Methoden der vision-basierten Nutzerwahrnehmung für eine natürliche Interaktion mit mobilen Servicerobotern

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)
vorgelegt der
Fakultät für Informatik und Automatisierung
der Technische Universität Ilmenau

von
Torsten Wilhelm

Gutachter:

Univ.-Prof. Dr.-Ing. Horst-Michael Groß	Technische Universität Ilmenau
Univ.-Prof. Dr.-Ing. Gerhard Rigoll	Technische Universität München
Prof. Dr.-Ing. M. Sc. Thomas Zielke	Fachhochschule Düsseldorf

Tag der Einreichung:	28. Juni 2005
Tag der wissenschaftlichen Aussprache:	8. Dezember 2005

urn:nbn:de:gbv:ilm1-2005000277

Zusammenfassung

Im Gegensatz zur zwischenmenschlichen Kommunikation, bei der die Beziehungsebene im Vergleich zur Sachebene den weitaus größeren Anteil einnimmt, wird diese bei der Mensch-Roboter-Interaktion bislang nur in Ansätzen berücksichtigt. Insbesondere die Nutzerwahrnehmung bleibt in der Regel auf eine reine Personendetektion oder ein einfaches Personen-Tracking beschränkt. Vor diesem Hintergrund wurde eine verbesserte Wahrnehmung des aktuellen Zustandes des Nutzers als Voraussetzung für eine Personalisierung des Dialogs als Zielstellung dieser Arbeit abgeleitet. Beim exemplarischen Anwendungsszenario handelt es sich um einen Shopping-Assistenten, der in einem Baumarkt den Kunden bei der Suche nach Produkten behilflich ist. Dieser sollte zumindest einen gewissen Grad an sozialer Kompetenz zeigen, indem er z.B. Personen in seiner Umgebung detektiert und während der Interaktion kontinuierlich Blickkontakt hält. Um Nutzermodelle erstellen, kurzzeitig verlorene Nutzer wiedererkennen und den Gemütszustand des Nutzers abschätzen zu können, sollen Geschlecht, Alter, Identität und Gesichtsausdruck des Nutzers aus einem Videobild ermittelt werden.

Für die Realisierung dieser Aufgabe wurde eine biologisch motivierte Aufteilung in ein peripheres und ein foveales Vision-System vorgeschlagen. Das periphere System arbeitet auf den Bildern einer omnidirektionalen Kamera und verfügt damit über einen sehr großen Sichtbereich, aber nur eine vergleichsweise geringe Auflösung. In diesem System werden zunächst Hypothesen über die Position von Personen im Umfeld des Roboters gebildet. Dafür werden Hautfarbe, Bewegung und Entfernung in einer Auffälligkeitskarte integriert und auffällige Bildbereiche mittels eines Multi-Target-Trackers verfolgt. Für die omnidirektionale Kamera wurde ein automatischer Weißabgleich entwickelt, der die Hautfarbdetektion unempfindlich gegen Änderungen der Chrominanz der Beleuchtung macht.

Nach Auswahl einer Nutzerhypothese wird der Kopf des Roboters kontinuierlich in die entsprechende Richtung ausgerichtet. Damit erhält der Nutzer zum einen eine Rückmeldung über die gerichtete Aufmerksamkeit des Roboters während der Interaktion. Zum anderen kann der Roboter hochaufgelöste Bilder der Person aufnehmen, so dass eine weitere nachfolgende Analyse ermöglicht wird. Diese ist wiederum in zwei Teilschritte unterteilt. Der erste Schritt besteht aus einer Detektion des Gesichtes und einer anschließenden Detektion der Augen, anhand derer eine normalisierte Darstellung des Gesichtes erzeugt wird. Für den Analyseschritt wurden das Elastic-Graph-Matching, die Independent Component Analysis und die Active-Appearance Models implementiert und vergleichend untersucht. Unter Berücksichtigung der Anforderungen einer Geschlechts-, Alters-, Mimik- und Identitätsschätzung wurde hierfür eine umfassende Gesichtsdatenbank zum Training und zum Test der Verfahren angelegt. Die Leistungsfähigkeit des Gesamtsystems wurde schließlich anhand von empirischen Experimenten demonstriert.

Abstract

In man-machine communication, particularly in the field of service robotics, the perception of the user is often constricted to people detection and tracking. This is in strong contrast to communication between people, where social information like gender, age, identity and facial expression is essential. The assumption of this thesis is that an improved perception of the user's state is necessary for future service robots to be successfully deployed in human centered service tasks. The example application is a service robot helping customers in a home store to find the desired products. During interaction, the robot should show a certain degree of social competence, e.g. by detecting persons and establishing and keeping eye contact. Furthermore, it should be able to build user models, identify known users robustly and estimate their affections by determining gender, age, identity and facial expression from video images.

To realize this functionality, a biologically motivated separation into a peripheral and a foveal vision system is proposed. The former uses images of an omnidirectional camera with a large field of view but relatively low resolution to generate hypotheses of the position of potential users in the surroundings of the robot. Therefore, skin color and movement as well as the measurements of sonar sensors are integrated into a saliency map. Salient structures are tracked by a multi target tracking system based on the CONDENSATION algorithm. To realize a skin color detection which is insensitive to changes of the illumination chrominance, an automatic white balance algorithm was developed which takes advantage of the special geometry of the omnidirectional objective.

After selecting a hypothesis, the head of the robot is continously directed in its direction. In this way, the user receives a feedback signal of the robots attention, while the robot is able to capture high resolution images of the users face suitable for a further two step analysis. The first step produces a normalized view of the users face by detecting the face and the eyes and applying affine image transformations. For the analysis itself, three methods were implemented and tested: Elastic Graph Matching, Independent Component Analysis and Active Appearance Models. With respect to the estimation of gender, age, facial expression and identity a comprehensive face image database was recorded for training and testing the different methods. The efficiency of the integrated system was demonstrated by empirical experiments.

Danksagung

Mein herzlicher Dank gilt allen, die zum Entstehen dieser Arbeit beigetragen haben.

An erster Stelle möchte ich meinem Betreuer und Leiter des Fachgebiets Neuroinformatik Prof. Dr. Horst-Michael Groß für die Möglichkeit danken, diese Arbeit unter seiner Regie in einem sehr herzlichen Arbeitsklima erstellen zu können. Die Geduld, die er und vor allem Herr PD Dr. Hans-Joachim Böhme aufgebracht haben, um mich von meinen zahlreichen Irrwegen auf den richtigen Pfad zurückzuführen, war bewundernswert. Die wertvollen Anregungen in unseren fachlichen Diskussionen haben mir geholfen, den roten Faden zu finden und dabei das große Ganze nicht aus dem Blick zu verlieren.

Mein herzlicher Dank gilt auch allen anderen Mitarbeitern des Fachgebiets, die mich bei meiner Arbeit unterstützt haben, ob durch die Bereitstellung der nötigen Arbeitsvoraussetzungen oder einfach nur ein aufmunterndes Wort zwischendurch: Klaus Debes, Ute Schütz, Heike Groß, Christof Schröter, Carsten Schauer, Alexander König, Steffen Müller, Christian Martin, Andrea Scheidig und Volker Stefan. Sabine, bei Dir muss ich mich fast schon entschuldigen.

Weiterhin möchte ich mich bei allen Studenten bedanken, die durch ihre Unterstützung zum Gelingen dieser Arbeit beigetragen haben, als da wären Rene Eckard, Norman Trapp, Alexander Bendlin, Andreas Backhaus, Jesko Ehrich, Jana Kludas und Norman Apel.

Mein ganz besonderer Dank gilt meinen Eltern Egon und Rosel Wilhelm. Ich hab es Euch wirklich oft nicht leicht gemacht.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Serviceroboter

Die Entwicklung von Servicerobotern hat in den letzten Jahren wesentlich an Bedeutung gewonnen. Obwohl diese Forschungsrichtung noch in den Kinderschuhen steckt und Anwendungen bis jetzt hauptsächlich demonstrativen Charakter haben, konnte der Einsatz von einigen Modellen auch schon in realen Einsatzfeldern gezeigt werden. Einige Beispiele sind [Burgard et al., 1998] [Wullschleger and Brega, 2001] und [Bischoff and Graefe, 2002]. In den letzten Jahren sind die ersten kommerziellen Systeme auf den Markt gekommen [King and Weiman, 1990].

Denkbare Anwendungsfälle für Serviceroboter sind:

Transportsysteme und *Produktionsassistenten*, die in Fabrik- und Montagehallen Transportaufgaben und einfache Montageaufgaben übernehmen und dabei nicht wie viele aktuelle Systeme an eine besondere Infrastruktur gebunden sind,

Tourguides, die autonom z.B. in Museen navigieren, Besucher zu Ausstellungsstücken führen und Erklärungen zu diesen abgeben,

Informationssysteme, die z.B. in Flughäfen oder anderen öffentlichen Einrichtungen anders als stationäre Informations-Terminals aktiv auf Besucher zugehen können und Informationen bereitstellen,

Einkaufsassistenten, die in Märkten aktiv auf Kunden zugehen können, diese ansprechen und ihnen bei der Suche nach bestimmten Produkten behilflich sind, Informationen zu Produkten bereitstellen und gegebenenfalls auch eine Verbindung zu einem Fachberater herstellen.

Haushaltshilfen, die als persönliche Diener fungieren, einfache Arbeiten übernehmen und ihre Nutzer mit alltäglichen Informationen wie dem Wetterbericht oder dem Fernsehprogramm versorgen, die sie aktuell aus dem Internet beziehen,

Pflegeroboter, die in Haushalten oder Pflegeheimen bei der Versorgung älterer oder hilfsbedürftiger Menschen eingesetzt werden, Erinnerungsfunktionen für die Einnahme von Medika-

menten übernehmen und unter Umständen auch Kontakt zu einem Arzt aufnehmen, falls eine Notsituation erkannt wird.

Unter dem Begriff Serviceroboter sind also autonome mobile Systeme zu verstehen, die in einem eingeschränkten Einsatzfeld bestimmte Dienstleistung für den Menschen erbringen. Durch diese Definition ergeben sich zwei Eigenschaften, die einen Serviceroboter charakterisieren:

Mobilität: Im Unterschied zu Industrierobotern oder Informations-Terminals sind Serviceroboter mobil, d.h. sie verfügen über die notwendige Aktuatorik, um sich in ihrem Einsatzgebiet fortbewegen zu können. Dabei sind sie in der Lage, ihre Umgebung selbständig zu kartieren. Sie können sich mittels der erstellten Karte selbst lokalisieren und Zielpunkte in ihrem Einsatzfeld ansteuern.

Service: Ein Serviceroboter unterstützt den Menschen, indem er bestimmte Dienstleistungen für ihn erbringt.

Im ersten Punkt wird der Frage nachgegangen, wie der Roboter seine Umgebung repräsentieren kann, so dass dieser jederzeit in der Lage ist, seinen eigenen Standort zu bestimmen und einen Pfad zu einem beliebigen anderen Standort zu planen [Schröter et al., 2004] [Gross and König, 2004]. Der zweite Punkt besagt, dass der Serviceroboter eine Dienstleistung für den oder mit dem Menschen erbringen soll. Diese Dienstleistungen können je nach Einsatzfeld sehr unterschiedlich sein. Auf jeden Fall aber, muss der Roboter in der Lage sein, die Wünsche seines Nutzers zu verstehen und ihm das Ergebnis seiner Arbeit in geeigneter Form zu präsentieren.

Die große Vielfalt an möglichen Dienstleistungen bedingt in ihrer Komplexität sehr unterschiedliche Mensch-Maschine-Schnittstellen. Ein autonomer Staubsauger z.B. benötigt nicht viel mehr als einen Ein- und Ausschalter, da seine Aufgabe lediglich darin besteht, den Boden eines Raumes autonom und möglichst gleichmäßig zu befahren. Ein Tourguide in einem Museum muss dagegen auch und vor allem Menschen wahrnehmen und mit ihnen kommunizieren können. Man sollte ihm mitteilen können, welche Ausstellungsstücke man sehen will, und er sollte in der Lage sein, diese sprachlich und eventuell graphisch zu erklären. Eine Haushaltshilfe oder ein Pflegeroboter muss nicht nur Menschen, sondern auch Objekte wahrnehmen und unterscheiden können und letztere gegebenenfalls auch mit Hilfe geeigneter Aktuatorik greifen, befördern und manipulieren können. In diesem Zusammenhang sollte der Roboter auch Zeigegesten verstehen können, damit eine Bezugsperson auf Objekte verweisen kann.

Um die in dieser Arbeit entwickelten Komponenten der Mensch-Maschine-Kommunikation zu motivieren, wird im nächsten Abschnitt das Einsatzszenario unseres Serviceroboters Perses vorgestellt.

1.2 Baumarktszenario

In dem untersuchten Szenario übernimmt der Roboter PERSES die Aufgabe eines Einkaufsassistenten in einem Baumarkt. Dieser soll in der Lage sein, mit einem Baumarktkunden Kontakt aufzunehmen, dessen Einkaufswünsche herauszufinden und ihn zu den gewünschten Produkten zu lotsen [Gross et al., 2000]. Neben den Problemen der Navigation, Selbstlokalisation und Pfadplanung in einer unstrukturierten Umgebung ergibt sich aus dem Einsatzszenario die Notwendigkeit einer intuitiven und natürlichen Kommunikation mit Nutzern.

Das Aufgabenspektrum während eines Einkaufs mit dem Serviceroboter ist in Abbildung 1.1 dargestellt. Dieses beginnt damit, dass der Roboter potentielle Kunden wahrnehmen und ansprechen muss, um sie zu einer Interaktion zu motivieren. Während des gesamten Interaktionszyklus, des Dialogs und der Lotsenfahrt, muss der Kunde kontinuierlich verfolgt werden. Für eine spätere Wiedererkennung soll die Identität des aktuellen Kunden als Referenz gespeichert werden. Falls der Roboter während der Interaktion kurzfristig den Kontakt zu diesem Nutzer verliert, soll der Roboter in der Lage sein, für eine Person zu entscheiden, ob es sich um den letzten Kunden handelt oder nicht. Auf diese Weise kann der Dialog fortgesetzt werden, falls der verlorene Kunde wiedergefunden wird. Das Alter und das Geschlecht des Kunden sind von Interesse, weil diese unter Umständen Einfluss auf die Gestaltung des Dialogs haben können. So ist es denkbar, dass ältere Personen anders angesprochen werden als jüngere oder dass je nach Geschlecht unterschiedliche Sonderangebote angezeigt werden. Mit Hilfe des Alters und des Geschlechts könnten auch Nutzerprofile erstellt werden. So könnte das Verhalten bei der Interaktion oder auch das Kaufverhalten je nach Alter und Geschlecht in bestimmte Kategorien eingeteilt werden und der Roboter könnte sich bei Kenntnis dieser Informationen entsprechend darauf einstellen und das Serviceangebot unterschiedlich präsentieren. Besondere Bedeutung kommt dem Gesichtsausdruck zu, da dieser Rückschlüsse auf den emotionalen Zustand des Nutzers zulässt und somit darauf, ob er mit dem Ablauf des Dialogs bzw. der durch den Roboter bereitgestellten Dienstleistung zufrieden ist oder nicht [Gross and Boehme, 2000].

Das Baumarktszenario stellt hinsichtlich der Mensch-Maschine-Schnittstelle große Ansprüche, da der Roboter mit Personen kommunizieren muss, die nicht von vornherein ein großes Interesse an der Nutzung eines Serviceroboters haben, wie dies z.B. bei einem Tourguide in einem technischen Museum der Fall wäre. Baumarktkunden decken nicht nur das volle Altersspektrum, sondern auch alle möglichen sozialen Schichten ab und bilden somit eine höchst heterogene Personengruppe. Es stellt sich also sofort die Frage, wie die Mensch-Maschine-Kommunikation beschaffen sein muss, damit der Roboter von den Kunden des Baumarktes akzeptiert und genutzt wird. Um diese Frage zu klären, wird im nächsten Abschnitt ein Blick auf die Kommunikation zwischen Menschen geworfen.

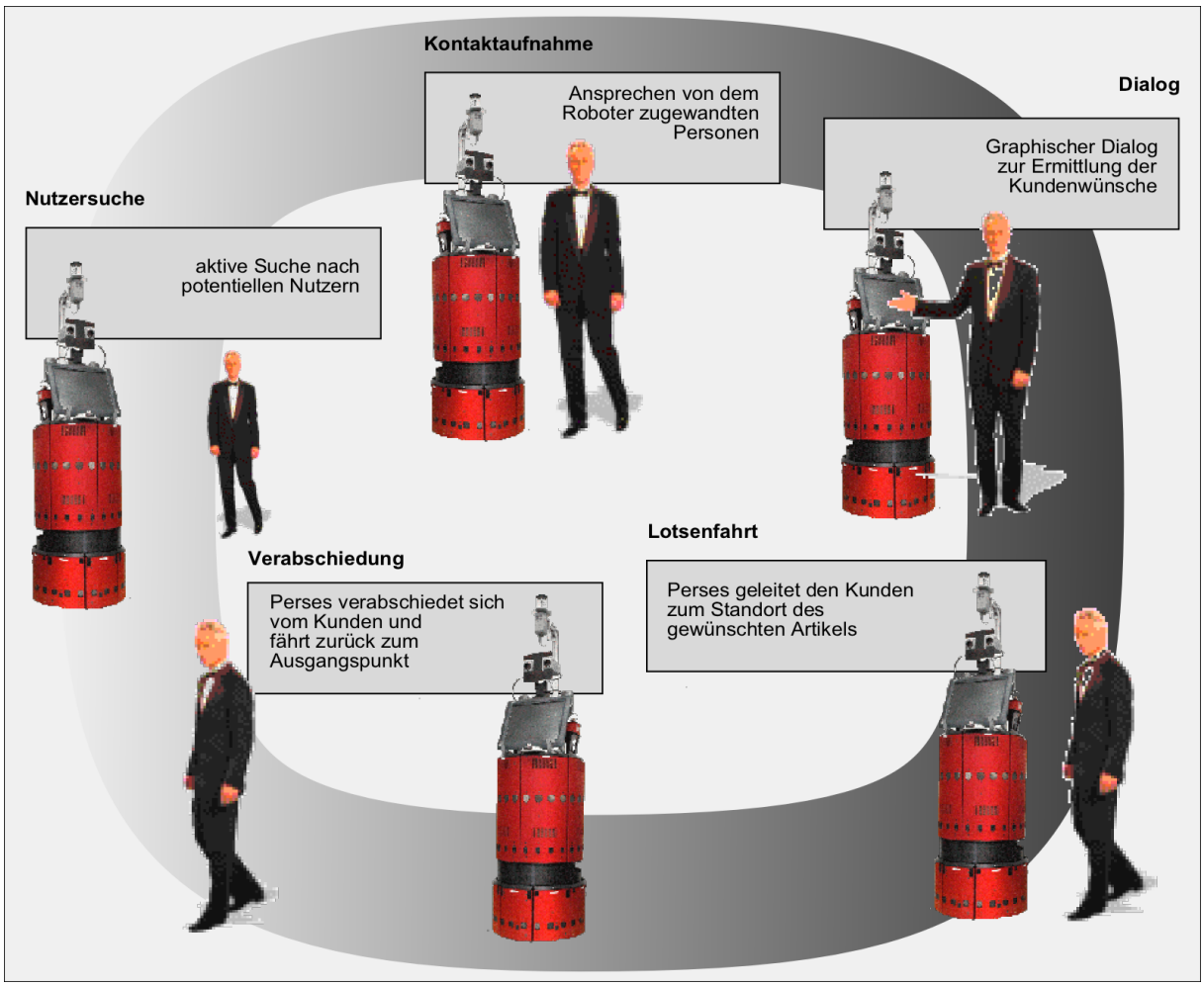


Abbildung 1.1: *Ein möglicher Interaktionszyklus mit dem Einkaufsassistenten* PERSES. *Zunächst wartet der Roboter im Eingangsbereich des Baumarktes auf potentielle Kunden. Sobald er einen solchen wahrnimmt, wendet er sich ihm zu und macht auf sich aufmerksam. Wenn der Kunde den Roboter anschaut und somit weiteres Interesse bekundet, wird er von* PERSES *begrüßt und dieser stellt sich ihm vor. Der eigentliche Dialog beginnt mit einer Erklärung der Funktionsweise des Roboters und hat das Ziel, die Einkaufswünsche des Kunden zu ermitteln. Dazu kann dieser auf verschiedene Weisen in einem graphischen Dialogsystem navigieren oder sich über eine Funkverbindung von einem telepräsenten Fachberater unterstützen lassen. In einer Lotsenfahrt geleitet der Roboter den Kunden schließlich zum gewünschten Produkt. Dialog und Lotsenfahrt können mehrmals wiederholt werden, bis der Kunde alle Artikel gefunden hat. Nachdem sich der Kunde abgemeldet hat, wird er von* PERSES *verabschiedet.*

1.3 Menschliche Kommunikation

Um Verfahren für die Mensch-Maschine-Kommunikation zu entwickeln, sollen in diesem Abschnitt die wesentlichen Grundzüge menschlicher Kommunikation aufgezeigt werden.

