von 24×30 Pixeln und 12 bzw. 3 versteckten Einheiten aufgeführt. Für die höhere Fehlerrate beim rechten Mundwinkel habe ich keine plausible Erklärung. Tendenziell blickten die Versuchspersonen während den Aufnahmen häufig links an der Kamera vorbei, so daß der rechte Mundwinkel (alle Richtungsangaben aus Sicht der Kamera) eher besser erkennbar sein müßte.

7.3 Verringerung des Rechenaufwandes durch rezeptive Felder

Die Grundidee bei der in diesem Abschnitt beschriebenen Beschleunigung der Erkennung liegt in einer Aufteilung des Eingabefeldes in rezeptive Felder. Die berechneten Aktivierungen der den rezeptiven Feldern zugeordneten versteckten Einheiten können für verschiedene Positionen des Eingabefensters des neuronalen Netzes verwendet werden. Damit dies möglich ist, müssen die Gewichte zwischen Eingabeschicht und versteckten Einheiten für alle rezeptiven Felder des Eingabefeldes gleich sein. Eine solche Nutzung von rezeptiven Feldern mit gleichen Gewichten

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	44,8	50,5	25,5
2-3	36,3	26,4	36,8
4-5	12,7	16,5	25,5
6-9	3,8	3,3	7,6
10-13	1,9	1,4	2,4
14-18	0,0	0,9	0,8
> 18	0,5	0,9	1,4

Tabelle 17: Fehlerverteilung linker Mundwinkel; Netz mit 12 versteckten Einheiten;Durchschnittsfehler dx: 2,5 dy: 2,3 insgesamt: 3,9 Pixel

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	32,1	45,3	17,9
2-3	36,3	25,9	34,4
4-5	16.0	16,0	24,5
6-9	9,9	9,4	13,7
10-13	1,9	0,5	4,2
14-18	0,5	1,0	0,6
> 18	3,3	1,9	4,7

Tabelle 18: Fehlerverteilung rechter Mundwinkel; Netz mit 12 versteckten Einheiten;Durchschnittsfehler dx: 5,1 dy: 2,9 insgesamt: 6,6 Pixel

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	41,0	46,7	22,2
2-3	39,1	27,4	37,7
4-5	15,1	17,9	28,8
6-9	2,4	5,2	8,0
10-13	0,5	0,0	0,5
14-18	0,0	0,9	0,5
> 18	1,9	1,9	2,4

Tabelle 19: Fehlerverteilung linker Mundwinkel; Netz mit 3 versteckten Einheiten;Durchschnittsfehler dx: 3,3 dy: 2,6 insgesamt: 4,7 Pixel

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	28,8	37,7	14,6
2-3	31,6	26,9	29,2
4-5	16,0	20,3	24,5
6-9	16,5	10,9	22,2
10-13	0,9	0,0	2,8
14-18	1,4	1,0	1,0
> 18	4,7	3,3	5,7

Tabelle 20: Fehlerverteilung rechter Mundwinkel; Netz mit 3 versteckten Einheiten:Durchschnittsfehler dx: 5,7 dy: 3,6 insgesamt: 7,4 Pixel

wurde auch in der Spracherkennung (Waibel, 89) und der Erkennung von handgeschriebenen Ziffern und Buchstaben (Le Cun et al., 90) verwendet. Dort war dies allerdings nicht in erster Linie mit der erhöhten Effizienz der Berechnung begründet worden. Die Mehrfachverwendung der rezeptiven Felder wird durch Abbildung 35 veranschaulicht. Das dort gezeigte Eingabefeld hat die Größe 4×4 und besteht aus 4 rezeptiven Feldern der Größe 2×2. Das Eingabefeld ist in vier verschiedenen Positionen zu sehen. In jeder dieser Positionen wird das rezeptive Feld in der Mitte des 6×6 Eingabebildes verwendet. Da die Gewichte des rezeptiven Feldes unabhängig von seiner Position im Template sind, müssen die Aktivierungen der dem rezeptiven Feld zugeordneten versteckten Einheiten nur einmal berechnet werden.