In [Watzlawick et al., 1996] wird eine Unterscheidung der Kommunikation in Sach- und Beziehungsebene getroffen. Die Sachebene, oder auch der Inhaltsaspekt, bezieht sich auf die klare und verständliche Mitteilung von Sachverhalten, also darauf, *worüber* informiert wird. Die Beziehungsebene dagegen bezieht sich darauf, wie Mitmenschen durch die Art und Weise der Kommunikation beeinflusst werden. Sie drückt aus, was man vom Kommunikationspartner hält und wie man zu ihm steht. Im allgemeinen wird bei zwischenmenschlicher Kommunikation zwischen verbaler und nonverbaler Kommunikation unterschieden, wobei erstere durch ihre logische Syntax eher auf der Sachebene und letztere vornehmlich auf der Beziehungsebene anzusiedeln ist. In Abbildung 1.2 werden die Komponenten menschlicher Kommunikation gegenübergestellt.

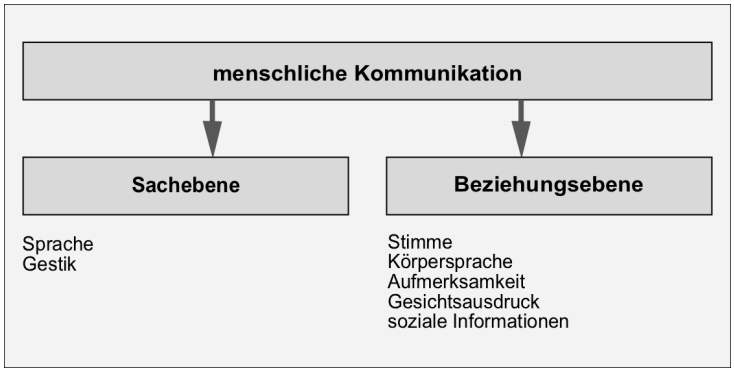


Abbildung 1.2: *Modalitäten menschlicher Kommunikation. Laut [Watzlawick et al., 1996] kann bei Kommunikation zwischen der Sach- und der Beziehungsebene unterschieden werden. Im Normalfall spielen diese Modalitäten sowohl auf Eingabe- als auch auf Ausgabeseite eine Rolle.*

Im folgenden wird auf die Bedeutung der einzelnen Modalitäten näher eingegangen.

Sprache: Sprache ist das komplexeste und vielseitigste Verständigungsmittel zwischen Menschen. Durch Sprache werden die konkreten zu vermittelnden Informationen ausgetauscht.

Gestik: Die Begriffe Gestik und Körpersprache werden oft synonym verwendet. Auf der Sachebene soll hier von Gestik gesprochen werden, die in Form von Zeigegesten oder Gebärdensprache ähnlich wie die Sprache zum Austausch von Inhalten verwendet wird.

Stimme: Hierunter fallen Modalitäten wie Tonfall, Sprechgeschwindigkeit, Pausen, Lachen, Seufzen.

Körpersprache: Die Körpersprache auf der Beziehungsebene dient dazu, bestimmte Aussagen zu unterstreichen bzw. Zuneigung oder Abneigung auszudrücken. Anders als Gesten bedient sich die Körpersprache keines eindeutig definierten Alphabets, sondern wird durch soziale Interaktionen erlernt.

Aufmerksamkeit: Ein ganz wesentlicher Punkt bei der zwischenmenschlichen Kommunikation ist die Aufmerksamkeit, die man seinem Gesprächspartner durch Blickkontakt oder Bestätigungen wie Kopfnicken entgegenbringt.

Gesichtsausdruck: Die nonverbale Kommunikation zwischen Menschen bedient sich in sehr starkem Maße des Gesichtsausdrucks. Durch Heben der Augenbrauen kann z.B. die Wichtigkeit einer Aussage unterstrichen werden. Menschen sind außerdem in der Lage, aus dem Gesichtsausdruck eines Gesprächspartners auf dessen emotionalen Zustand zu schließen.

Soziale Informationen: Unbewusst nehmen Menschen alle möglichen Informationen über ihre Gesprächspartner wahr und passen ihre Kommunikation entsprechend an. An erster Stelle zu nennen wäre in diesem Zusammenhang die Identität des Gegenübers. Bei unbekannten Personen sind das Alter und das Geschlecht ausschlaggebend für die Gestaltung und den Ablauf der Interaktion.

1.4 Kommunikation mit Servicerobotern

Da auch in der Servicerobotik eine natürliche Kommunikation angestrebt wird, gelten die oben genannten Punkte im Prinzip auch hier. Bei Servicerobotern bleibt die Kommunikation aber in der Regel auf die Sachebene beschränkt. Der Grund hierfür ist ganz einfach, dass es oftmals ausreicht, wenn Nutzer die gewünschte Serviceleistung spezifizieren. Durch das Fehlen von Kommunikationsmechanismen auf der Beziehungsebene wirkt die Kommunikation mit solchen Robotern allerdings sehr unnatürlich. Hinzu kommt, dass auch auf der Sachebene eher technische Ein- und Ausgabemedien verwendet werden. Dies ist darin begründet, dass natürliche Ein- und Ausgabemodalitäten für viele Einsatzbereiche nicht geeignet sind, sei es, weil sie in bestimmten Umgebungen nicht robust funktionieren (Spracherkennung bei lauten Hintergrundgeräuschen) oder weil der Einsatz anderer Modalitäten schneller und effektiver zum Ziel führt. Abbildung 1.3 veranschaulicht die Verhältnisse bei heutigen Servicerobotern.

Sprache: Wenn ein Roboter auf natürliche Art und Weise mit einem Menschen kommunizieren soll, muss er in der Lage sein, natürliche Sprache unabhängig vom jeweiligen Sprecher zu verstehen. Bei der Kommunikation mittels Sprache kann der Nutzer seine Hände für andere Aufgaben verwenden. Umgekehrt kommt der Sprache auch als Ausgabemedium eine entscheidende Bedeutung bei. So können Text- oder graphischen Ausgaben durch Sprachausgaben unterstützt werden, so dass wichtige Informationen redundant und damit sicherer übermittelt werden. Sprachausgaben kommen auch dann beim Nutzer an, wenn ein Display nicht sichtbar ist, wie es z.B. während der Lotsenfahrt vorkommen kann, so dass sie auch besonders in Gefahrensituationen geeignet sind.

Gestik: In diesem Zusammenhang wird in der Regel mit Zeigegesten für das Referenzieren von Objekten operiert bzw. mit Handgesten aus einem definierten Alphabet.

Touch-Display: Da bei der Instruierung eines Serviceroboters in der Regel ein größerer Informationsaustausch nötig ist und dieser möglichst schnell und sicher vonstatten gehen

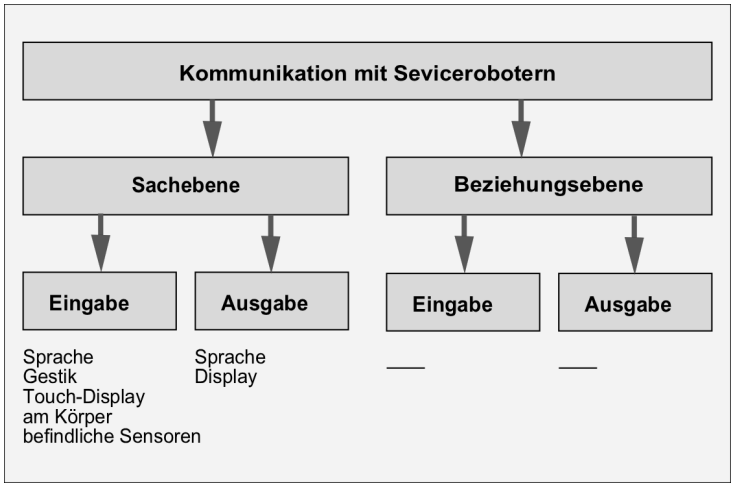


Abbildung 1.3: *Kommunikation mit Servicerobotern. Im Gegensatz zur menschlichen Kommunikation beschränkt sich die Kommunikation mit Servicerobotern im wesentlichen auf die Sachebene. Hier kommen Modalitäten zum Einsatz, mit denen konkrete Informationen schnell und effizient übermittelt werden können. Außerdem ist im Vergleich zur menschlichen Kommunikation bislang eine starke Asymmetrie zwischen Eingabe- und Ausgabemodalitäten zu verzeichnen.*

soll, wird die Präsentation von Informationen in der Regel über Displays durchgeführt. Auf diese Weise können viele Informationen übersichtlich dargestellt werden. Bei Verwendung eines berührungsempfindlichen Displays kann das selbe Medium gleichzeitig für die Eingabe verwendet werden.

Am Körper befindliche Sensoren: In zahlreichen Arbeiten werden Sensoren verwendet, die direkt am Körper des Interaktionspartners angebracht werden. Beispiele sind Datenhandschuhe für die Detektion von Gesten, Mikrofone, die der Kommunikationspartner tragen muss, da die heutigen Spracherkennungssysteme zu empfindlich auf Umgebungsgeräusche reagieren oder Sensoren zur Bestimmung des Hautleitwertes, mit denen auf den emotionalen Zustand oder auf Stress beim Gesprächspartner geschlossen werden kann.

Andere in der Mensch-Maschine-Kommunikation häufig eingesetzte Geräte wie Tastaturen sind in der Servicerobotik eher unüblich, da sie zum einen relativ sperrig sind und zum anderen die Eingabe mit Tastaturen verhältnismäßig zeitintensiv ist. Zudem sollte die Information immer so präsentiert werden, dass eine einfache Auswahl mit wenigen Tasten möglich ist und keine längeren Texte einzugeben sind.

Die Forschung an Servicerobotern konzentriert sich also, bis auf wenige Ausnahmen, auf die zu erbringende Dienstleistung, betrachtet die Kommunikation eher als Mittel zum Zweck und beschränkt sich somit im Wesentlichen auf deren Sachebene. Wie bereits erläutert, ist aber die zwischenmenschliche Kommunikation in starkem Maße durch die Beziehungsebene charakterisiert. Im nächsten Abschnitt wird eine Forschungsrichtung vorgestellt, die besonderes Augenmerk auf die Beziehungsebene bei der Mensch-Roboter-Interaktion legt.

1.5 Kommunikation mit sozialen Robotern

Seit Menschen von mechanischen Wesen träumen, die mit und für den Menschen arbeiten, werden diese in Form und Funktionsweise als menschenähnlich gedacht. Insbesondere die Kommunikation mit solchen Robotern unterscheidet sich nicht von der zwischenmenschlichen Kommunikation. Die Forschungsrichtung der *sozialen* oder auch *sozial interaktiven Roboter* soll diesen Traum einer menschenähnlichen Kommunikation mit Maschinen wahr machen, indem deren Beziehungsaspekt vermehrt in den Vordergrund gerückt wird. In [Fong et al., 2002] geben Fong et al. eine Übersicht über die aktuelle Forschung auf diesem Gebiet. Soziale Roboter werden hier wie folgt definiert:

> Soziale Roboter sind körperliche Agenten, die Teil einer heterogenen Gruppe sind, einer Gesellschaft von Robotern und/oder Menschen. Sie sind in der Lage, sich gegenseitig zu erkennen und soziale Interaktionen einzugehen. Sie besitzen eine Geschichte (die Wahrnehmung und die Interpretation der Welt erfolgen entsprechend ihrer eigenen Erfahrungen). Sie kommunizieren explizit miteinander und lernen voneinander.

Die zugrunde liegende Annahme bei der Entwicklung sozialer Roboter besteht darin, dass Menschen es bevorzugen, mit Maschinen auf die gleiche Weise zu interagieren, wie sie auch mit anderen Menschen interagieren würden. Bei sozialen Robotern muss es sich nicht um Serviceroboter handeln. Sie können ganz unterschiedliche Formen und Funktionen haben. Insbesondere können sie ausschließlich dazu dienen, Menschen zur sozialen Interaktion zu motivieren (z.B. KISMET und LEONARDO, siehe Abschnitt 6.2). Abbildung 1.4 zeigt eine Auflistung der bei sozialen Robotern eingesetzten Kommunikationsmodalitäten.

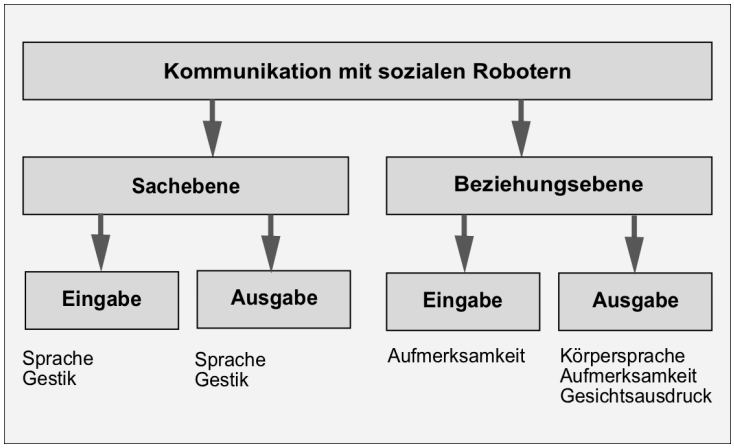


Abbildung 1.4: *Modalitäten bei der Kommunikation mit sozialen Robotern.*

Dazu ist zu bemerken, dass die Modalitäten auf der Beziehungsebene in der Regel nicht symmetrisch eingesetzt werden. Sozial interaktive Roboter versuchen diese Modalitäten so nachzubilden, dass der Roboter bei seinem Interaktionspartner als sozial kompetent angesehen wird. Das bedeutet, dass Körpersprache und Gesichtsausdrücke verwendet werden, um emotionale

Zustände des Roboters zu verdeutlichen. In umgekehrter Richtung aber sind auch sozial interaktive Roboter in der Regel nicht in der Lage, den emotionalen Zustand oder andere Informationen eines menschlichen Kommunikationspartners zu ermitteln.

Es gibt prinzipiell zwei Entwurfsprinzipien für sozial interaktive Roboter. Dabei handelt es sich zum einen um stark biologisch inspirierte und zum anderen um eher funktionale Entwürfe.

Biologisch inspirierte Entwürfe: Hier wird versucht, Roboter zu entwickeln, welche die soziale Intelligenz von Lebewesen nachbilden. Die Grundidee liegt darin, dass die Natur das beste Modell für lebensechte Verhaltensweisen ist. Insbesondere wird vermutet, das ein Roboter, damit er von Menschen verstanden wird, einen naturalistischen Körper besitzen muss. Er muss mit seiner Umgebung auf die selbe Weise interagieren, wie dies Lebewesen tun, und er muss die selben Dinge wahrnehmen und auf die selben Dinge seine Aufmerksamkeit richten, wie dies Menschen tun. Solche Entwürfe erlauben es, die wissenschaftlichen Theorien, welche die Basis für den Entwurf bilden, zu kontrollieren, zu testen und zu verbessern. Diese Theorien sind die Ethologie, Theorien über die Struktur von Interaktionen [Magnusson, 2005], die Erkenntnistheorie und die Entwicklungspsychologie. Typische Vertreter sind die am MIT entwickelten Roboter KISMET und LEONARDO, siehe Abschnitt 6.2.

Funktionale Entwürfe: Die nach diesem Prinzip entworfenen Roboter sollen nach außen sozial intelligent wirken, obgleich die internen Mechanismen nicht zwangsläufig biologisch oder psychologisch inspiriert sind. Die Grundidee ist hier, dass es möglich ist, den Eindruck eines künstlichen sozialen Agenten zu schaffen, ohne notwendigerweise zu wissen, wie die zugrunde liegenden biologischen Mechanismen funktionieren. Im Gegensatz zu den biologisch inspirierten Architekturen haben funktionale Entwürfe in der Regel eingeschränkte operationale Zielstellungen. Diese „entworfenen“ Roboter müssen nur oberflächlich sozial kompetent sein und nur bestimmte Effekte bei der Kommunikation mit Nutzern zeigen. Sie besitzen in der Regel nur eingeschränkte sensorische und aktuatorische Fähigkeiten für die Interaktion mit Menschen. Typische Vertreter sind der an der Universität der Bundeswehr in München entwickelte Roboter HERMES und der an der Carnegie Mellon University entwickelte Roboter MINERVA (siehe Abschnitt 6.1).

Aus diesen Charakterisierungen ergibt sich ein fließender Übergang zwischen biologisch inspirierten sozialen Robotern auf der einen Seite über eher funktionale soziale Roboter bis hin zu Servicerobotern auf der anderen Seite.

1.6 Anspruch der Arbeit

In den letzten Abschnitten wurden das Baumarktszenario vorgestellt und die Anforderungen an die Mensch-Roboter-Interaktion definiert. Insbesondere wurde darauf hingewiesen, dass die Interaktion mit Personen, die nicht in den Umgang mit dem Serviceroboter eingewiesen sind, eine besonders intuitive und natürliche Interaktion voraussetzt. Weiterhin wurde dargelegt, dass die Kommunikation bei aktuellen Servicerobotern in der Regel auf die Sachebene beschränkt bleibt, was eine erhebliche Einschränkung darstellt und insbesondere in dem vorgestellten Einsatzfeld zu Akzeptanzproblemen führen dürfte.

Sozial interaktive Roboter widmen sich zwar verstärkt der Beziehungsebene der Kommunikation, tun dies aber hauptsächlich auf der Ausgabeseite. Sie besitzen Charaktere und Emotionen, die sie mit realistischen Gesichtsausdrücken darstellen können und vermitteln dem Interaktionspartner durch Blickkontakt ihre Aufmerksamkeit während der Interaktion. Auf der Eingabeseite auf der Beziehungsebene sind die Fähigkeiten aber in der Regel auf eine Detektion des Interaktionspartners beschränkt.

Diese Arbeit versucht, diese Lücke zu schließen. Dabei geht es insbesondere darum, Methoden für die Beziehungsebene der Kommunikation zu realisieren, die den Roboter befähigen, für den Interaktionsprozess relevante Informationen über seinen Interaktionspartner zu extrahieren. Eine wesentliche Maßgabe bei der Entwicklung der Methoden besteht darin, dass keine Sensoren eingesetzt werden sollen, die der Nutzer am Körper tragen muss und die ihn somit in seiner Bewegungsfreiheit hemmen und einer natürlichen Interaktion im Wege stehen. Es werden daher ausschließlich Methoden der visuellen Informationsverarbeitung zum Einsatz kommen.

Das anvisierte Aufgabenspektrum umfasst die Erkennung der Identität, des Geschlechtes, des Alters und des Gesichtsausdrucks des Nutzers. Damit soll der Roboter befähigt werden, bekannte Personen zu erkennen bzw. eine grobe Kategorisierung seiner Nutzer vorzunehmen, also Nutzermodelle zu erstellen und die Kommunikationsstrategie entsprechend anzupassen. Mit dem Gesichtsausdruck des Nutzers verfügt der Roboter über eine Schätzung seines emotionalen Zustandes, der für eine Bewertung des Interaktionsprozesses herangezogen werden kann.

Die Erkennung solcher Informationen über den Nutzer spielt bei heutigen Robotersystemen eine eher untergeordnete Rolle. Sie ist jedoch nicht nur für das Baumarktszenario von Interesse, sondern allgemein für Service- und soziale Roboter. Ausgehend von den gestellten Anforderungen an eine natürliche Interaktion wird im nächsten Kapitel ein Interaktionskonzept für den Serviceroboter Perses erstellt.

1.7 Interaktionskonzept für PERSES

1.7.1 Interaktionsmodalitäten

Die Zielstellung dieser Arbeit ist die Entwicklung von Teilkomponenten für eine intuitive und möglichst natürliche Interaktion zwischen einem mobilen Serviceroboter und seinen Nutzern. In der Einleitung wurde eine Charakterisierung von Servicerobotern und von sozialen Robotern vorgenommen. Während Serviceroboter eine bestimmte Aufgabe zu erfüllen haben und die Interaktion mit Menschen nur Mittel zum Zweck ist, werden soziale Roboter nur für diesen einen Zweck entwickelt. Die Forschungsarbeiten zu sozial interaktiven Robotern geben jedoch wichtige Hinweise darauf, welche Komponenten der Mensch-Roboter-Interaktion von wesentlicher Bedeutung für eine natürliche Interaktion sind. Im folgenden werden eine Reihe von Modalitäten für die Mensch-Roboter-Interaktion betrachtet und es werden Überlegungen angestellt, inwieweit diese im Baumarktszenario sinnvoll eingesetzt werden können.

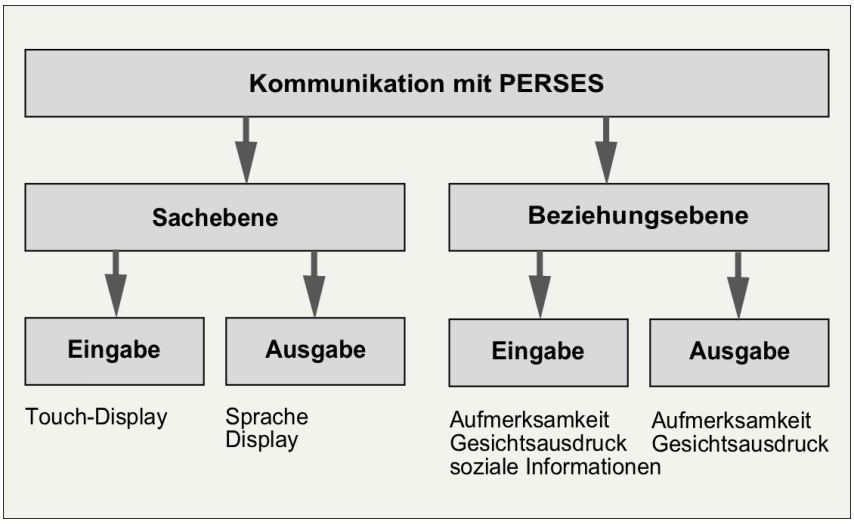


Abbildung 1.5: *Modalitäten bei der Kommunikation mit* PERSES.

Sprache: Auf den Einsatz von Spracherkennern wurde in dieser Arbeit aus zwei Gründen verzichtet. Zum einen arbeiten heute verfügbare Spracherkenner bei lauten Hintergrundgeräuschen, wie sie im Baumarkt die Regel sind, nicht zuverlässig genug. Deshalb müsste das Mikrophon direkt am Nutzer angebracht werden. Zum anderen ist das Baumarktszenario denkbar ungeeignet für den Einsatz einer Spracherkennung, da bei der Bestimmung der Kundenwünsche mit extrem vielfältigen Problembeschreibungen zu rechnen ist. Das nötige Vokabular umfasst sicherlich mehrere tausend Wörter. Auch die Verwendung eines eingeschränkten Vokabulars ist nicht sinnvoll, da dann der Nutzer vor oder während der Interaktion über das vom Roboter verstandene Vokabular unterrichtet werden müsste. Die Ausgabe von Sprache ist dagegen durchaus sinnvoll, da sie dazu dienen soll, die Aufmerksamkeit einer Person auf den Roboter zu lenken oder in Situationen, in denen das Display nicht sichtbar ist, Informationen an den Nutzer zu vermitteln.

Gestik: Für den Einsatz im Baumarkt wäre die Interpretation von Gesten sicherlich von großem Vorteil. So könnte ein Nutzer z.B. mit einer Zeigegeste auf ein im Regal befindliches Produkt verweisen und den Roboter um Zusatzinformationen bitten. Die in der Servicerobotik verwendeten Gesten sind in der Regel keine natürlichen Gesten, mit denen auf bestimmte Objekte verwiesen wird oder die gesprochene Aussagen unterstreichen. Es wird vielmehr mit Gesten aus einem fest vorgegebenen, mehr oder weniger natürlichem Vokabular gearbeitet. Wie bei der Sprache ist es auch hier nicht wünschenswert, das für die Interaktion gültige Vokabular im Vorfeld festzulegen, auszuhandeln oder zu trainieren. Eine Erkennung von natürlichen Gesten unter realen Bedingungen ist bis Heute Gegenstand der Forschung und nicht Bestandteil dieser Arbeit.

Touch-Display: Das Touch-Display wird für den Einsatz im Baumarkt als besonders geeignet betrachtet. Mit ihm ist es möglich, große Informationsmengen übersichtlich zu präsentieren und Eingaben vom Nutzer entgegenzunehmen. Heutzutage sind die meisten Menschen durch den Umgang mit Geld- oder Fahrscheinautomaten mit der Bedienung von Touch-Displays vertraut. Für die große Vielzahl an Produkten und Produktkategorien in einem Baumarkt scheint eine Auswahl über ein Touch-Display sinnvoll.

Aufmerksamkeit: Die Fähigkeit des Roboters zu erkennen, ob ein Nutzer in der Nähe ist und ob dieser aufmerksam an der Interaktion teilnimmt, wurde bereits als sehr wichtig für eine natürliche Interaktion herausgestellt. So sollte es möglich sein zu erkennen, ob sich eine Person dem Roboter nähert, um diese gezielt anzusprechen. Während der Interaktion sollte PERSES in der Lage sein, seinen Interaktionspartner kontinuierlich zu verfolgen und gegebenenfalls, wenn dieser den Roboter verlässt, die Interaktion nicht stupide zu Ende zu führen, sondern geeignet zu reagieren. Umgekehrt soll auch der Roboter in der Lage sein, seine Aufmerksamkeit an den Nutzer zu vermitteln. So ist auch im Baumarkt zu erwarten, dass Nutzer eher dazu bereit sind, mit dem Roboter in Kontakt zu treten, wenn dieser seinerseits durch Anschauen der Baumarktkunden Interaktionsbereitschaft signalisiert.

Gesichtsausdruck: Um die kurz- und mittelfristige Zufriedenheit eines Nutzers zu ermitteln, soll dessen Gesichtsausdruck ausgewertet werden. Auf diese Weise können vom Roboter ausgeführte Aktionen unmittelbar evaluiert werden. Der aus der Mimik abgeleitete emotionale Zustand des Nutzers könnte hierbei als Reinforcement-Signal einer lernfähigen Kommunikationsarchitektur herangezogen werden.

Soziale Informationen: Wie bereits erwähnt, soll der Roboter ein Modell von seinem aktuellen Nutzer anlegen, damit er im Falle, dass er diesen aus dem Sichtfeld verliert, entscheiden kann, ob er es wieder mit der selben Person zu tun hat. Mit weiteren Informationen wie Alter und Geschlecht des Nutzers sollen perspektivisch Nutzerprofile angelegt werden.

1.7.2 Hardware

Nachdem ein Interaktionskonzept entwickelt wurde, sollen die Plattformen mit ihren Hardware-Komponenten vorgestellt werden, auf denen dieses Konzept realisiert wird. Dabei handelt es sich zum einen um den Prototypen des Shopping-Assistenten PERSES und zum anderen um den Standalone-Arbeitsplatz MIMIR.

1.7.2.1 Der Shopping-Assistent PERSES

Bei PERSES handelt es sich um einen B21-Roboter der Firma RWI. Dieser ist standardmäßig mit zwei Ringen von jeweils 24 Sonarsensoren ausgestattet, mit denen die Entfernung zu Objekten bis rund $5m$ bestimmt werden kann. PERSES wurde mit einem Activ-Vision Kopf ausgestattet, der auf einer Pan-Tilt-Unit angebracht ist. In diesem Kopf befinden sich zwei Kameras mit einem Basisabstand von $10cm$, von denen jedoch für das vorgestellte Interaktionssystem nur eine eingesetzt wird. Der Kopf verfügt über ein einfaches Gesicht, dass aus einer Reihe von LED-Arrangements besteht, mit dem Emotionen dargestellt werden können. PERSES besitzt eine über dem Kopf angebrachte Kamera mit omnidirektionalem Objektiv. Für die Interaktion mit seinen Nutzern wurde PERSES mit einem Touch-Display ausgestattet. Sprachausgaben erfolgen über zwei in der Nähe des Kopfes angebrachte Lautsprecher. Abbildung 1.6 zeigt PERSES in der aktuellen Hardware-Konfiguration.

1.7.2.2 Der Standalone-Arbeitsplatz MIMIR

Um das Interaktionssystem auch unabhängig von PERSES entwickeln und demonstrieren zu können, wurde der Arbeitsplatz MIMIR entworfen, der alle notwendigen Komponenten enthält. Er besteht aus einer omnidirektionalen Kamera und einem auf einer Pan-Tilt-Unit angebrachten Kopf mit Frontalkamera. Der Kopf verfügt ebenfalls über ein Gesicht mit LED-Arrays zur Darstellung von Emotionen. Im Gegensatz zu PERSES verfügt MIMIR nicht über Sonarsensoren und kann somit nicht mit dem in dieser Arbeit vorgestellten Verfahren den Abstand zu einem Nutzer bestimmen. Das Touch-Display wird bei MIMIR durch den Bildschirm und die Maus ersetzt. Mit der Möglichkeit Sprache auszugeben, verfügt MIMIR schließlich über fast alle Interaktionsmöglichkeiten, die auch bei PERSES realisiert wurden. Nicht verfügbar sind alle Interaktionsmöglichkeiten, die Bewegungen der Roboterplattform einbeziehen, wie z.B. das Annähern an einen Nutzer.

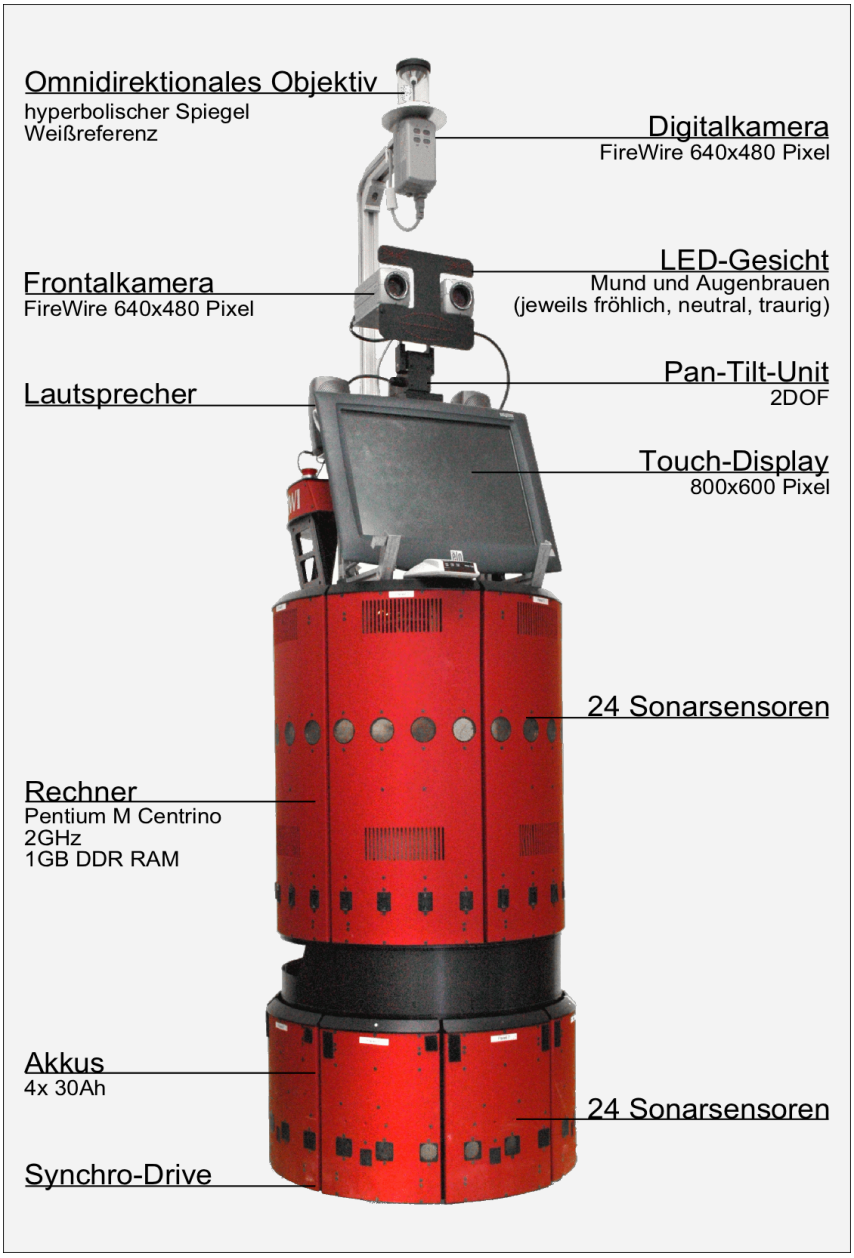


Abbildung 1.6: *Bei der mobilen Forschungsplattform* PERSES *handelt es sich um einen B21-Roboter der Firma RWI, der für die Interaktion mit Nutzern mit Erweiterungen wie einem Touch-Display, einer Pan-Tilt-Unit, einem Gesicht, Lautsprechern, einer Kamera mit omnidirektionalem Objektiv und zwei Frontalkameras ausgestattet wurde.*

Abbildung 1.7: MIMIR *ist ein Standalone-Arbeitsplatz mit nahezu identischen Interaktionsmöglichkeiten wie* PERSES*, siehe Text.*

1.7.3 Systemarchitektur

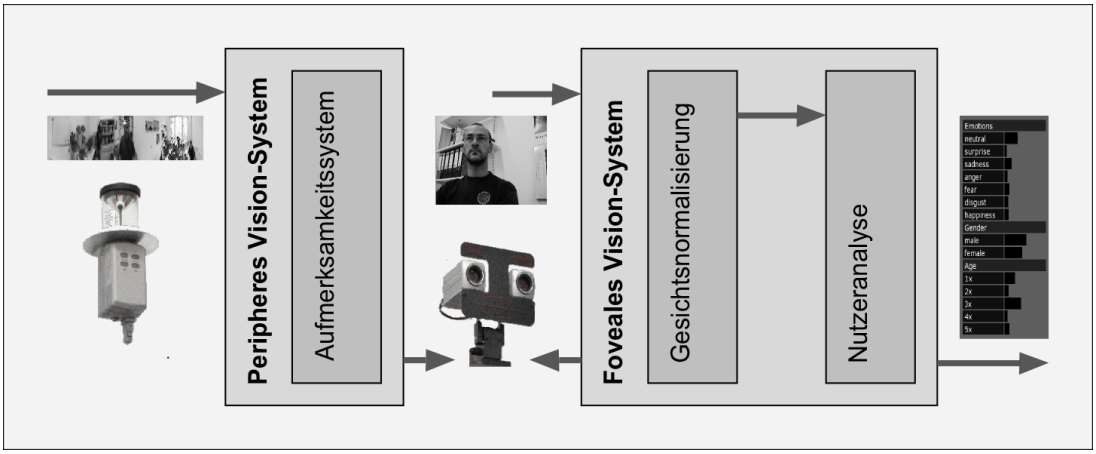


Abbildung 1.8: *Überblick über die Systemarchitektur, bestehend aus peripherem und fovealem Vision-System. Das periphere Vision-System arbeitet auf Bildern der omnidirektionalen Kamera und erzeugt Hypothesen für potentielle Nutzer. Da der Roboterkopf auf die aktuelle Nutzerhypothese ausgerichtet wird, können mit der Frontalkamera hochaufgelöste Bilder aufgenommen werden, die vom fovealen Vision-System weiter ausgewertet werden. Hier wird zunächst nach einem Gesicht gesucht, um die Hypothese des peripheren Vision-Systems zu verifizieren. Daraufhin werden Merkmale im Gesicht gesucht, mit deren Hilfe das Gesicht vor der weiteren Analyse in eine normalisierte Darstellung gebracht werden kann. Im letzten Schritt wird das Gesicht hinsichtlich Geschlecht, Alter, Gesichtsausdruck und Identität analysiert.*

Abbildung 1.8 zeigt die für die Lösung der gestellten Aufgaben im Rahmen dieser Arbeit entwickelte Systemarchitektur. Sie gliedert sich in ein peripheres und in ein foveales Vision-System. Diese Aufteilung hat einen direkten biologischen Bezug und wird besonders bei sozialen Robotern angewandt. Der Hintergedanke ist, dass in biologischen Sehsystemen der periphere Bereich des Gesichtsfeldes, der zwar sehr groß ist, aber nur eine verhältnismäßig geringe Auflösung besitzt, für die visuelle Aufmerksamkeit genutzt wird. Um auffällige Bildregionen genauer analysieren zu können, müssen sie durch Fixieren in den fovealen Bereich mit einer höheren Auflösung gebracht werden. Die Idee einer solchen Aufteilung bei Verwendung von omnidirektionalen Sensoren wurde zum ersten Mal in [Wilhelm et al., 2003b] veröffentlicht.

1.7.3.1 Peripheres Vision-System

Die Aufgabe des peripheren Vision-Systems besteht darin, die unmittelbare Umgebung des Roboters nach potentiellen Nutzern abzusuchen und diese kontinuierlich zu verfolgen. Es übernimmt also die Funktion eines Aufmerksamkeitssystems. Da Personen auch erfasst werden sollen, wenn sie sich dem Roboter von hinten nähern, kommen Sensoren zum Einsatz, die das gesamte Umfeld des Roboters erfassen. Dabei handelt es sich um die omnidirektionale Kamera und um die Sonarsensoren des Roboters. Als Merkmale für die Personendetektion dienen Hautfarbe, Bewegung und Entfernung. Das periphere Vision-System soll seine Hypothesen über die Zeit

verfolgen, eine Hypothese für einen möglichen Nutzer auswählen und den Kopf des Roboters auf diese ausrichten. Das Aufmerksamkeitssystem wird in Kapitel 2 vorgestellt.

1.7.3.2 Foveales Vision-System

Um weitere Informationen über den Nutzer zu erhalten, werden mit der Frontalkamera hochaufgelöste Bilder von dessen Gesicht aufgenommen und mit dem fovealen Vision-System analysiert.

Als erstes wird die Nutzerhypothese mit einem Gesichtsdetektor verifiziert. Detaillierte Betrachtungen zum Gesichtsdetektor befinden sich im Abschnitt 3.2. Wenn ein Gesicht gefunden wurde, wird die Frontalkamera kontinuierlich nachgeführt, so dass es immer möglichst in der Mitte des Bildes erscheint. Für die weitere Verarbeitung wird das Gesicht aus dem Bild der Frontalkamera ausgeschnitten. Anschließend werden die Positionen der Augen im Bild geschätzt und das Gesicht wird anhand einer affinen Transformation in eine normalisierte Darstellung gebracht. Dieser Verarbeitungsschritt wird in Abschnitt 3.3 beschrieben.

Das eigentliche Ziel besteht darin, detaillierte Informationen über den Nutzer aus dem Bild zu extrahieren. Dazu werden aus der normalisierten Darstellung des Gesichtes Merkmale extrahiert und hinsichtlich Identität, Geschlecht, Alter und Gesichtsausdruck klassifiziert. Für diesen Analyseschritt werden in Kapitel 4 drei Verfahren vergleichend untersucht. Es handelt sich dabei um das Elastic-Graph-Matching (Abschnitt 4.4), die Independent-Component-Analysis (Abschnitt 4.5) und die Active-Appearance-Models (Abschnitt 4.6).

1.8 Gliederung der Arbeit

Die Gliederung der Arbeit folgt dem Aufbau der entwickelten Systemarchitektur. In Kapitel 2 wird das Aufmerksamkeitssystem beschrieben. Kapitel 3 befasst sich mit der Erzeugung einer normalisierten Darstellung des Gesichtes und Kapitel 4 mit der Analyse. Für jeden dieser Teilschritte wird zunächst die Aufgabenstellung definiert und es werden mögliche aus der Literatur bekannte Lösungsansätze aufgezeigt. Die Theorie und vergleichende Untersuchungen zu den realisierten Verfahren werden ebenfalls in den jeweiligen Teilabschnitten abgehandelt.

In Kapitel 5 wird die Software-Architektur für die Integration der Teilsysteme zu einem Gesamtsystem beschrieben und Untersuchungen zum Zusammenspiel der Teilkomponenten vorgestellt.