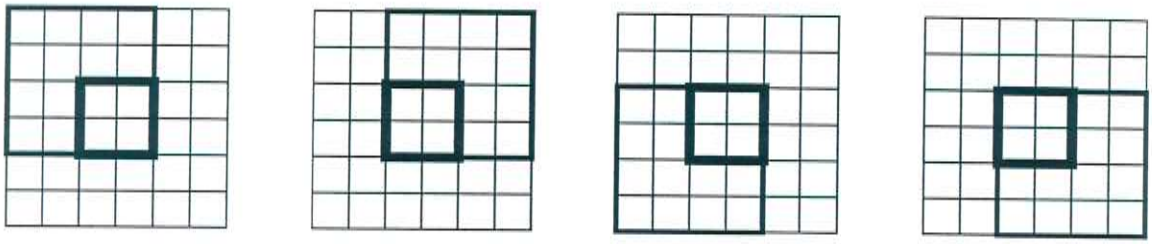

Abbildung 35: Mehrfachverwendung rezeptiver Felder

Dieses Verfahren wurde für die Templategröße 24×30 und eine Aufteilung in 3×3 rezeptive Felder, die jeweils die Größe 8×10 hatten, für verschiedene Anzahlen von versteckten Einheiten pro rezeptivem Feld getestet. Die Ergebnisse sind in Abbildung 36 aufgeführt. Wie ein Vergleich mit Abbildung 33 zeigt, beschleunigt die Verwendung von rezeptiven Feldern nicht nur die Erkennung, sie führt auch zu etwas besseren Ergebnissen.