In Kapitel 6 werden in den letzten Jahren entwickelte Serviceroboter und soziale Roboter daraufhin untersucht, inwieweit die in der Einleitung definierten Anforderungen an eine natürliche Mensch-Maschine-Interaktion erfüllt sind. Dabei erfolgt eine Gegenüberstellung mit und eine Abgrenzung zu den in dieser Arbeit entwickelten Mechanismen.

Kapitel 2

Aufmerksamkeitssystem

2.1 Aufgabenstellung

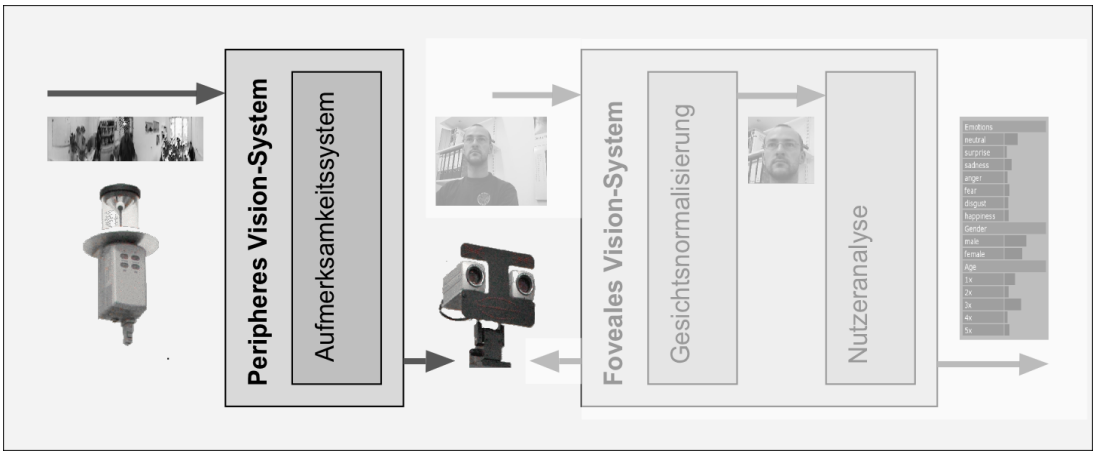


Abbildung 2.1: *Das Aufmerksamkeitssystem hat die Aufgabe, potentielle Nutzer in der Umgebung des Roboters zu finden. Es wertet Bilder der omnidirektionalen Kamera und Messungen der Sonarsensoren aus und steuert die Pan-Tilt-Unit so an, dass ein potentieller Nutzer mit der Frontalkamera angeschaut wird, so dass ein hochaufgelöstes Bild aufgenommen und weiter ausgewertet werden kann.*

Wie in den vorhergehenden Kapiteln gezeigt, ist es für einen Serviceroboter von entscheidender Bedeutung, seine Aufmerksamkeit auf eine Person in seiner Umgebung richten zu können und sich während der Kommunikation nicht von anderen Personen ablenken zu lassen. Dazu muss er seine unmittelbare Umgebung nach potentiellen Nutzern absuchen und diese kontinuierlich verfolgen. Sobald die Kommunikation mit einem Nutzer beendet ist, soll sich der Roboter einer anderen Person zuwenden. Das Aufmerksamkeitssystem muss also in der Lage sein, mehrere Personen verfolgen und bei Bedarf zwischen diesen umschalten zu können. Abbildung 2.1 zeigt eine Einordnung des Aufmerksamkeitssystems in die entwickelte Systemarchitektur.

Zunächst erfolgt eine Aufarbeitung der Literatur zum Thema Aufmerksamkeitssysteme. Darauf-

hin werden für die Personendetektion geeignete Merkmale definiert und erläutert, wie diese im realisierten System berechnet werden. Weitere Abschnitte befassen sich mit einer geeigneten Fusion der gefundenen Merkmale und mit der Bildung von Nutzerhypothesen aus der fusionierten Merkmalskarte. Am Ende des Kapitels wird beschrieben, wie die Pan-Tilt-Unit des Roboters aufgrund der gefundenen Hypothese so ausgerichtet wird, dass ein hochaufgelöstes Bild der Hypothese für die weitere Analyse aufgenommen werden kann.

2.2 Literatur

Zum einen stellt sich hier die Frage, welche Merkmale typischerweise verwendet werden, um auffällige Bildregionen zu definieren und zum anderen, wie letztendlich eine Region ausgewählt wird, auf die das System seine Aufmerksamkeit richtet.

Merkmale: Bei Systemen zur Personendetektion werden in der Regel entfernungsmessende, visuelle oder auditorische Sensoren verwendet. Beispiele für entfernungsmessende Sensoren sind Laser-Scanner und Ultraschallsensoren. In [Schulz and Burgard, 2001] wird aus den lokalen Minima in einem Laser-Scan auf die Anwesenheit von Personen geschlossen. Visuelle Merkmale sind Hautfarbe, Bewegung und Gesichtsstrukturen. Häufig wird auch eine Geräuschquellenlokalisation für die Detektion von Personen herangezogen. Typischerweise werden zur Erhöhung der Sicherheit die Ergebnisse mehrerer Sensorsysteme in so genannten Auffälligkeitskarten kombiniert [Corty and Marchand, 2003].

Selektion Die Auswahl einer auffälligen Struktur kann z.B. durch eine Maximumsauswahl in der Auffälligkeitskarte geschehen [Feyrer and Zell, 1999], was allerdings zu einem sprunghaften Wechseln zwischen Hypothesen führen kann. Mit Hilfe von neuronalen Feldern, die auf den Auffälligkeitskarten arbeiten, kann eine zeitliche Stabilisierung erreicht werden. Durch den Einsatz von Tracking-Verfahren wird nicht nur eine zeitliche Stabilisierung, sondern auch eine Reduzierung der Rechenzeit erreicht.

Im Folgenden werden einige Beispiele für Auffälligkeitssysteme für mobile Roboter beschrieben. In [Schulz et al., 2001] und [Schulz et al., 2003] werden die lokalen Minima in einem Laser-Scan als Nutzerhypothesen betrachtet. Mittels eines Multi-Target-Trackers können mehrere bewegte Objekte im Umfeld des mobilen Roboters verfolgt werden. In [Feyrer and Zell, 1999] wird eine Kombination aus Hautfarbe, Bewegung, Kontur- und Stereoinformationen verwendet, um die Position einer Person in einem Bild zu bestimmen. Die durch eine Maximumsauswahl selektierte Person wird mit der Kamera kontinuierlich angeschaut und der Roboter nähert sich ihr bis auf einen Mindestabstand an. Um bewegte Objekte auch während der Fahrt des Roboters detektieren zu können, wird eine Eigenbewegungskompensation durchgeführt. Beim Roboter RoboVie [Shiomi et al., 2004] werden auffällige Bildstrukturen durch eine Kombination aus Hautfarb- und Bewegungsdetektion gebildet. Diese werden mit Hilfe eines Partikelfilters über die Zeit verfolgt. Das System arbeitet auf Bildern einer omnidirektionalen Kamera. Eine Verifizierung der Hypothesen erfolgt durch einen Gesichtsdetektor, der auf Bildern einer Frontalkamera arbeitet. Der

sozial interaktive Roboter KISMET [Breazeal and Scassellati, 1999] verwendet einen Gesichts-, einen Hautfarb- und einen Bewegungsdetektor. Der Auswahlmechanismus berechnet eine Linearkombination der einzelnen Auffälligkeitskarten, wobei die Gewichte für die einzelnen Merkmale durch Motivationen und Emotionen des Roboters gesteuert werden. Mit Hilfe einer zeitlich zunehmenden Abwichtung von auffälligen Strukturen werden Habituationseffekte realisiert, so dass der Roboter nach einiger Zeit seine Aufmerksamkeit auf andere Objekte richtet. Bei den Robotern ASIMO und QRIO wird außer visuellen Cues auch eine Geräuschquellenlokalisation eingesetzt, bei WAKAMARU außerdem eine Wärmebildkamera.

Im Folgenden werden die Strukturen definiert, die in dieser Arbeit für die Personendetektion verwendet werden sollen.

2.3 Auffällige Strukturen

Damit auch Personen erfasst werden, die sich dem Roboter von hinten nähern, werden im Gegensatz zu vielen sozialen Robotern Sensoren eingesetzt, die das gesamte Umfeld des Roboters erfassen. Dabei handelt es sich zum einen um eine omnidirektionale Kamera und zum anderen um 24 Sonarsensoren, die rund um den Roboter angebracht sind. Dies bietet, wie sich zeigen wird, einige entscheidende Vorteile gegenüber der ausschließlichen Nutzung von Frontalkameras.

An dieser Stelle sollen zunächst die Sensordaten spezifiziert werden, die die „Aufmerksamkeit“ des Roboters erregen sollen. Da die Aufgabe in der Detektion von Personen besteht, die sich in der Nähe des Roboters befinden, bieten sich die Merkmale Hautfarbe, Bewegung und Entfernung an. In ersten Arbeiten wurden für das Aufmerksamkeitssystem nur die Merkmale Hautfarbe und Entfernung verwendet [Wilhelm et al., 2003c]. Die Hautfarb- und die Bewegungsdetektion werden auf dem Bild der omnidirektionalen Kamera berechnet. Die Bestimmung der Entfernung könnte ebenfalls visuell über eine Stereogeometrie bestimmt werden [Erich, 2003], für die hier vorgestellte Realisierung wird allerdings auf die Sonarsensoren des Roboters zurückgegriffen, zum einen, weil die Auswertung wesentlich weniger zeitintensiv ist und zum anderen, weil die Sonarsensoren wie die omnidirektionale Kamera ebenfalls das gesamte Umfeld des Roboters erfassen können. Zwar existieren erste Ansätze für eine Stereobildverarbeitung mit omnidirektionalen Kameras, ein entsprechendes Objektiv war aber bis zum jetzigen Zeitpunkt noch nicht verfügbar.

2.3.1 Hautfarbe

Hautfarbe ist ein häufig verwendeter Cue bei der Suche nach Personen in Bildern. Sie bietet den Vorteil, unabhängig von Eigenbewegungen des Roboters berechnet werden zu können und eignet sich dadurch besonders für den Einsatz auf mobilen Systemen. Weiterhin handelt es sich bei der Hautfarbdetektion um ein Pixel basiertes Verfahren, das keinen räumlichen Kontext im Bild benötigt und somit größen- und rotationsinvariant ist. Die Detektion von Hautfarbe bringt allerdings auch zwei wesentliche Probleme mit sich. Zum Ersten sind dies die extrem unter-

schiedlichen Ausprägungen von Hautfarbe für verschiedene Individuen. Zum Zweiten variiert die Hautfarbe für die selbe Person sehr stark in Abhängigkeit der Beleuchtung.

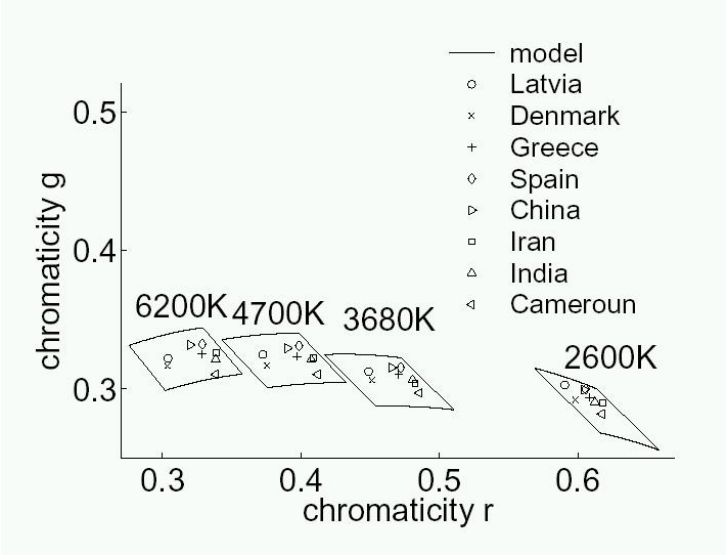


Abbildung 2.2: *Während eines normalen Bürotages mit einer Mischung aus natürlichem und Kunstlicht bewegt sich die Beleuchtung in einem Bereich von 2700 bis 6500K. Die Abbildung zeigt die Auswirkungen unterschiedlicher Beleuchtungen auf die Abbildung der Hautfarben verschiedener Personen in den rg-Farbraum. Die Abbildung wurde entnommen aus [Störring, 2004].*

Abbildung 2.2 veranschaulicht die Lage von Hautfarbverteilungen für Personen verschiedener ethnischer Gruppen bei verschiedenen Beleuchtungsbedingungen. Ein Hautfarbdetektor müsste die Varianzen zwischen den ethnischen Gruppen abdecken und gleichzeitig bei einer bestimmten Beleuchtung nur auf Hautfarbe ansprechen. Daraus wird ersichtlich, dass die Hautfarbdetektion nur dann zufriedenstellend funktionieren kann, wenn die Beleuchtungsverhältnisse bei der Aufnahme der Trainingsdaten für die Erstellung des Modells und bei der Anwendung desselben hinreichend ähnlich sind, was in der Regel nicht gewährleistet werden kann. Insbesondere im Baumarkt können die Beleuchtungsverhältnisse von reinem Kunstlicht bis zu reiner natürlicher Beleuchtung variieren. Die Aufgabe besteht also zunächst darin, auch unter solchen extremen Bedingungen eine weitgehende Farbkonstanz zu gewährleisten.

Eine mögliche Lösung für dieses Problem besteht in der kontinuierlichen Adaption des Farbmodells an die jeweiligen Beleuchtungsbedingungen [Böhme et al., 1998] [Yang and Waibel, 1996] [Yang and Waibel, 1998]. Dabei dient das detektierte Objekt selbst als Referenz für eine Adaption des Modells. Allerdings ist diese Vorgehensweise nach eigener Erfahrung problematisch, da das Gesicht des Nutzers für die Adaption in jedem Zeitschritt absolut sicher detektiert werden muss. Bei nur geringfügigen Positionsfehlern neigen solche adaptiven Modelle dazu, von der Beschreibung der eigentlichen Zielregion wegzudriften. Ein zu starkes Wegdriften des Farbmodells kann verhindert werden, indem eine Adaption nur innerhalb eines allgemeinen Hautfarbmodells zugelassen wird, welches unter verschiedenen Beleuchtungsbedingungen aufgenommen wurde [Fritsch et al., 2002] [Jang and Kweon, 2001] [Soriano et al., 2003].

Es gibt auch Bemühungen, die Lokalisation der Hautfarbregionen im Bild durch die Verwendung zusätzlicher Merkmale zu stabilisieren. So wird in [Fritsch et al., 2002] ein Gesichtsdetektor eingesetzt und das Farbmodell immer nur dann adaptiert, wenn an der entsprechenden Stelle auch ein Gesicht erkannt wurde. Solche Verfahren können das genannte Problem zwar entschärfen, nicht jedoch beseitigen, denn bei einer Falsch-Positiv-Detektion des Gesichtsdetektors wird das Modell mit fehlerhaften Farbwerten adaptiert. Auf der anderen Seite erhöhen sie die Berechnungskomplexität, was sich wiederum negativ auf ein kontinuierliches Verfolgen der Hautfarbregionen auswirkt.

Ein anderer Weg, Beleuchtungseinflüsse zu beseitigen, besteht in der Vorverarbeitung des Bildes mit Farbkonstanzalgorithmen. Die Aufgabe solcher Algorithmen besteht in der Erzeugung eines Bildes, wie es bei einer so genannten kanonischen Beleuchtung entstanden wäre, ausgehend von einem Bild, das bei einer beliebigen unbekannten Beleuchtung aufgenommen wurde. In [Funt et al., 1998] wurden verschiedene Farbkonstanzalgorithmen auf ihre Tauglichkeit im Zusammenhang mit einer farbbasierten Objekterkennung getestet. Dabei konnte zwar eine signifikante Steigerung der Erkennungsraten durch die Anwendung dieser Algorithmen nachgewiesen werden, trotzdem war keine robuste Erkennung unter variablen Beleuchtungen möglich.

Im Gegensatz zu den bisher genannten Verfahren werden in dieser Arbeit die speziellen Abbildungseigenschaften des omnidirektionalen Objektivs ausgenutzt. Dadurch kann ein automatischer Weißabgleich durchgeführt und weitgehende Farbkonstanz bei unterschiedlichen Beleuchtungsbedingungen erreicht werden, siehe Abschnitt 2.3.1.3.

2.3.1.1 Farbraum

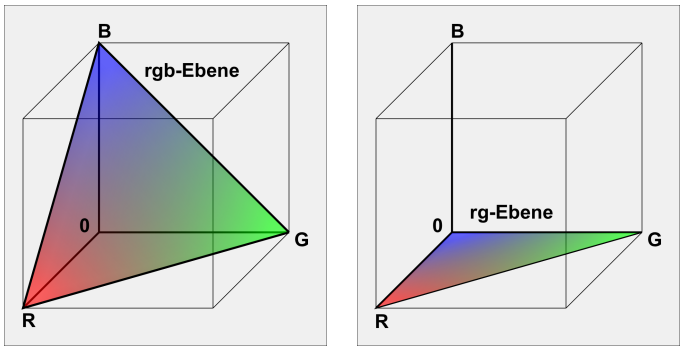
Bevor ein geeigneter Hautfarbdetektor entwickelt werden kann, muss zunächst entschieden werden, welcher Farbraum hierfür besonders geeignet ist. In diesem Abschnitt wird mit Hilfe von Untersuchungen aus der Literatur eine Entscheidung getroffen.

Während die Intensität des von einer diffus reflektierenden Oberfläche wie Haut reflektierten Lichtes von der Geometrie abhängt, z.B. der Entfernung zur Lichtquelle, gilt dies für die Farbe des reflektierten Lichtes in der Regel nicht. Daher eignen sich Farbräume, in denen Helligkeits- und Farbinformationen unkorreliert vorliegen, am Besten. Beispiele sind der HSI-Farbraum und der rg-Farbraum. Es ist allerdings in keinem Farbraum möglich, eine Invarianz gegenüber verschiedenen Beleuchtungsbedingungen zu erreichen. Entscheidende Fragen bei der Wahl des Farbraums sind, ob die Hautfarbverteilungen für verschiedene Beleuchtungsbedingungen entlang einer einfachen Kurve liegen und wie dicht diese Hautfarbverteilungen dann beieinander liegen.

In verschiedenen Publikationen wurde nachgewiesen, dass der helligkeitsnormierte dichromatische rg-Farbraum besonders gut geeignet ist, Hautfarbe in einem großen Bereich unterschiedlicher Beleuchtungsverhältnisse zu repräsentieren [Yang and Waibel, 1998] [Terrillon et al., 2000]. In [Menser and Bräunig, 1999] wird explizit der Einfluss variabler Beleuchtung untersucht, wobei gezeigt wird, dass in anderen Farbräumen keine besseren Ergebnisse erzielt werden, als im rg-Farbraum. Dieser wird wie folgt definiert:

$$r = R/(R+G+B) \tag{2.1}$$
$$g = G/(R+G+B) \tag{2.2}$$
$$b = B/(R+G+B) \tag{2.3}$$

Damit gilt $r+g+b=1$. Es handelt sich also um eine Abbildung aus einem dreidimensionalen in einen zweidimensionalen Raum, bei der die Intensitätsinformationen beseitigt werden und alle Farbwerte auf einer Ebene liegen, siehe Abbildung 2.3. Da sich im rgb-Farbraum ein Farbwert immer aus den beiden anderen berechnen lässt, werden im Folgenden nur die Komponenten r und g verwendet.



(a) (b)

Abbildung 2.3: *(a) rgb-Ebene im RGB-Farbraum. (b) Durch Weglassen der b-Komponente, die aus r und g berechnet werden kann, entsteht der rg-Farbraum.*

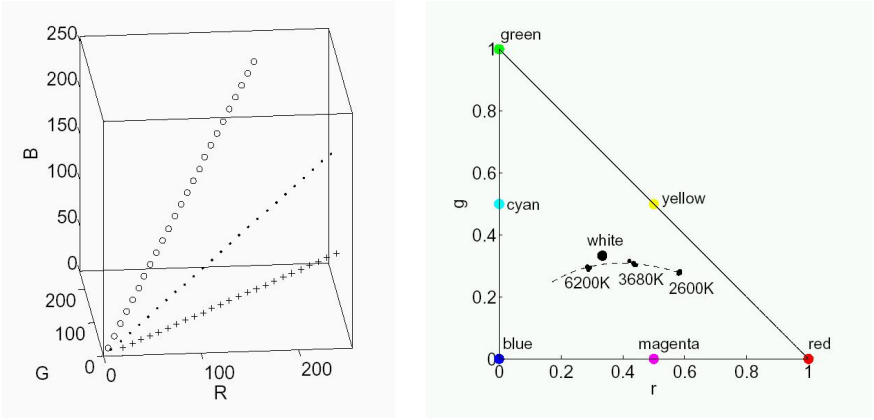
Weniger geeignet sind laut [Störring, 2004] Farbdifferenzräume wie der $Y'C_BC_R$- oder der YES-Farbraum, da hier anders als in Abbildung 2.4 unterschiedliche Beleuchtungsintensitäten nicht auf einen Punkt abgebildet werden. Eine experimentelle Bestätigung hierfür wurde in [Zarit et al., 1999] erbracht.

2.3.1.2 Farbmodell

Als nächstes stellt sich die Frage nach einer geeigneten Repräsentation von Hautfarbe. Die möglichen Modelle können wie in Abbildung 2.5 systematisiert werden. Einige Beispiele werden im folgenden näher beschrieben.

nicht-adaptive nicht-parametrische Modelle: Schiele und Waibel verwenden eine Lookup-Tabelle für Hautfarbe im rg-Farbraum [Schiele and Waibel, 1995].

nicht-adaptive parametrische Modelle: In [Fieguth and Terzopoulos, 1997] wird ein nicht-adaptives parametrisches Modell beschrieben, das aus dem Mittelwert einer Hautfarbregion im RGB-Farbraum besteht. In der Anwendung des Modells werden für jeden Farbkanal die Verhältnisse zum Mittelwert einer zu testenden Hautfarbregion bestimmt. Sind diese



(a) (b)

Abbildung 2.4: *(a) Modellierte Hautfarbe im RGB-Farbraum für drei verschiedene Lichtquellen (2600K, 3680K und 6200K) bei jeweils 25 verschiedenen Beleuchtungsstärken. (b) Im rg-Farbraum fallen die Hautfarben bei einer Beleuchtungsfarbe und unterschiedlichen Intensitäten nahezu auf einen Punkt. Hautfarbe kann somit relativ unabhängig von der Beleuchtungsintensität detektiert werden. Die Abbildung wurde entnommen aus [Störring, 2004].*

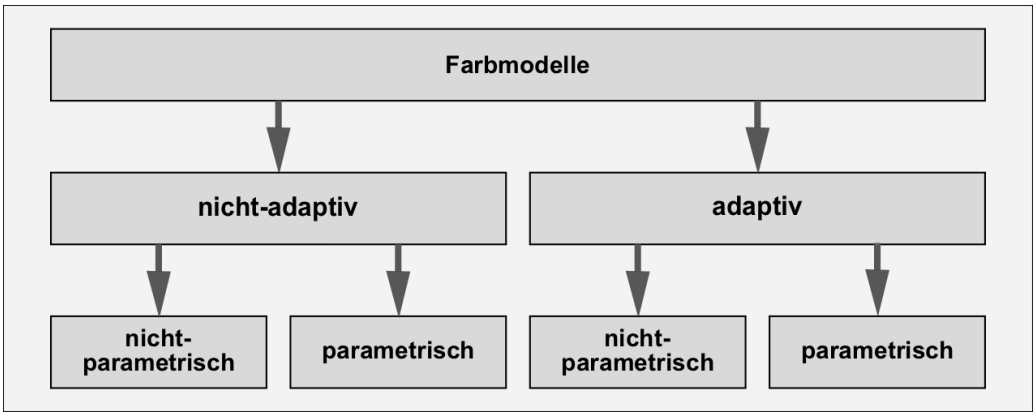


Abbildung 2.5: *Taxonomie von Farbmodellen.*

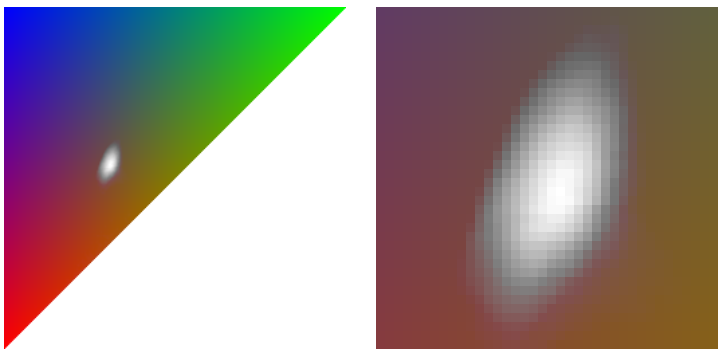
untereinander ähnlich, handelt es sich um Hautfarbe. In [Sanger et al., 1997] wurde die Hautfarbverteilung mit einer unimodalen Gaußfunktion modelliert. Dazu wurden der Mittelwert und die Kovarianzmatrix im rg-Farbraum berechnet. In [Jebara et al., 1998] wird ein dreidimensionales Mixture-of-Gaussian-Modell im RGB-Farbraum verwendet.

adaptive nicht-parametrische Modelle: Um mit veränderlichen Beleuchtungsbedingungen umgehen zu können, wurden adaptive Modelle entwickelt, bei denen wiederum zwischen nicht-parametrischen und parametrischen unterschieden werden kann. In [Bradski, 1998b] wird der Continuously Adaptive Mean Shift-Algorithmus (CAMSHIFT) vorgestellt, der auf dem Hue-Kanal des HSV-Farbraumes angewendet wird. Das Hautfarbmodell besteht aus einem normierten Histogramm, also einer Dichtefunktion. Also handelt es sich bei der Rückprojektion des Histogramms um eine Wahrscheinlichkeitsverteilung des Modells im Bild.

CAMSHIFT detektiert durch die Anwendung des Mean Shift Algorithmus das Maximum der Verteilung, wobei die Zielverteilung dynamisch angepasst wird. In [Soriano et al., 2000] [Soriano et al., 2003] wird ein adaptives Hautfarbhistogramm verwendet, das durch Histogrammrückprojektion aktualisiert wird.

adaptive parametrische Modelle: In [Yang and Waibel, 1996] und [Yang and Waibel, 1998] wurde ein adaptives parametrisches Hautfarbmodell vorgestellt, das aus einer unimodalen zweidimensionalen Gaußfunktion im rg-Farbraum besteht. Der Mittelwert und die Kovarianzmatrix werden unter Verwendung der Hautfarbdetektionen der letzten Zeitschritte adaptiert.

Das in dieser Arbeit verwendete Farbmodell besteht aus einer Lookup-Tabelle mit manuell als Hautfarbe klassifizierten Pixeln im rg-Farbraum. Damit handelt es sich um ein nichtparametrisches und nicht-adaptives Farbmodell. Ein ähnlicher Ansatz, allerdings im YUV-Farbraum, wurde in [Feyrer and Zell, 1999] vorgestellt. Hier wurde auf den rg-Farbraum zurückgegriffen, da dieser besonders gut geeignet ist, Hautfarbe in einem großen Bereich unterschiedlicher Beleuchtungsverhältnisse zu repräsentieren [Yang and Waibel, 1998] [Terrillon et al., 2000]. Das verwendete Farbmodell ist in Abbildung 2.6 zu sehen. Auf eine Approximation der Hautfarbregion mit einem parametrischen Modell wie in [Böhme et al., 1998] oder [Braumann, 2001] wurde bewusst verzichtet, da dies lediglich die Genauigkeit der Abbildung reduziert. Allerdings muss bei dieser Vorgehensweise darauf geachtet werden, dass die Stichprobe hinreichend groß ist, so dass das entstehende Hautfarbmodell keine „Löcher“ aufweist.



(a) (b)

Abbildung 2.6: *(a) Lookup-Tabelle für Hautfarbe im dichromatischen rg-Farbraum. (b) Vergrößerung des Hautfarbmodells. Bei hinreichend großer Stichprobe entsteht ein Farbmodell mit einer glatten Oberfläche.*

Die Varianzen dieser Verteilung hängen zum einen von den Variationen der Hautfarben der Probanden ab, aber auch von den Beleuchtungsschwankungen während der Aufnahme der Trainingsdaten. Nur wenn letztere möglichst gering gehalten werden, kann ein spezifisches Hautfarbmodell erzeugt werden. Dies wurde erreicht, indem die Daten für dieses Modell unter Verwendung des im nächsten Abschnitt beschriebenen automatischen Weißabgleichs aufgenommen wurden. Die Werte des Histogramms wurden auf den Bereich [0..1] normiert. Der Hautfarbdetektor liefert

für jeden Pixel im Panoramabild ein Maß für das Vorhandensein von Hautfarbe $w_{skin}(\mathbf{x})$ im Bereich [0..1], siehe Abbildung 2.7.

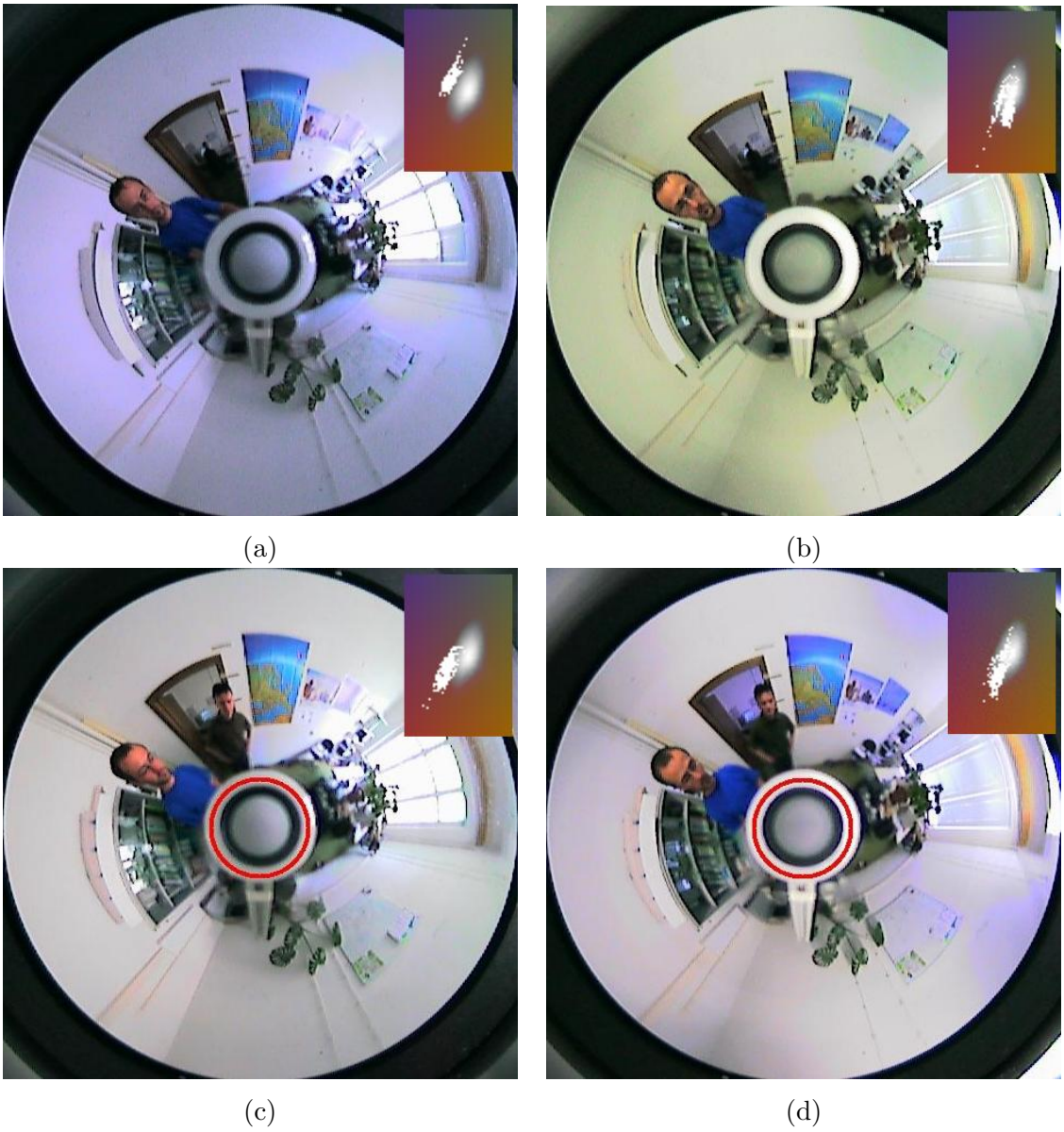
Abbildung 2.7: *Panoramabild und zugehörige Ausgabe des Hautfarbdetektors (dunkle Pixel stehen für große Werte). Neben dem Gesicht der Person werden auch andere Bildregionen als Hautfarbe erkannt, wie z.B. Objekte aus Holz.*

2.3.1.3 Automatischer Weißabgleich

Abbildung 2.8 verdeutlicht die drastischen Unterschiede der Kamerabilder bei verschiedenen Beleuchtungsbedingungen. Um in der Lage zu sein, trotz veränderlicher Beleuchtungsbedingungen Hautfarbe robust detektieren zu können, wurde ein automatischer Weißabgleich für die omnidirektionale Kamera entwickelt [Wilhelm et al., 2003a].

Weißreferenz Dazu wurde die Kamera mit einem weißen Ring ausgestattet, der als Weißreferenz im Bild erscheint. Abbildung 2.9 zeigt die Kamera mit omnidirektionalem Objektiv und Weißreferenz und ein mit dieser Kamera aufgenommenes Bild, das die Referenz auf einem inneren Radius enthält. Die Oberfläche der Weißreferenz weist eine leichte konvexe Krümmung auf, damit auch von der Seite kommendes Licht erfasst wird.

Regelkreis Abbildung 2.10 zeigt den Aufbau des Regelkreises. Es wird der Mittelwert für R, G und B über alle Pixel berechnet, die sich innerhalb der Weißreferenz befinden. Dieser Mittelwert wird dann in den YUV-Farbraum transformiert. Mit der Differenz zu den Sollwerten $U = 0$ und $V = 0$ (weiß) werden mit Hilfe von zwei separaten PID-Reglern die Stellgrößen für die Verstärkungsfaktoren für U und V für die verwendete Digitalkamera (SONY DFW VL500) bestimmt. Außerdem wird der Mittelwert von Y verwendet, um die Iris der Kamera so zu steuern, dass eine annähernd gleich bleibende Helligkeit im Bild erreicht wird. Die einzelnen Komponenten des Regelkreises sowie die Vorgehensweise für die Parametrierung der Regler werden im Anhang A.1 beschrieben.



(a) (b)

(c) (d)

Abbildung 2.8: *Bilder der omnidirektionalen Kamera, aufgenommen bei (a) weitgehend natürlicher Beleuchtung ohne Weißabgleich, (b) weitgehend künstlicher Beleuchtung ohne Weißabgleich, (c) weitgehend natürlicher Beleuchtung mit Weißabgleich und (d) weitgehend künstlicher Beleuchtung mit Weißabgleich. Ohne Weißabgleich hat das Bild bei natürlicher Beleuchtung einen deutlichen Blau- und bei künstlicher Beleuchtung einen deutlichen Grünstich. In der rechten oberen Ecke jedes Bildes ist die Hautfarbverteilung des Gesichtes der Person im rg-Farbraum dargestellt. Besonders bei der natürlichen Beleuchtung wird deutlich, wie die Hautfarbverteilung durch den Weißabgleich weiter in Richtung des Farbmodells verschoben wird.*

(a) (b)

Abbildung 2.9: *(a) Weißreferenz zwischen Kamera und Objektiv. (b) Mit dieser Kamera aufgenommenes Bild, wobei die Weißreferenz in der Nähe des Zentrums des Bildes erscheint. Dieser Bildbereich entspricht dem Fußboden in der unmittelbaren Umgebung des Roboters und ist für die Nutzerdetektion irrelevant.*

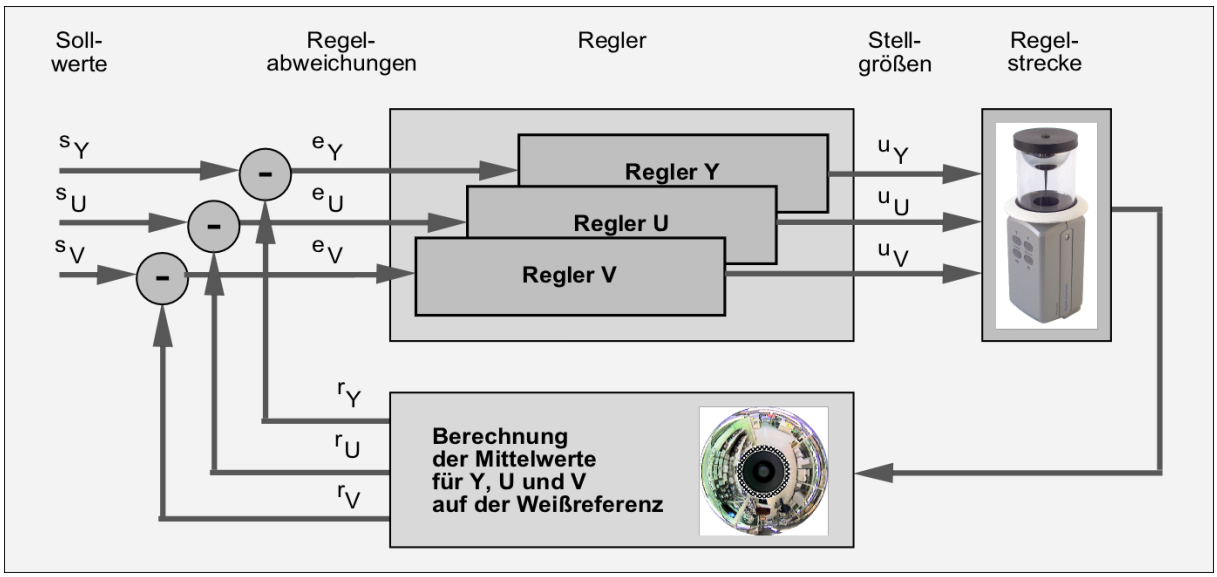


Abbildung 2.10: *Struktur des digitalen Regelkreises für den automatischen Weißabgleich. Aus den Mittelwerten für Y, U und V aller Pixel, die auf der Weißreferenz liegen und den Sollgrößen $s_U = 0$, $s_V = 0$ und $s_Y = 0$ werden die Regelabweichungen e_U, e_V und e_Y berechnet. Drei separate PID-Regler berechnen die Stellgrößen für den Weißabgleich u_U, u_V und die Iris u_Y der Kamera.*

2.3.2 Bewegung

Als zweites Merkmal für die Personendetektion dient die Bewegung. Dazu wird ein Differenzbild zwischen zwei aufeinander folgenden Panoramabildern berechnet. Ist die Länge des RGB-Differenzvektors für einen Pixel größer als ein Schwellwert, so wird für diesen Pixel Bewegung angenommen. Die Berechnung wird nur für jeden zweiten Pixel durchgeführt und für alle Pixel in der Nachbarschaft eines Pixels mit Bewegung wird ebenfalls Bewegung angenommen. Da hier eine Eigenbewegung der Kamera nicht kompensiert wird, kann die Bewegungsdetektion nur bei stehendem Roboter angewendet werden. Dies stellt keine große Einschränkung dar, da ein Interaktionszyklus in der Regel bei stehendem Roboter beginnt. Da die Berechnung der Bewegungsinformation auf dem Bild der omnidirektionalen Kamera erfolgt, kann der Kopf des Roboters kontinuierlich dem getrackten Objekt folgen. Auf diese Weise müssen die Bilder nicht mit temporär stationären Frontalkameras aufgenommen werden wie in [Feyrer and Zell, 1999]. Der Bewegungsdetektor liefert für jeden Pixel im Panoramabild einen Wert $w_{movement}(\mathbf{x})$ von 1, falls dieser Bewegung enthält und ansonsten 0, siehe Abbildung 2.11.