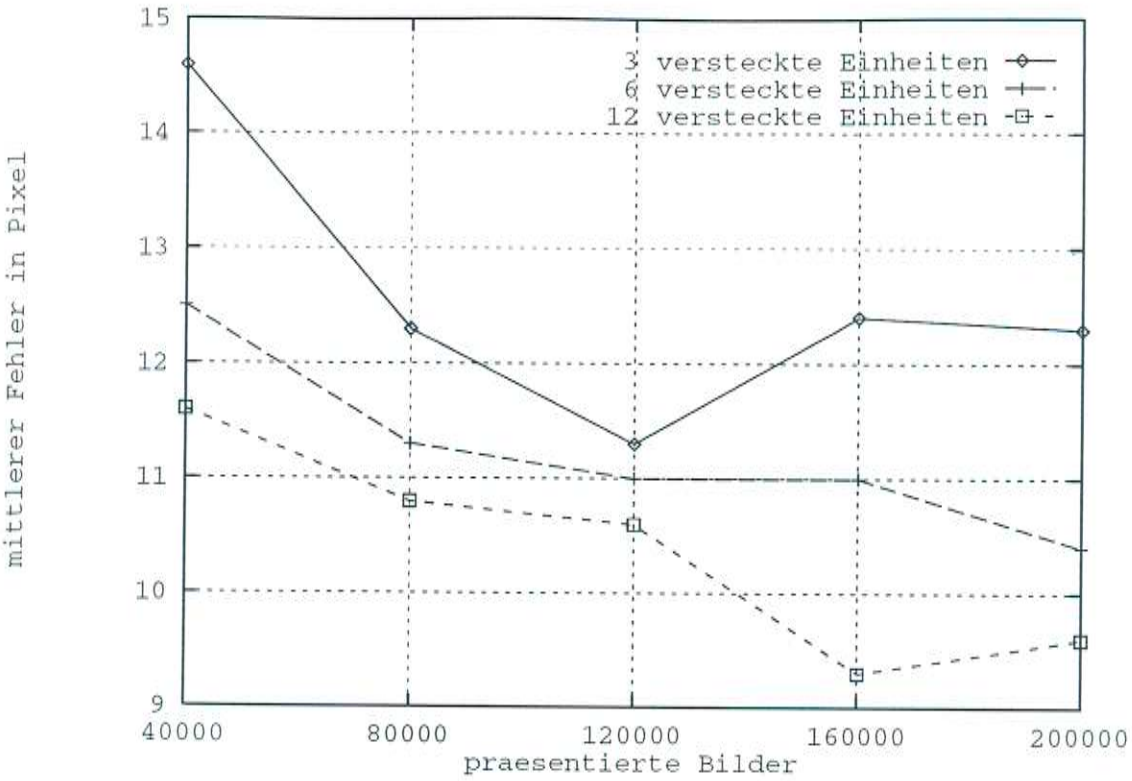

Abbildung 36: Erkennungsfehler bei der Verwendung rezeptiver Felder

Schließlich wurde für diese Architektur mit rezeptiven Feldern noch die Größe des Eingabefeldes auf 36×45 (mit 9 rezeptiven Feldern) gesteigert, um die Erkennungsleistungs möglicherweise noch weiter zu steigern. Wie man in Abbildung 37 sieht, ist aber keine wesentliche Verbesserung zu erkennen.

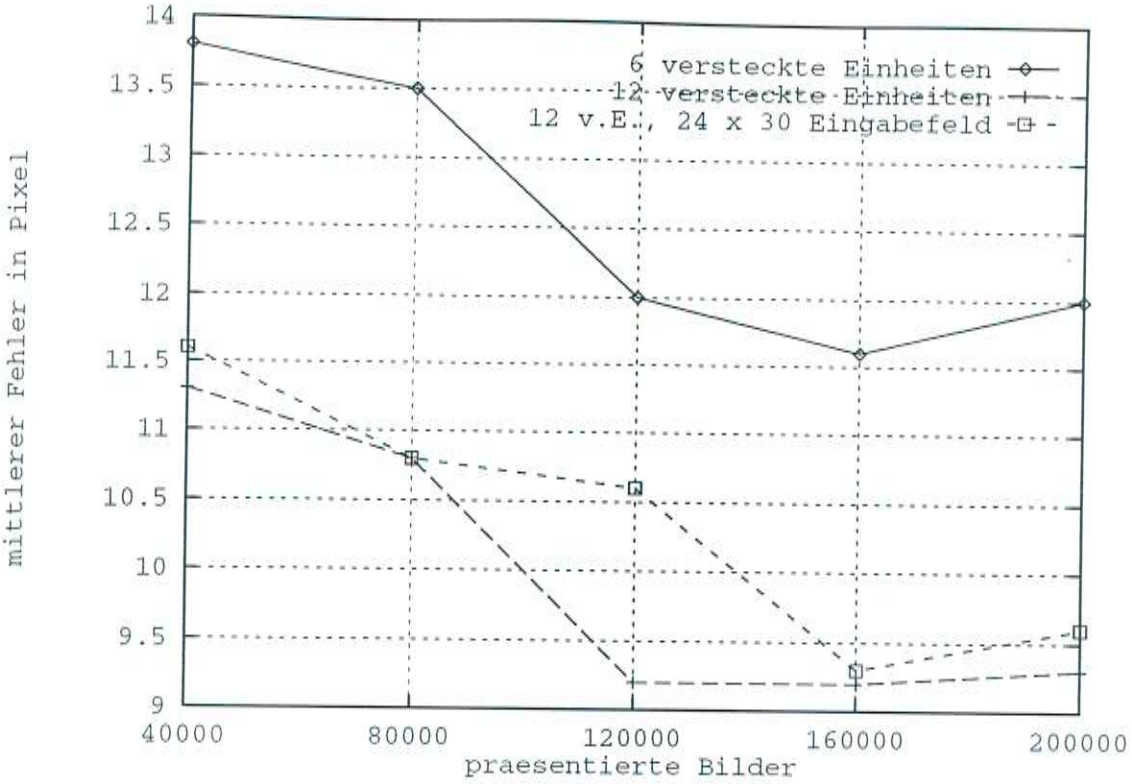

Abbildung 37: Erkennungsfehler bei der Verwendung rezeptiver Felder und einem 36×45 Eingabefeld

7.4 Vergleich der verschiedenen Architekturen zur Feinerkennung

Beim Vergleich der verschiedenen Verfahren zur Feinerkennung erscheinen folgende Kriterien wichtig zu sein:

- Qualität der Erkennung
- Geschwindigkeit
- Flexibilität gegenüber Änderungen des Suchbereichs
- Trainingszeit

Zum Vergleich der Qualität der Erkennungsgenauigkeit geben die Tabellen 22 und 21 Auskunft. Wie man sieht, sind bei der personenunabhängigen Erkennung die auf Kantendetektion und neuronalen Netzen basierenden Verfahren (mit Einpaßarchitektur und Templatearchitektur) besser als normalisierte Kreuzkorrelation. Bei der personenabhängigen Erkennung fällt der Vergleich nicht so eindeutig aus. Wie in den Tabellen 24 und 23 zu sehen ist, funktioniert die Feinerkennung mit normalsierter Kreuzkorrelation im Prinzip gut und wird lediglich durch eine Anzahl

Gesamtergebnis mit Feinerkennungsmethode	Summe mittlerer Fehler der beiden Mundwinkel
Neuronales Netz (Einpaßarchitektur)	15,0
Neuronales Netz (Templatearchitektur)	11,1
normalisierte Kreuzkorrelation (ein Template pro Mundwinkel)	34,0

Tabelle 21: Gesamtfehler bei der personenunabhängigen Erkennung

Gesamtergebnis mit Feinerkennungsmethode	Summe mittlerer Fehler der beiden Mundwinkel
Neuronales Netz (Einpaßarchitektur)	5,0
Neuronales Netz (Templatearchitektur)	6,6
normalisierte Kreuzkorrelation (ein Template pro Mundwinkel)	9,8

Tabelle 22: Gesamtfehler bei der personenabhängigen Erkennung

grober Fehler bei der Erkennung des linken Mundwinkels gestört. Beim Test der normalisierten Kreuzkorrelation wurden Durchschnittstemplates verwendet, mit denen bei den Versuchen von Hutchinson und Welsh (89) besonders gute Ergebnisse erzielt wurden. Dazu wurden jeweils fünf Ausschnitte um den jeweiligen Lippenwinkel von einer (personenabhängige Erkennung) und verschiedenen (personenunabhängige Erkennung) Personen gemittelt.

Dieser Unterschied in der Leistungsfähigkeit der Korrelation zwischen personenunabhängiger und personenabhängiger Erkennung wird vermutlich dadurch verursacht, daß bei der personenabhängigen Erkennung die Variationen zwischen den Grauwerten in der Mundregion verschiedener Bilder nur gering sind. Bei der personenunabhängigen Erkennung sind diese Unterschiede wesentlich größer. Der menschliche Beobachter hat jedoch trotzdem keine Probleme, den Mund wiederzuerkennen. Ich vermute, daß dies daran liegt, daß der Mensch bei der Objekterkennung komplizierte Features als einfache Grauwerte benutzt. Beispielsweise wurden im infereotemporalen Cortex von Makkakenaffen Zellen gefunden, die auf geometrische Merkmale mittlerer Komplexität (z.B. eine dunkle horizontal ausgerichtete Ellipse) reagieren (Tanaka, 93).