Abbildung 2.11: *Panoramabild und zugehörige Ausgabe des Bewegungsdetektors (dunkle Werte stehen für Bewegung). Je nach Schwellwert wird auch Pixelrauschen im Bild als Bewegung detektiert.*

2.3.3 Entfernung

Mit der sonarbasierten Komponente des Auffälligkeitssystems wird die Entfernung zum nächsten Objekt für 24 Bereiche um den Roboter gemessen. Da die Rohdaten der Sonarsensoren sehr stark rauschen und zudem von der Orientierung und dem Material der Objekte um den Roboter abhängen, werden diese wie folgt vorverarbeitet: ungültige Messwerte, d.h. Entfernungen größer $22.5m$, werden durch den jeweiligen vorhergehenden Messwert ersetzt. Eine räumliche Tiefpassfilterung benachbarter Messwerte und eine zeitliche Tiefpassfilterung aufeinander folgender Messwerte dienen der Rauschreduktion. Die Entfernungswichtung liefert für jeden Bereich c einen Wert $w_{distance}(c)$, der umgekehrt proportional zur gemessenen Entfernung ist,

$$w_{distance}(c) = \frac{1}{1 + e^{s(d_{sonar}(c) - d_{max})}}, \tag{2.4}$$

wobei es sich bei $d_{sonar}(c)$ um den vorverarbeiteten Messwert für den Bereich c im Scan handelt, bei d_{max} um die maximal zu berücksichtigende Entfernung $(2.0m)$ und bei s um den Anstieg der Funktion, siehe Abbildung 2.12. Die Position des Maximums im resultierenden Gewichtsvektor entspricht der Richtung zum nächstgelegenen Objekt. Falls auf einem System keine Entfernungssensoren zur Verfügung stehen, wie beim Standalone-Arbeitsplatz MIMIR, werden die Gewichte für alle Richtungen auf 1 gesetzt.

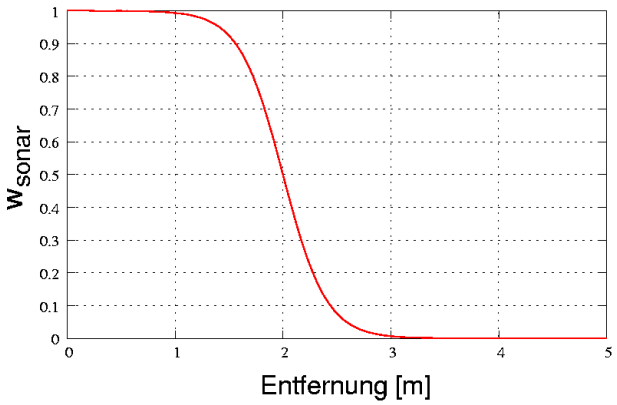


Abbildung 2.12: *Abhängigkeit der Wichtung $w_{distance}(c)$ von der gemessenen Entfernung $d_{sonar}(c)$, siehe Gleichung 2.4. Der Anstieg s der Funktion beträgt hier 5 und die maximale Entfernung d_{max} beträgt 2.*

Abbildung 2.13: *Panoramabild und zugehörige Sonargewichte (dunkle Werte stehen für Bewegung). Neben der Person befindet sich auch der Schrank auf der rechten Seite in der Nähe des Roboters.*

2.4 Fusion der Auffälligkeitskarten

Da sowohl das Panoramabild als auch die Entfernungswichtung eine 360°-Beschreibung der Umgebung des Roboters darstellen, ist es möglich, jeder Position $\mathbf{x}$ im Bild ein Sonar-Gewicht $w_{distance,t}(c)$ an der Position c im Scan zuzuordnen. Auf folgende Weise werden die Auffälligkeitskarten verknüpft:

$$w_t(\mathbf{x}) = (w_{skin,t}(\mathbf{x}) + w_{movement,t}(\mathbf{x}))\, w_{distance,t}(c) + w_{face} \tag{2.5}$$

Wenn sich eine Person nicht bewegt, ist $w_{movement,t}$ Null. Dieser Wert wird neben der Hautfarbe als zusätzlicher Hinweis verwendet und entsprechend zum Hautfarbwert addiert. Da nur Personen detektiert werden sollen, die sich in unmittelbarer Nähe des Roboters befinden, geht das Sonargewicht multiplikativ ein. Falls keine Sonarmessungen zur Verfügung stehen, wird dieser Wert auf 1 gesetzt. Bei w_{face} handelt es sich um eine Rückkopplung der Gesichtsdetektion aus dem fovealen Vision-System, welche in Kapitel 3 vorgestellt wird. Auf diese Weise wird die Bewertung einer auffälligen Bildregion zusätzlich verstärkt, wenn es sich tatsächlich um eine Person handelt. Abbildung 2.14 veranschaulicht die Fusion der Auffälligkeitskarten.

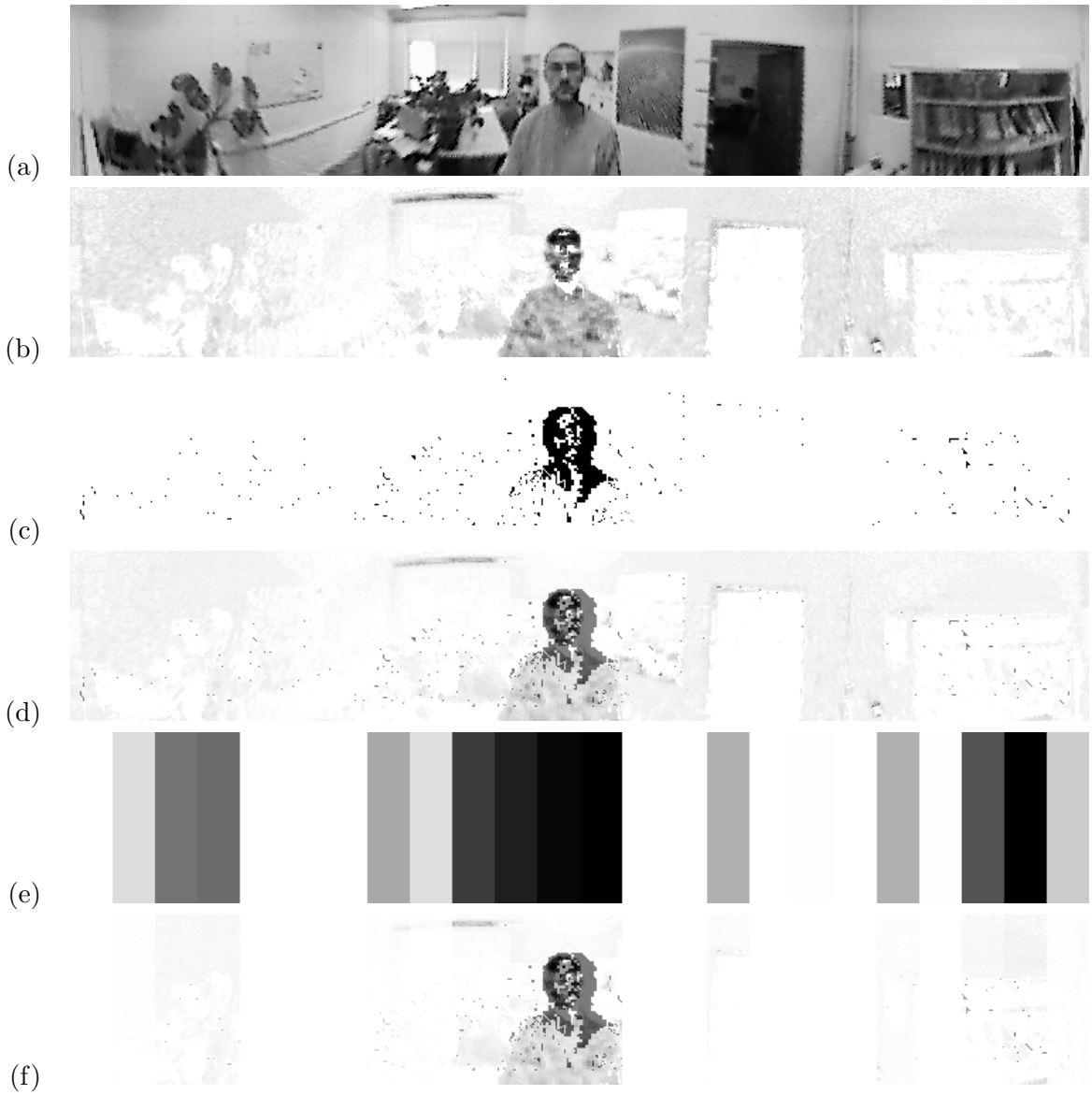


Abbildung 2.14: *Fusion der Merkmale für das Aufmerksamkeitssystem. (a) Panoramabild (b) Hautfarbe (c) Bewegung (d) Additive Überlagerung von Hautfarbe und Bewegung (e) Entfernung (f) Durch die multiplikative Verknüpfung von d und e werden zum einen Personen unterdrückt, die zu weit vom Roboter entfernt sind und deshalb für eine Interaktion nicht in Frage kommen. Zum anderen werden viele Falsch-Positiv-Hypothesen beseitigt.*

2.5 Bildung von Nutzerhypothesen

Hier geht es darum, aus der Auffälligkeitskarte Hypothesen für die Anwesenheit von Personen zu bilden und über die Zeit zu verfolgen. Wie im Abschnitt 2.2 beschrieben, existieren zu diesem Zweck eine Reihe von Verfahren. In dieser Arbeit fiel die Auswahl auf das so genannte CONDENSATION-Tracking [Isard and Blake, 1998] [Isard and Blake, 1996], da es im Vergleich zu neuronalen Feldern sehr rechenzeit-effektiv ist. Um mehrere Hypothesen verwalten zu können, wurden zwei Multi-Target-Tracker implementiert und vergleichend untersucht. Im folgenden Abschnitt wird zunächst der CONDENSATION-Algorithmus, der die Basis dieser Multi-Target-Tracker bildet, vorgestellt.

2.5.1 CONDENSATION-Algorithmus

Die Aufgabe der Berechnung der Wahrscheinlichkeit, ob sich an einem bestimmten Bildpunkt eine auffällige Struktur befindet, und die Verfolgung der resultierenden Dichtefunktion über die Zeit t, wird durch eine Approximation der Dichtefunktion $p(\mathbf{x}_t)$ durch eine relativ kleine Anzahl von N Samples $\mathbf{s}_t^{(i)}$ realisiert:

$$p(\mathbf{x}_t) \propto \left\{ \mathbf{s}_t^{(i)} = \left\langle \mathbf{x}_t^{(i)}, w_t^{(i)} \right\rangle | i = 1, .., N \right\} \tag{2.6}$$

wobei jedes Sample $\mathbf{s}_t^{(i)}$ durch eine Position $\mathbf{x}_t^{(i)}$ und ein Gewicht $w_t^{(i)}$ charakterisiert wird. Ein Aktualisierungsschritt des rekursiven Filters läuft wie folgt ab [Isard and Blake, 1998]:

$$P(\mathbf{s}_t|\mathbf{Y}_t) = \underbrace{P(\mathbf{y}_t|\mathbf{s}_t)}_{\text{Beobachtungsmodell}} \int \underbrace{P(\mathbf{s}_t|\mathbf{s}_{t-1})}_{\text{Bewegungsmodell}} \cdot P(\mathbf{s}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{s}_{t-1} \tag{2.7}$$

Am Anfang steht eine Menge von Samples $\mathbf{s}$, welche die a posteriori Dichte $p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})$ aus dem Zeitschritt $t-1$ beschreibt, wobei $\mathbf{Y}_{t-1}$ die Menge aller bisherigen Beobachtungen $\{\mathbf{y}_0, ..., \mathbf{y}_{t-1}\}$ ist. Aus den Samples $\mathbf{s}^{(i)}$ werden mit der Wahrscheinlichkeit $w_t^{(i)}$ neue Samples $\mathbf{s}'$ erzeugt. D.h. ein Sample $\mathbf{s}^{(i)}$ mit einem hohen Gewicht $w_t^{(i)}$ erzeugt mit höherer Wahrscheinlichkeit Nachkommen in $\mathbf{s}'$. Die so entstandenen Samples werden entsprechend des Bewegungsmodells $P(\mathbf{s}_t|\mathbf{s}_{t-1})$ propagiert. Dieses besteht aus einer stochastischen Komponente für unvorhergesehene Bewegungen der Person und aus einer deterministischen Komponente für bekannte Bewegungen des Roboters. Die neue Sample-Menge $\mathbf{s}'$ beschreibt die a priori Dichte $p(\mathbf{x}_t|\mathbf{Y}_{t-1})$. Im letzten Schritt werden die neuen Sample-Gewichte $w_t^{(i)}$ entsprechend der Beobachtungen im aktuellen Zeitschritt $P(\mathbf{y}_t|\mathbf{s}_t)$ zugewiesen. Die Gewichte der Sample-Menge werden schließlich auf 1 normiert. Abbildung 2.15 veranschaulicht den Ablauf des CONDENSATION-Algorithmus.

Wie bereits erwähnt, besteht der Vorteil von CONDENSATION darin, dass die Dichtefunktion $p(\mathbf{x}_t)$ nicht vollständig berechnet werden muss, sondern mit nur N Samples approximiert wird. Im Falle der Feature-Extraktion auf dem Panoramabild mit 720×106 Pixeln wird durch die Approximation mit nur 500 Samples eine Reduktion auf 0.655% erreicht. Aufgrund der oft

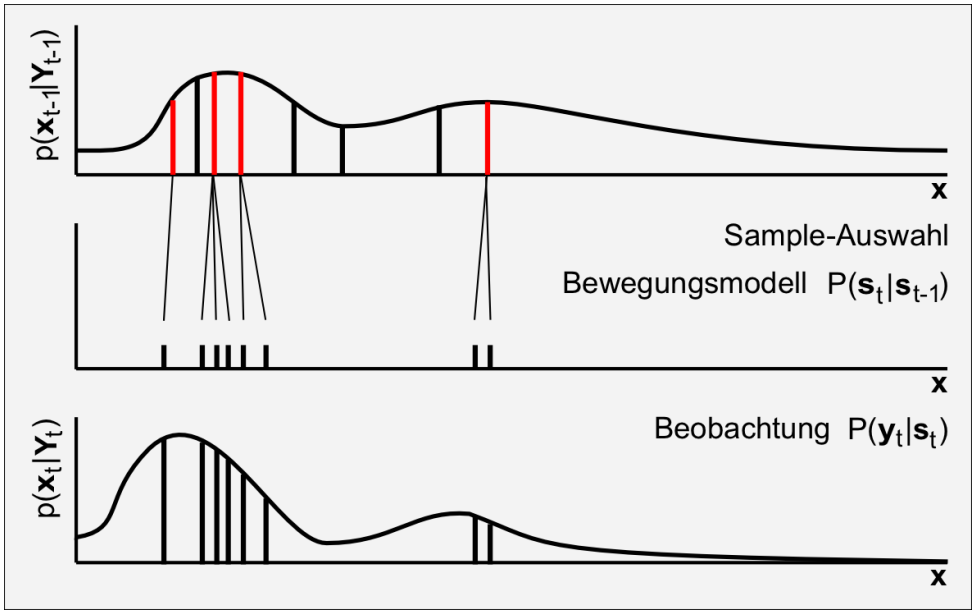


Abbildung 2.15: *Veranschaulichung des CONDENSATION-Algorithmus. Ausgangspunkt ist die Sample-Menge* $\mathbf{s}$*, die die Dichtefunktion zum Zeitpunkt* $t-1$ *repräsentiert. Aus dieser wird wahrscheinlichkeitsbasiert eine neue Sample-Menge* $\mathbf{s}'$ *erzeugt, dergestalt, dass Samples mit einem hohen Gewicht mit größerer Wahrscheinlichkeit Nachkommen in* $\mathbf{s}'$ *erzeugen dürfen. Auf diese neuen Samples wird das Bewegungsmodell angewendet. Schließlich werden den Samples aus* $\mathbf{s}'$ *neue Gewichte entsprechend der Beobachtung zum Zeitpunkt* t *zugewiesen.*

unvorhersagbaren Bewegungen von Personen muss das stochastische Bewegungsmodell eine hinreichend große Varianz in x-Richtung aufweisen, damit Personen nicht verloren werden. Die Anzahl von 500 Samples hat sich als sinnvolle Größe erwiesen.

2.5.2 Multi-Target-Tracking

Obwohl es für den Serviceroboter eigentlich immer nur ein interessantes zu trackendes Objekt gibt, nämlich seinen aktuellen Nutzer, sollte das Tracking-System in der Lage sein, mehrere Objekte gleichzeitig zu verfolgen. Meldet sich der aktuelle Nutzer ab oder verlässt den Einsatzbereich des Roboters, soll dieser sich unmittelbar einer anderen Person zuwenden. Im Vergleich zum Tracken eines einzelnen Objektes ergeben sich beim Multi-Target-Tracking eine ganze Reihe von neuen Problemstellungen:

Sample-Verarmung: CONDENSATION-Tracker sind in der Lage, beliebige multimodale Dichtefunktionen zu approximieren und können kurzfristig ausbleibende Beobachtungen kompensieren. Damit sind sie aber noch nicht geeignet, mehrere Objekte gleichzeitig zu verfolgen, denn in der Realität bildet ein CONDENSATION-Tracker sehr schnell eine unimodale Verteilung über dem stärksten Stimulus, da hier die meisten Samples neu entstehen und die Anzahl der Samples auf anderen Regionen immer kleiner wird. Um mit diesem Problem der Sample-Verarmung umgehen zu können, wurden eine Reihe von Erweiterungen des CONDENSATION-Algorithmus vorgestellt [MacCormick and Blake, 1999]

[Tweed and Calway, 2002] [Tao et al., 1999]. In [Tweed and Calway, 2002] werden Cluster in der Dichtefunktion gebildet, wobei angenommen wird, dass in jedem Cluster ein Objekt getrackt wird. Dann erfolgt eine Skalierung der Sample-Gewichte derart, dass das größte Gewicht innerhalb jedes Clusters 1 ist. In [Tao et al., 1999] wird eine so genannte Konfigurationsabdeckung berechnet, die angibt, wie gut eine Sample-Verteilung die gesamte Beobachtung abdeckt.

Verdeckungen: Ein anderes Problem besteht im Umgang mit sich gegenseitig verdeckenden Objekten. Zum einen gibt es den Ansatz, solche Verdeckungen explizit zu modellieren und die zugehörigen Objekte weiterzutracken [MacCormick and Blake, 1999] [Tweed and Calway, 2002]. Dabei besteht die Gefahr, dass mehrere Tracker für lediglich ein Objekt in der Szene verwendet werden. In einer anderen Lösungsstrategie wird auf die Trennung mehrerer sich verdeckender Objekte verzichtet. Im Zeitpunkt der Verdeckung wird die Beobachtung also nur einem Objekt zugeschrieben, so lange, bis wieder separate Beobachtungen gemacht werden. So wird in [Tao et al., 1999] durch die Konfigurationskompaktheit verhindert, dass mehrere Objekte an einer Bildposition getrackt werden, siehe Abschnitt 2.5.2.1.

Datenassoziation: Das Problem der Datenassoziation ist eng mit dem der Verdeckung von Objekten verbunden. Dabei geht es darum, die Einzelbeobachtungen jeweils einer Hypothese zuzuordnen, was insbesondere dann zum Problem wird, wenn zwei Objekte nahe beieinander liegen bzw. sich gegenseitig verdecken. Eine übliche Methode, mit diesem Problem umzugehen, sind die Joint Probabilistic Data Association Filters (JPDAF) [Schulz et al., 2003] [Rasmussen and Hager, 2001]. JPDAFs realisieren eine Art Ausschlussprinzip, welches verhindert, das zwei oder mehr Tracker auf dem selben Objekt liegen, indem Wahrscheinlichkeiten für Assoziationen zwischen Targets und Messungen berechnet werden.

Im Rahmen dieser Arbeit wurde auf eine Behandlung von Verdeckungen und eine Datenassoziation verzichtet, da es im Auffälligkeitssystem nur darum geht, saliente Bildstrukturen zu verfolgen und nicht, diese auseinander zu halten. Für den Zeitpunkt einer Verdeckung ist eine Modellierung als eine auffällige Bildstruktur hinreichend. Für eine Realisierung eines Multi-Target-Trackers wurde der Ansatz aus [Tao et al., 1999] reimplementiert und mit einer eigenen Entwicklung vergleichend untersucht, bei der mehrere separate CONDENSATION-Tracker verwendet werden. Ein erster solcher Vergleich, allerdings auf weniger realistischen Daten, wurde in [Wilhelm and Martin, 2004] veröffentlicht. In den nächsten beiden Abschnitten werden diese beiden Verfahren vorgestellt. Im Abschnitt 2.5.3 folgen Untersuchungen und Ergebnisse.