In dieser Arbeit schnitten bei der personenunabhängigen Erkennung Verfahren besser ab, die die etwas abstrakteren Merkmale der orientierten Kantenfelder nutzen. Diese Merkmale sind deshalb komplexer, weil sie Änderungen der Intensität unabhängig von deren Absolutwert anzeigen. Interessant ist, daß im primären Sehfeld von Primaten die meisten Zellen am besten auf Lichtbalken einer festen Orientierung

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	58,6	69,6	49,9
2-3	20,9	11,7	29,2
4-5	4,0	1,0	2,0
6-9	2,3	1,0	0,0
10-13	0,5	10,5	0,0
14-18	1,0	0,7	0,7
> 18	12,7	5,5	18,2

Tabelle 23: Fehlerverteilung linker Mundwinkel bei der personenabhängigen Erkennung mit normalisierter Kreuzkorrelation; Durchschnittswerte in Pixel dx: 5,1; dy: 3,5; gesamt: 6,9

Fehler in Pixeln	Prozent der Bilder mit Abstand		
	horizontal	vertikal	gesamt
0-1	20,2	78,1	16.0
2-3	63,1	20,7	67,1
4-5	15,5	0,5	15,0
6-9	0,7	0,2	1,5
10-13	0,0	0,0	0,0
14-18	0,0	0,3	0,0
> 18	0,5	0,2	0,5

Tabelle 24: Fehlerverteilung rechter Mundwinkel bei der personenabhängigen Erkennung mit normalisierter Kreuzkorrelation; Durchschnittswerte in Pixel dx: 2,6; dy: 1.1: gesamt 3,0

Erkennungsverfahren	CPU User Time in Sekunden
Groberkennung	0,28
Feinerkennung (Einpaß)	0,20
Feinerkennung Template 24×30 6 versteckte Einheiten ohne rez. Felder	2,64
dito mit 3×3 rezeptiven Feldern	0,65
Korrelation Template 24×30	1,4

Tabelle 25: Rechenzeiten für die verschiedenen Erkennungsverfahren auf einer DEC AXP 600

reagieren. Diese selektive Empfindlichkeit hat gewisse Ähnlichkeiten mit den hier verwendeten gerichteten Kantenfeldern.

Die Geschwindigkeiten der verschiedenen Verfahren ist ist Tabelle 25 zu sehen. Die Zeit für das Matching mit normalisierter Kreuzkorrelation bezieht sich auf ein Template (einen Lippenwinkel). Im Gegensatz zu anderen Verfahren wurde hier jedoch eine höhere Auflösung (alle Positionen statt nur jede vierte) verwendet. Für die Groberkennung mit Einpaßarchitektur wurde mehr Zeit benötigt als für die Feinerkennung mit Einpaßarchitektur, obwohl das verwendete Netz wesentlich kleiner ist. Der Hauptaufwand bei der Groberkennung liegt in der Verkleinerung des Bildes von 256×256 Pixel auf 32×32 Pixel.

Die Einpaßarchitektur zur Feinerkennung hat den Nachteil, daß sie recht unflexibel bzgl. der verwendeten Bildgröße ist. Wenn die Größe des Suchbereiches geändert werden soll, muß ein neues Netz mit anderen Anzahlen von Ein- und Ausgabeeinheiten trainiert werden. Der Suchbereich darf auch nicht zu groß sein, da durch die volle Verknüpfung von versteckter Schicht und Ausgabeschicht die Anzahl der Verbindungen zwischen diesen Schichten - wie in Abschnitt 6.6. beschrieben - quadratisch mit der Fläche des Suchbereichs steigt. Bei der templatebasierten Erkennung mit neuronalen Netzen ist bei einer Änderung des Suchbereichs ein neuer Traininglauf auf den im neuen Suchbereich vorkommenden Bildausschnitten sinnvoll aber nicht zwingend. Das Training der Einpaßarchitektur mit 10 versteckten Einheiten pro rezeptivem Feld auf einer DEC AXP 600, die nicht durch weitere rechenintensive Prozesse belastet ist, dauert ca. 2 Tage. Für das Training der templatebasierten Architektur wurde bei 12 versteckten Einheiten etwas mehr als eine Stunde benötigt.

Zusammenfassend ist nach den Ergebnissen der durchgeführten Experimente für die personenabhängigen Feinerkennung keines der drei Verfahren von der Erken-

nungsqualität her unbrauchbar. Die auf neuronalen Netzen basierenden Methoden liefern allerdings etwas bessere Ergebnisse. Da die Einpaßarchitektur um einiges schneller ist als die templatebasierte Architektur, würde ich sie für Online-Tests des Gesamtsystems zum Lippenlesen empfehlen. Bei der personenunabhängigen Erkennung weisen die mit Vorverarbeitung und neuronalen Netzen arbeitenden Verfahren deutlich bessere Erkennungsergebnisse als die normalisierte Kreuzkorrelation auf.

8 Ausblick

In der vorliegenden Diplomarbeit wurde versucht, mit dem relativ einfachen Werkzeug des mit Backpropagation trainierten mehrschichtigen Perzeptrons möglichst gute Resultate zu erzielen. Die Qualität der Erkennung wird durch die Vorverarbeitung (Detektion gerichteter Kantenfelder) sehr positiv beeinflußt. Meiner Meinung nach sind durch weitere Optimierungen von Vorverarbeitung oder Netzarchitektur nur noch geringe Verbesserungen zu erwarten. Der Vergleich von personenabhängiger mit personenunabhängiger Erkennung läßt es jedoch möglich erscheinen, daß durch eine Vergrößerung der Trainingsmenge die Erkennungsqualität bei der personenunabhängigen Erkennung noch gesteigert werden kann.

Daneben sehe ich noch zwei Hauptansatzpunkte für weitere Verbesserungen:

1. Reduktion der Anzahl grober Fehler (outlier) durch Nachverarbeitung der Netzausgaben

   Hier sind Verfahren unterschiedlicher Komplexität denkbar. Durch die Festlegung eines Maximalwertes für den vertikalen Abstand der Mundwinkel konnte beispielsweise in einem Test die Summe der mittleren Fehler für die beiden Mundwinkel von 12,8 auf 10,3 Pixel gesenkt werden. Bei 11 von 212 Bildern wurde der Maximalwert für den vertikalen Abstand der Mundwinkel überschritten und die Schätzung verworfen. Es wäre aber auch möglich, wie in Vincent et al. (91) neben den Positionen mit maximalen Netzausgaben auch andere Positionen mit hohen Ausgaben zu betrachten und dann eine Schätzung für die Position der Mundwinkel auszuwählen, die durch hohe Netzausgaben nahegelegt wird und geometrisch plausibel ist.

2. Erhöhung der Genauigkeit durch einen weiteren Erkennungsschritt mit feinerer Auflösung der Ausgabeeinheiten

   Bei der Verwendung von Bildern der Größe 256×256 Pixel beträgt der Gitterabstand bei der in dieser Arbeit beschriebenen Feinerkennung mit neuronalen Netzen (bei der Einpaßarchitektur und der templatebasierten Architektur) 3,75 Pixel. Fehler von knapp zwei Pixel in x- und y-Richtung sind also schon aufgrund der Architektur unvermeidlich. Diese Fehler könnten durch eine "Ultrafeinerkennung" in der unmittelbaren Umgebung der von der Feinerkennung geschätzten Position der Lippenwinkel verringert werden. Dabei wäre dann ein Abstand der Ausgabeeinheiten von einem Bildpixel vom Rechenaufwand her kein Problem. Möglicherweise ist aber eine Genauigkeit von einem Pixel deshalb nicht erreichbar, weil es beim Labeln der Trainingsbilder unmöglich ist, die Position der Lippenwinkel eindeutig auf den Pixel genau zu erkennen. Weiterhin ist bei weiten Öffnungsgraden der Mund relativ kreisförmig und eine genaue Erkennung der Lippenwinkel (vor allem bei Variationen der Kopfneigung) zusätzlich erschwert.