2.5.2.1 Hochdimensionale Sample-Konfigurationen

In diesem Abschnitt wird ein alternatives Verfahren zum Multi-Objekt-Tracking basierend auf der Arbeit von Tao et al. [Tao et al., 1999] vorgestellt. Hier wird eine so genannte Sample-

Konfiguration eingeführt, die eine Kombination aller beobachteten Objekte in einem Bild beschreibt. Eine solche Konfiguration kann wie folgt ausgedrückt werden:

$$\mathbf{c}_t^{(i)} = \left\langle \left\{ \mathbf{x}_{t,1}^{(i)}, \mathbf{x}_{t,2}^{(i)}, ..., \mathbf{x}_{t,m}^{(i)} \right\}, w_t^{(i)} \right\rangle | i = 1, .., N; j = 1, .., M \tag{2.8}$$

wobei $\mathbf{x}_{t,j}$ der Zustand des Objektes j zum Zeitpunkt t (z.B. die Koordinaten einer Person im Bild) und M die Anzahl der Samples in der Konfiguration ist. Eine solche hochdimensionale Sample-Konfiguration $\mathbf{c}_t$ beschreibt also den Zustand von M Objekten in einer einzigen Variablen. Ziel des Verfahrens ist es, die a posteriori Wahrscheinlichkeit der Konfiguration mit Hilfe eines geeigneten Bayes-Filters zu bestimmen:

$$P(\mathbf{c}_t|\mathbf{Y}_t) = \underbrace{P(\mathbf{y}_t|\mathbf{c}_t)}_{\text{Konfigurationsgüte}} \int \underbrace{P(\mathbf{c}_t|\mathbf{c}_{t-1})}_{\text{Konfigurationsdynamik}} \cdot P(\mathbf{c}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{c}_{t-1} \tag{2.9}$$

Der Term $P(\mathbf{y}_t|\mathbf{c}_t)$ stellt die Konfigurationsgüte dar und ist ein Maß dafür, wie gut die aktuelle Beobachtung $\mathbf{y}_t$ durch die Konfiguration $\mathbf{c}_t$ beschrieben wird. Dieser Term entspricht dem Beobachtungsmodell im klassischen CONDENSATION-Algorithmus, vgl. Gleichung 2.7. Die Konfigurationsdynamik $P(\mathbf{c}_t|\mathbf{c}_{t-1})$ beschreibt, wie sich eine Konfiguration $\mathbf{c}_{t-1}$ zur Konfiguration $\mathbf{c}_t$ verändert. Im klassischen CONDENSATION-Algorithmus ist dies das Bewegungsmodell. In den beiden folgenden Abschnitten werden diese beiden Terme ausführlicher beschrieben.

Konfigurationsgüte Die Konfigurationsgüte $P(\mathbf{y}_t|\mathbf{c}_t)$ ist eine komplexe und unter Umständen auch schwierig zu berechnende Verteilung. In unserer Arbeit haben wir daher die gleiche Dekomposition der Konfiguration wie in [Tao et al., 1999] eingesetzt. Tao et al. haben dazu eine Energiefunktion definiert, die erwünschten Konfigurationen hohe Werte und nicht erwünschten Konfigurationen niedrige Werte zuweist. Die Funktion besteht aus drei Faktoren:

Objektwahrscheinlichkeit $\lambda(\mathbf{c}_t)$: Dieser Faktor gibt an, wie gut die Objekte der Konfiguration $\mathbf{c}_t$ mit Hilfe der aktuellen Beobachtung $\mathbf{y}_t$ erklärt werden können. Dazu werden die Gewichte der einzelnen Samples $\mathbf{x}_{t,j}$ der Sample-Konfiguration genutzt. Das Gewicht $w(\mathbf{x}_{t,j})$ entspricht der Beobachtung an der Position des Samples $\mathbf{x}_{t,j}$, siehe Gleichung 2.5.

$$\lambda(\mathbf{c}_t) = \left(\prod_{j=1}^{m} w(\mathbf{x}_{t,j}) \right)^{\frac{1}{m}} \tag{2.10}$$

Konfigurationsabdeckung $\gamma(\mathbf{c}_t)$: Dieser Term gibt an, wie gut eine Konfiguration die gesamte Beobachtung abdeckt. Er wird wie folgt berechnet:

$$\gamma(\mathbf{c}_t) = \frac{\left| A \cap \left(\bigcup_{j=1}^{m} B_j \right) \right| + b}{|A| + b} \tag{2.11}$$

wobei A die Menge aller Einzelbeobachtungen und B_j der Zustand des Objektes j der Konfiguration ist. Die Schnittmenge $A \cap \left(\bigcup_{j=1}^{m} B_j\right)$ ist somit die Menge aller Einzelbeobachtungen, die auch durch die zu bewertende Konfiguration abgedeckt werden. Der Faktor γ geht dann gegen 1, wenn alle Einzelbeobachtungen auch durch die entsprechende Konfiguration erfasst werden. Wenn dagegen keine einzige Beobachtung erfasst wird, geht γ gegen 0. Die kleine positive Konstante b verhindert eine Division durch 0. Um γ berechnen zu können, muss die Anzahl der Objekte in der Szene bekannt sein. Um diese zu schätzen, wird das Bild stark unterabgetastet und für jeden Pixel dieses unterabgetasteten Bildes die Hautfarbzugehörigkeit mit einem Schwellwert ermittelt. Die Summe dieser Zugehörigkeiten ist eine Schätzung für die Größe A.

Konfigurationskompaktheit $\xi(\mathbf{c}_t)$: Dieser Faktor ist definiert als das Verhältnis der abgedeckten Einzelbeobachtungen zur Komplexität der Konfiguration $\mathbf{c}_t$. Sie wird wie folgt berechnet:

$$\xi(\mathbf{c}_t) = \frac{\left|A \cap \left(\bigcup_{j=1}^{m} B_j\right)\right| + d}{\left|\bigcup_{j=1}^{m} B_j\right| + a} \tag{2.12}$$

Dieser Wert geht dann gegen 1, wenn eine effektive Abdeckung der Beobachtung erfolgt. Wenn in einer Konfiguration zu viele Samples verwendet werden, um eine bestimmte Menge von Einzelbeobachtungen zu repräsentieren, wird ξ klein. d ist eine positive Konstante, die so gewählt wird, dass wenn $|A| = 0$, Konfigurationen mit weniger Samples eine höhere Bewertung bekommen. Der kleine positive Wert a verhindert eine Division durch 0.

Letztendlich wird die Güte einer Konfiguration $\mathbf{c}_t$ approximiert durch:

$$P(\mathbf{y}_t|\mathbf{c}_t) \approx \lambda(\mathbf{c}_t) \cdot (\gamma(\mathbf{c}_t) \cdot \xi(\mathbf{c}_t))^{\delta} \tag{2.13}$$

wobei δ eine positive Konstante ist, die die relative Wichtigkeit der letzten beiden Faktoren beeinflusst. Die Werte $P(\mathbf{y}_t|\mathbf{c}_t)$ werden normiert und dienen im nachfolgenden Zeitschritt als Gewichte w_t für die Sample-Konfigurationen im Resampling-Schritt.

Konfigurationsdynamik Das Bewegungsmodell entspricht dem des normalen CONDENSATION-Algorithmus und besteht aus einer stochastischen und einer deterministischen Komponente für die Positionsänderung der Samples. Für das Einfügen bzw. Löschen von Samples wird das Bewegungsmodell um zwei Wahrscheinlichkeiten erweitert. Mit der Wahrscheinlichkeit α wird ein neues Sample in eine Konfiguration eingefügt, wobei dessen Position zufällig initialisiert wird. Mit der Wahrscheinlichkeit β wird ein Sample aus einer Konfiguration gelöscht. In der verwendeten Implementierung werden jeweils konstante Werte verwendet ($\alpha = 0.01$ und $\beta = 0.01$). Mit α und β wird die Wahrscheinlichkeit des Erscheinens bzw. Verschwindens von Objekten in der Szene modelliert. Dieses erweiterte Bewegungsmodell wird als Konfigurationsdynamik bezeichnet, siehe Gleichung 2.9.

Schätzung der Objektanzahl und -positionen Die geschätzte Anzahl von Objekten zum Zeitpunkt t kann wie folgt berechnet werden:

$$\sum_{i=1}^{N} \left|\mathbf{c}_t^{(i)}\right| w_t^{(i)} \quad \text{, mit} \quad 0 \leq \left|\mathbf{c}_t^{(i)}\right| \leq M \tag{2.14}$$

wobei $\left|\mathbf{c}_t^{(i)}\right| \in N$ die Anzahl der Samples $\mathbf{x}_{t,j}^{(i)}$ der Konfiguration $\mathbf{c}_t^{(i)}$ ist. Die Position der einzelnen Objekte im Bild wird wie folgt geschätzt: da die Samples in den Konfigurationen nicht nach ihrer Objektzugehörigkeit geordnet vorliegen, wird versucht, diese entsprechend ihrer räumlichen Lage einander zuzuordnen. Hierzu wird der Abstand θ_d definiert, den zwei zu einem Objekt gehörende Samples nicht überschreiten dürfen.

2.5.2.2 Mehrere Einzel-Tracker

Das alternative selbst entwickelte Verfahren verwendet mehrere voneinander unabhängige CONDENSATION-Tracker, von denen jeder ein Objekt in der Szene verfolgt. Die Anzahl der Samples pro Tracker und die maximale Anzahl der verwendeten Tracker ist dabei prinzipiell beliebig. Alle verwendeten Tracker werden bei diesem Verfahren in einer Liste verwaltet, wobei der Tracker mit dem größten mittleren Sample-Gewicht vor der Normierung als aktuelle Nutzerhypothese verwendet wird. In diesem Fall sind die Fragen zu klären, wann ein Tracker eingefügt und wann ein Tracker gelöscht werden muss.

Einfügen eines neuen Trackers Befindet sich ein Objekt im Bild, dass noch nicht von einem Tracker erfasst wird, wird an der entsprechenden Bildstelle ein neuer Tracker eingefügt. Hierzu wird, wie beim Verfahren von Tao, die Auffälligkeitskarte unterabgetastet. Liegt der mittlere Auffälligkeitswert für ein Rasterelement über einem Schwellwert θ_i und wird dieses Rasterelement noch nicht durch einen Tracker erfasst, wird dort ein neuer Tracker initialisiert. Dieses Vorgehen entspricht der Berechnung der *Konfigurationsabdeckung* $\gamma(\mathbf{c}_t)$ nach Tao. Allerdings wird hier nach der Detektion eines neuen auffälligen Bildbereichs zielstrebig an genau dieser Stelle ein Tracker platziert, während bei Tao Konfigurationen, die einen solchen Hautfarbbereich nicht erfassen, schlechter bewertet werden. Ein neuer Hautfarbbereich wird bei Tao erst dann erfasst, wenn in einer Konfiguration mit der Wahrscheinlichkeit α zufällig ein neues Sample an der entsprechenden Position im Bild erzeugt wird.

Löschen von vorhandenen Trackern Ein Tracker wird in den folgenden Fällen aus der Liste gelöscht:

1 Die mittlere Sample-Wichtung eines Trackers unterschreitet einen Mindestwert θ_e. Dieser Fall tritt z.B. dann ein, wenn sich eine getrackte Person aus dem Umfeld des Roboters entfernt und Hautfarb- und Distanzwichtung entsprechend kleiner werden oder verschwinden. Dieses Maß entspricht der *Objektwahrscheinlichkeit* $\lambda(\mathbf{c}_t)$ bei Tao, wobei hier für die

Berechnung das arithmetische Mittel verwendet wird, da einzelne abweichende Samples das Gesamtergebnis nicht zu sehr beeinflussen sollen.

$$\lambda_j = \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{x}_{t,j}^{(i)}) \tag{2.15}$$

2 Der Mindestabstand θ_d zwischen den Schwerpunkten zweier Tracker wird unterschritten. Damit ein Objekt nicht von mehreren Trackern verfolgt wird, sobald sich zwei Tracker zu nahe kommen, wird derjenige mit der kleineren mittleren Sample-Wahrscheinlichkeit gelöscht. So wird sichergestellt, dass die aktuelle Nutzerhypothese nicht durch andere Tracker verdrängt werden kann. Dieser Fall wird bei Tao nur indirekt über die *Konfigurationsabdeckung* $\gamma(\mathbf{c}_t)$ berücksichtigt. Bei der Unterabtastung des Bildes und der Schätzung der Anzahl der Objekte werden zwei sehr nahe beieinander liegende Objekte als eines gezählt. In diesem Fall werden Konfigurationen mit mehr Samples schlechter bewertet und sterben nach kurzer Zeit aus.

Schätzung der Objektanzahl und -positionen Die geschätzte Anzahl an Objekten in der Szene entspricht der Anzahl der zu jedem Zeitpunkt verwendeten Tracker. Die geschätzten Objektpositionen $\mathbf{x}_j$ entsprechen bei diesem Verfahren dem Schwerpunkt der Sample-Verteilungen der einzelnen Tracker j. Diese werden wie folgt berechnet:

$$\mathbf{x}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{t,j}^{(i)} w_{t,j}^{(i)} \tag{2.16}$$

2.5.3 Vergleichende Untersuchungen

Um die Leistungsfähigkeit der beiden Multi-Target-Tracker gegenüberzustellen, wurden die Genauigkeit der Schätzung der Personenanzahl, die Treffergenauigkeit und der Zeitbedarf als Vergleichskriterien herangezogen. Abbildung 2.16 zeigt einige Bilder des für den Vergleich verwendeten Datensatzes. Bei diesem Test wurde die Auffälligkeitskarte lediglich aufgrund der Hautfarbe gebildet. Zu Beginn befindet sich eine Person in der Szene. Nacheinander kommen zwei weitere Personen hinzu, wobei die am Anfang in der Szene befindliche Person kurzzeitig verdeckt wird. Es sollte hierbei untersucht werden, inwieweit die beiden Tracking-Verfahren in der Lage sind, zu jedem Zeitpunkt die Anzahl der Personen und deren Position in der Szene korrekt zu schätzen.

Personenanzahl Abbildung 2.17 zeigt die von den beiden Tracking-Verfahren geschätzte Personenzahl auf dem Testdatensatz. Prinzipiell sind beide Verfahren in der Lage, alle in der Sequenz auftauchenden Personen zu erfassen. Es fällt auf, dass das Verfahren nach Tao die korrekte Anzahl schneller erfasst, wenn eine zusätzliche Person die Szene betritt.

Positioniergenauigkeit Die Positioniergenauigkeit der Tracker wurde wie in Abbildung 2.18 erläutert ermittelt. Dabei wurden die Anzahl der Fehldetektionen und die Genauigkeit der Tref-

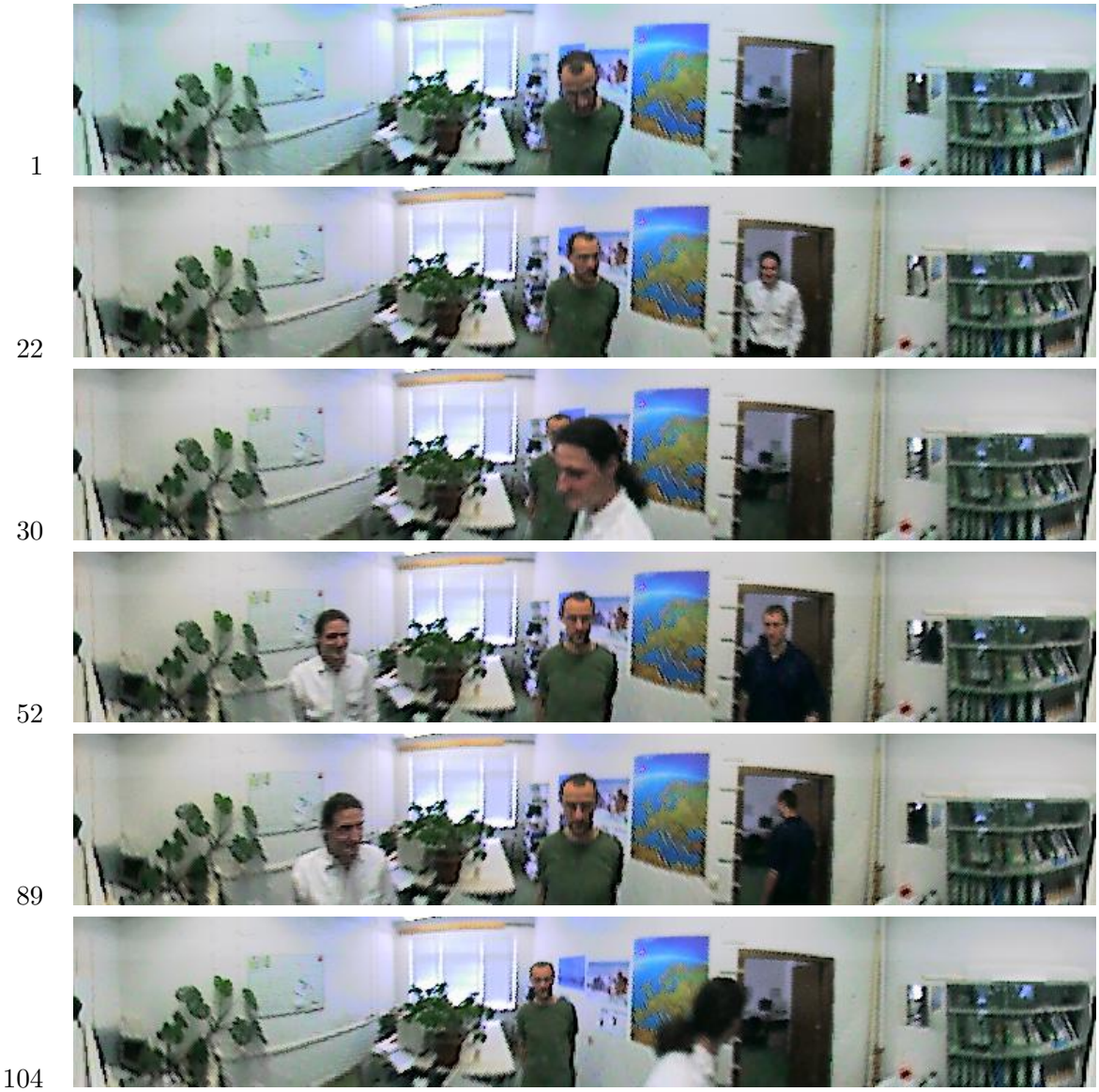


Abbildung 2.16: *Einzelne Bilder aus dem Testdatensatz. Auf der linken Seite ist die Nummer des Frames in der Bildfolge zu sehen. Die Frames entsprechen den Zeitpunkten, zu denen Personen die Szene betreten bzw. verlassen. (1) Zu Beginn befindet sich nur Person A in der Szene. (22) Person B betritt die Szene. (30) Person A wird kurzzeitig von Person B verdeckt. (52) Person C betritt die Szene. (89) Person C verlässt die Szene. (104) Person B verlässt die Szene.*

fer bestimmt. Hierfür wurden für alle Bilder des Datensatzes die Positionen aller Gesichter von Hand markiert. Eine Person gilt als detektiert, wenn die vom Tracker geschätzte Position innerhalb eines bestimmten Radius liegt. Wenn dies der Fall ist, wird außerdem die mittlere quadratische Abweichung zwischen tatsächlicher und geschätzter Position ermittelt. Ansonsten handelt es sich um eine Fehldetektion. Die Messung wurde für verschiedene Suchradien wiederholt. Die Ergebnisse sind in Tabelle 2.1 dargestellt. Die geringere Positioniergenauigkeit des Verfahrens nach Tao kann durch die Art und Weise der Positionsbestimmung erklärt werden. Da

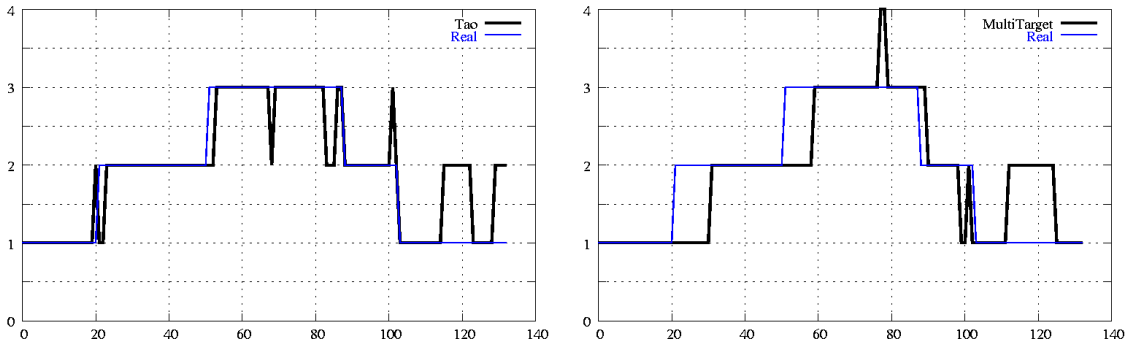


Abbildung 2.17: *Reale und geschätzte Personenzahl auf dem Testdatensatz (a) für das Verfahren nach Tao und (b) für das Verfahren mit mehreren Einzeltrackern. Wenn eine neue Person die Szene betritt, liefert der Tracker nach Tao schneller die korrekte Personenzahl. Beide Verfahren liefern am Ende der Sequenz für einen kurzen Moment eine zu große Schätzung. Dabei handelt es sich um die Arme von Person A, die korrekt als Hautfarbregion erkannt werden. Wichtig ist, dass möglichst immer alle Gesichter gefunden und verfolgt werden.*

in den Sample-Konfigurationen keine Zuordnung der einzelnen Samples zu Objekten existiert, müssen diese anhand ihrer räumlichen Verteilung gruppiert werden, was insbesondere bei dicht beieinander liegenden Verteilungen problematisch sein kann.