9 Literatur

Ballard, P., Stockman, G. C., 1992. Computer operation via face orientation. In Proc. International Conference on Pattern Recognition, 1992, 407-410.

Beymer, D. J., 1993. Face regonition under varying pose. Massachusetts Institute of Technology, Artificial Intelligence Laboratory, A.I. Memo No. 1461.

Bichsel, M., Pentland, A. P., 1994. Human face recognition and the face image set's topology. Vol. 59, No. 2, March 94, 254-261.(CVGIP - Image Understanding)

Bregler, C., Hild, H., Manke, S., Waibel, A., 1993. Improving connected letter recognition by lipreading. In Proc. ICASSP 1993.

Brunelli, R., Poggio, T., 1993. Face recognition: features versus templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15., No. 10, 1042-1052.

Cootes, T. F., Taylor, C. J., Lanitis, A., Cooper, D. H., Graham, J.,1993. Building and using flexible models incorporating grey-level information. In Proc. Fourth International Conference on Computer Vision, Berlin, 1993.

Craw, I., Tock, D., Bennet, A., 1992. Finding face features. In Proc. European Conference on Computer Vision, 1992, 92-96.

Duchnowski, P., Meier, U., Waibel, A., 1994. See me, hear me: integrated automatic speech recognition and lipreading. In Proc. ICSLP 94.

Garcia, O. N., Goldschen, A. J., Petajan, E. D., 1992. Feature extraction for optical automatic speech recognition or automatic lipreading. George Washington University Technical Report GWU-IIST-92-32, Washington, D.C. 1992.

Hertz, J., Krogh, A., Palmer, R. G., 1991. Introduction to the theory of neural computation, Addison-Wesley, 1991.

Hunke, M., 1994. Locating and tracking of human faces with neural networks. Technical Report CMU-CS-94-155, Carnegie Mellon University 1994.

Hutchinson, R. A., Welsh, W. J. 1989. Comparison of neural networks and conventional techniques for feature location in facial images. First IEE International Conference on Artificial Neural Networks, London 1989.

Kitchen, L. J., Malin, J. A. 1989. The effect of spatial discretization on the magnitude and direction response of simple differential edge operators on a step edge. Computer Vision, Graphics and Image Processing 47, 243-258.

Le Cun, Y., Boser, B., Denker, J. S., Solla, S., Howard, R., Jackel, L., 1990. Back-Propagation applied to handwritten zipcode recognition. Neural Computation 1 (4), 541-551.

Reisfeld, D., Wolfson, H., Yeshurun, Y., 1990. Detection of interest points using symmetry. In Proc. Third International Conference on Computer Vision, 1990, 62-65.

Reisfeld D., Yeshurun, Y., 1992. Robust detection of facial features by generalized symmetry. In Proc. International Conference on Pattern Recognition, 1992, 117-120.

Rumelhart, D. E., McClelland, J. L. and the PDP Research Group, 1986. Parallel Distributed Processing: Exploration in the Microstructure of Cognition, 2 vol. Cambridge: MIT Press.

Stork, D. G., Wolff, G., Levine, E., 1992. Neural network lipreading system for improved speech recognition. In Proc. IJCNN 92, Baltimore.

Tanaka, K. 1993. Neuronal mechanisms of object recognition. Science 262, 685-688.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of Cognitive Neuroscience, Vol. 3, Nr. 1, 71-86.

Vincent, J. M., Waite, J. B., Myers, D. J., 1991. Precise location of facial features by a hierarchical assembly of neural nets. In Proc. Second IEE International Conference on Artificial Neural Networks, 1991.

Waibel, A., 1989. Modular construction of time-delay neural networks for speech recognition. Neural Computation, 1 (1), 39.

Yuille, A. L., Hallinan, P. W., Cohen, D. S., 1992. Feature extraction from faces using deformable templates. International Journal of Computer Vision, 8:2, 99-111.