Abbildung 2.18: *Bestimmung der Positioniergenauigkeit der Tracking-Verfahren. Für den Datensatz wurden die Gesichtspositionen für jedes Bild von Hand bestimmt. Lag die vom Tracker geschätzte Position innerhalb eines definierten Radius, wurde dies als Treffer gewertet, ansonsten galt das Gesicht als nicht detektiert (Fehldetektion). Für die Treffer wurde zusätzlich die mittlere euklidische Distanz zwischen Label-Punkt und Schätzung bestimmt.*

	MT	**Tao**	**MT**	**Tao**	**MT**	**Tao**
Radius	10		20		30	
Fehldetektionen	119	191	50	64	39	45
Abweichung	5.68	6.83	8.55	11,54	9.37	12,59

Tabelle 2.1: *Positioniergenauigkeit der Tracking-Verfahren auf dem Testdatensatz. Für die Radien 10, 20 und 30 Pixel wurde die Anzahl von Fehldetektionen und der mittlere euklidische Abstand zwischen manuell gesetztem Label-Punkt und Schätzung bestimmt. Das Verfahren mit mehreren separaten Trackern (MT) trifft für alle Radiengrößen öfter die tatsächliche Position und erreicht außerdem eine höhere Positioniergenauigkeit.*

Zeitbedarf Abbildung 2.19 zeigt den Zeitbedarf der beiden Tracking-Systeme auf dem Datensatz. Zwar steigt für beide Systeme die benötigte Rechenzeit mit der Anzahl der getrackten Objekte, sie können jedoch beide als echtzeitfähig eingestuft werden und sind für die Anwendung im realen Einsatzfeld geeignet. Das Verfahren mit mehreren Einzel-Trackern ist etwas schneller.

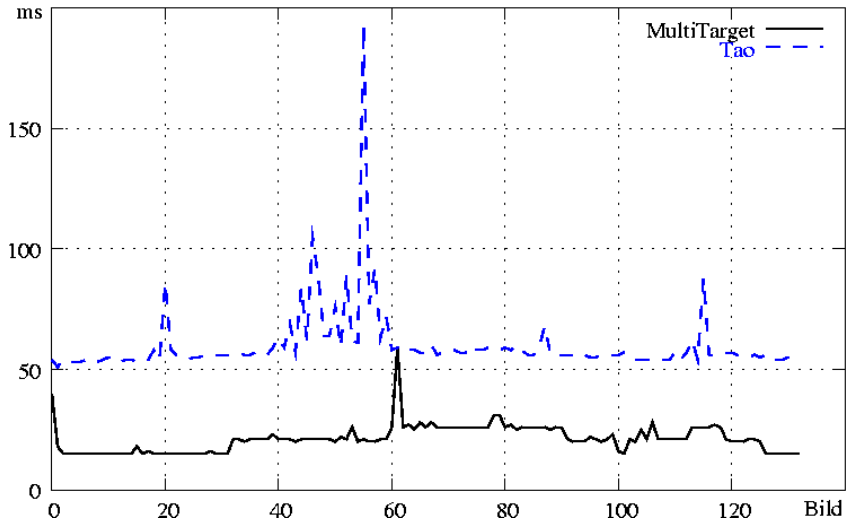


Abbildung 2.19: *Rechenzeitbedarf der beiden Multi-Target-Tracker im Millisekunden (AMD Athlon XP 3000+). Obwohl beide Verfahren als echtzeitfähig eingestuft werden können, ist das Verfahren mit mehreren Einzel-Trackern etwas schneller.*

2.5.4 Fazit

Bei beiden Ansätzen handelt es sich auf den ersten Blick um sehr unterschiedliche Verfahren, die aber tatsächlich an vielen Stellen sehr ähnliche Berechnungsmodelle verwenden. Das Verfahren mit mehreren Einzel-Trackern ist etwas schneller und genauer als das Verfahren von Tao und wird deshalb für den Einsatz im Gesamtsystem ausgewählt. Ein ganz entscheidender Vorteil bei der Verwendung von Einzel-Tracken besteht darin, dass prinzipiell beliebige Verfahren zum Tracken einzelner Objekte verwendet werden können, d.h. es besteht keine Beschränkung auf den CONDENSATION-Algorithmus. Die Verfahren müssen dazu lediglich ein Gütemaß zurückliefern. Mögliche Alternativen sind z.B. Kalman-Filter [Kalman, 1960] oder der CAMSHIFT-Algorithmus [Bradski, 1998a]. Ein ähnlicher Ansatz zum Tracking von Gesichtern wird von Shakhnarovich et al. verwendet [Shakhnarovich et al., 2002], wobei hier die Auffälligkeitskarten, auf denen die Tracker arbeiten, auf der Basis einer Gesichtsdetektion erstellt werden.

2.6 Ansteuerung der PTU

2.6.1 Motivation

In [Bruce et al., 2002] wird gezeigt, dass der direkte Blickkontakt auch bei der Interaktion zwischen Mensch und Roboter entscheidend dafür ist, dass die Interaktion als natürlich empfunden wird. Hier bestand die Aufgabenstellung für den Roboter darin, eine Umfrage durchzuführen und Passanten aktiv zur Interaktion mit dem Roboter aufzufordern. Der Roboter besaß einen Laser-Range-Finder zur Personendetektion und einen auf einer Pan-Tilt-Unit angebrachten Flachbildschirm für die Animation eines Gesichtes. Ziel der Untersuchungen war es, die Bedeutung des Gesichtes bzw. der PTU zu ermitteln. Dazu wurde das Verhalten der beobachteten Personen ausgewertet. Es wurde ermittelt, ob diese vom Roboter begrüßt wurden, ob sie beim Roboter stehen blieben, ob sie an der Umfrage teilnahmen oder ob sie die Umfrage vollständig beendeten. Die Ergebnisse belegen, dass sowohl das Gesicht als auch die PTU einen statistisch signifikanten Einfluss auf den Erfolg der Kontaktaufnahme haben. Beim Einsatz dieser Modalitäten stieg die Interaktionsbereitschaft der Passanten stark an.

Der Tourguide-Roboter Minerva besitzt ebenfalls ein Gesicht, auf das der Nutzer seine Aufmerksamkeit richten kann [Schulte et al., 1999]. Dieses ist mechanisch und kann in eingeschränktem Maße emotionale Zustände darstellen, die davon abhängen, ob dem Roboter der Weg zu einem Zielpunkt versperrt ist. Folgende Aussagen werden über den Einsatz des Gesichtes getroffen: „Das Gesicht als Blickpunkt bei der Interaktion hilft dabei, dass der Nutzer seine Aufmerksamkeit auf den Roboter richten kann und diesen als Kommunikationspartner akzeptiert. Der Roboter wird so als ein glaubwürdiger sozialer Agent wahrgenommen, der grundlegende soziale Konventionen befolgt, z.B. indem er seinen Nutzer anschaut.“

2.6.2 Realisierung

Nachdem eine auffällige Bildregion im Panoramabild gefunden wurde, wird der Kopf des Roboters in die entsprechende Richtung gedreht. Dies geschieht aus zwei Gründen. Zum einen sollte die gefundene Hypothese mittels eines Gesichtsdetektors verifiziert werden, bevor der vermeintliche Nutzer z.B. mittels einer Sprachausgabe kontaktiert wird. Da die geringe Auflösung im Panoramabild keine genauere Analyse der Gesichtsstrukturen zulässt, wird mit der Frontalkamera ein höher aufgelöstes Bild aufgenommen und im fovealen Vision-System ausgewertet. Zum anderen dient die Bewegung des Kopfes dazu, dem Nutzer die Aufmerksamkeit und damit die Kommunikationsbereitschaft des Roboters zu vermitteln. Aus eigener Erfahrung weiß jeder, wie essentiell solche elementaren Aufmerksamkeitsbekundungen in der zwischenmenschlichen Kommunikation sind.

Um den Kopf ausrichten zu können, müssen der Schwenk- und der Neigewinkel der Pan-Tilt-Unit berechnet werden, auf die der Kopf montiert ist. Wie von den Koordinaten der Nutzerhypothese im Panoramabild auf die räumliche Position des Nutzers relativ zum Gesicht des Roboters geschlossen werden kann, wird im Folgenden beschrieben.

Schwenkwinkel Wie in Abbildung 2.20 zu sehen ist, ergibt sich der relative horizontale Winkel direkt aus der x-Koordinate der Hypothese im Panoramabild:

$$\phi_{pan} = \frac{2\pi}{w}x - \pi \tag{2.17}$$

Dieser einfache Zusammenhang ist unabhängig vom verwendeten Spiegeltyp. Da der Kopf genau unter der omnidirektionalen Kamera angebracht ist, muss keine weitere Koordinatentransformation durchgeführt werden. Der ermittelte Winkel kann also unmittelbar zur Steuerung der PTU verwendet werden.

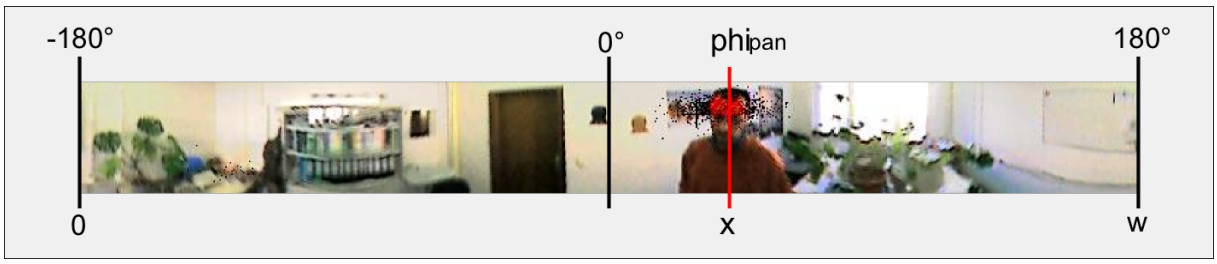


Abbildung 2.20: *Zusammenhang zwischen x-Koordinate und Schwenkwinkel der PTU. Unter dem Panoramabild ist die x-Koordinate der Hypothese und darüber der zugehörige Schwenkwinkel ϕ_{pan} aufgetragen. w ist die Breite des Panoramabildes in Pixeln.*

Neigewinkel Die Berechnung des Neigewinkels der PTU ist dagegen nicht nur abhängig vom Abstand d der Person zum Roboter, sondern auch von der Form des verwendeten Spiegels. In ersten Untersuchungen wurde mit einem sphärischen Spiegel gearbeitet [Wilhelm et al., 2003b]. Später kam ein hyperbolischer Spiegel zum Einsatz, wie er in Abbildung 2.21 schematisch dargestellt ist. Für die Transformation vom omnidirektionalen Bild in ein Panoramabild wird ein virtueller Einheitszylinder um den omnidirektionalen Spiegel gelegt, auf den das Panoramabild abgebildet wird. In Anhang A.2 wird die Transformation eines Bildes einer omnidirektionalen Kamera mit hyperbolischem Spiegel in ein Panoramabild erläutert.

Durch die Assoziation des Einheitszylinders mit Raumkoordinaten kann von der y-Koordinate im Panoramabild direkt auf Raumkoordinaten geschlossen werden. Dazu wird der Einheitszylinder so definiert, dass er mit einem Radius von $1m$ um den Spiegelbrennpunkt liegt. Es ist ausreichend, die Höhe eines Objektes im Abstand von $1m$ zum Kameraobjektiv für die Pixel am oberen Panoramabildrand z_{max} und für die Pixel am unteren Rand z_{min} zu bestimmen. Dazu kann das in Anhang A.2 beschriebene Testmuster verwendet werden. Wie in Abbildung 2.22 gezeigt, kann für alle Pixel im Panoramabild die Höhe z in Metern auf dem Einheitszylinder wie folgt linear interpoliert werden.

$$z = \left(\frac{y}{h}\left(z_{min} - z_{max}\right) + z_{max}\right) \tag{2.18}$$

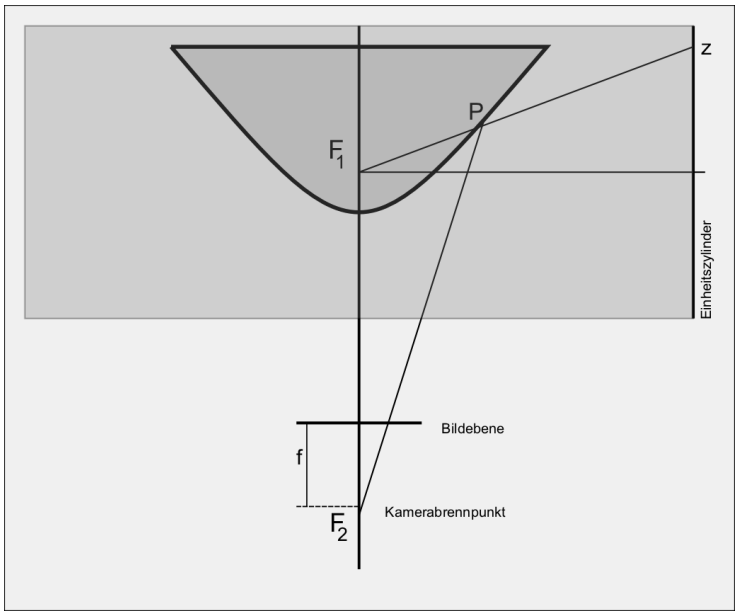


Abbildung 2.21: *Schematische Darstellung eines hyperbolischen Spiegels. Das Bild der omnidirektionalen Kamera wird auf einen virtuellen Einheitszylinder projiziert, der um den Brennpunkt des omnidirektionalen Objektivs gelegt wird.*

Abbildung 2.22: *Zusammenhang zwischen y-Koordinate eines Pixels im Panoramabild und der Höhe z relativ zur Höhe des omnidirektionalen Spiegels. h ist die vertikale Auflösung des Panoramabildes.*

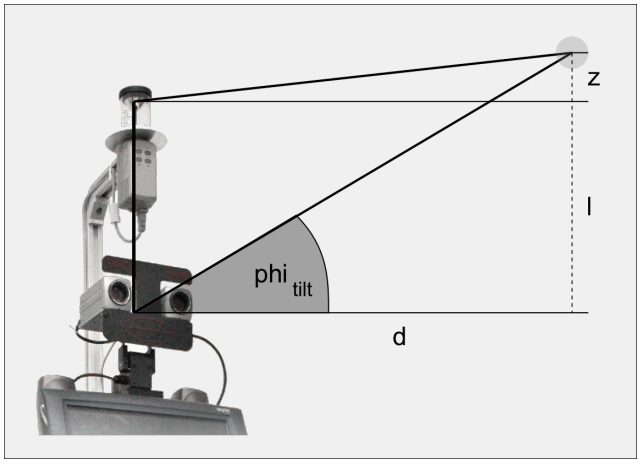


Abbildung 2.23: *Der Neigewinkel ϕ_{tilt} für die PTU kann aus dem Abstand d zur Hypothese und aus deren Höhe berechnet werden. Letztere kann direkt aus der y-Koordinate im Panoramabild ermittelt werden.*

Abbildung 2.23 zeigt, wie aus der Entfernung d zum Objekt und der Höhe der Hypothese z auf dem Einheitszylinder der Winkel ϕ_{tilt} berechnet werden kann. Dazu wird zunächst die Höhe der Hypothese relativ zur PTU bestimmt. Der Winkel ϕ_{tilt} kann schließlich mit Hilfe der durch die Sonarsensoren gemessenen Entfernung zur Hypothese berechnet werden:

$$\phi_{tilt} = \frac{l + z}{d} \tag{2.19}$$

Durch das Aufmerksamkeitssystem ist der Roboter in der Lage, Personen in seinem Umfeld zu detektieren, eine Nutzerhypothese zu bilden und diese mit der Frontalkamera anzuschauen. Auf diese Weise können hochaufgelöste Bilder des Nutzers aufgenommen und durch das foveale Vision-System ausgewerdet werden. Die nächsten beiden Kapitel beschreiben diese Komponente der Systemarchitektur.

Kapitel 3

Gesichtsnormalisierung

3.1 Aufgabenstellung

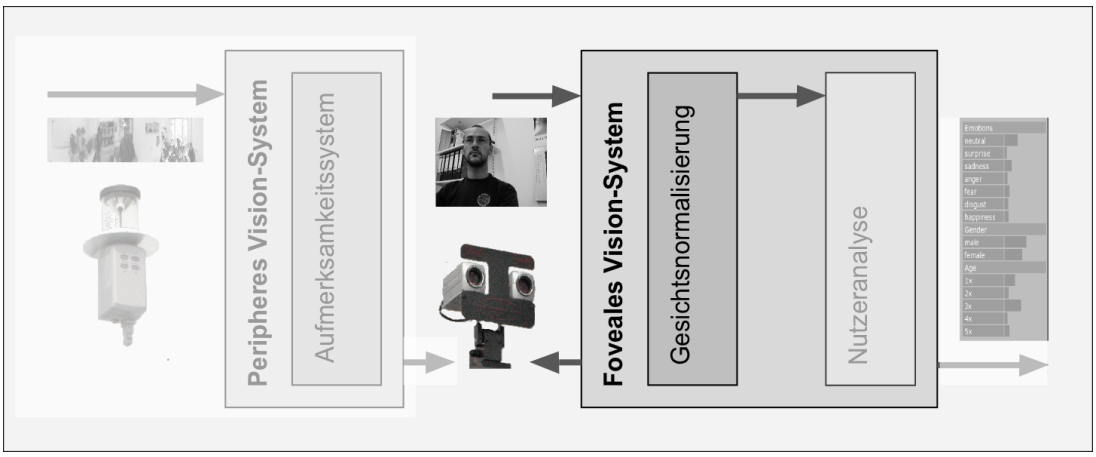


Abbildung 3.1: *Einordnung der Gesichtsnormalisierung in die Systemarchitektur. Sie bildet die erste Komponente des fovealen Vision-Systems. Hier wird das von der Frontalkamera aufgenommene Bild so aufbereitet, dass es von den nachfolgenden Modulen verarbeitet werden kann.*

Die Gesichtsnormalisierung ist die erste Komponente des fovealen Vision-Systems. Sie schafft die Voraussetzungen für die nachfolgende Nutzeranalyse. Dazu gehört zum einen die Detektion des Gesichts im Bild der Frontalkamera und zum anderen dessen genaue Ausrichtung anhand markanter Punkte im Gesicht. Diese beiden Teilschritte werden in den Abschnitten 3.2 und 3.3 näher betrachtet.

3.2 Gesichtsdetektion

3.2.1 Aufgabenstellung

Im letzten Kapitel wurde gezeigt, wie die Frontalkamera des Roboters auf die Hypothese ausgerichtet und wie ein hochaufgelöstes Bild vom entsprechenden Objekt aufgenommen werden kann. Die erste Aufgabe des fovealen Vision-Systems besteht darin, die vom peripheren Vision-System ausgewählte Hypothese zu verifizieren. Dazu wird in dem mit der Frontalkamera aufgenommenem Bild eine Gesichtsdetektion durchgeführt. Diese Aufgabe wird laut [Yang et al., 2002] wie folgt definiert:

> Das Ziel der Gesichtsdetektion besteht darin, für ein beliebiges Bild zu entscheiden, ob dieses Gesichter enthält und falls ja, die Koordinaten und Größen dieser zu ermitteln.

Die Gesichtsdetektion kann auch als 2-Klassen-Problem aufgefasst werden, bei dem für jedes Pixel im Bild entschieden wird, ob sich in dessen Umgebung ein Gesicht befindet oder nicht. Im Vergleich zu anderen Klassifikationsaufgaben ist die Gesichtsdetektion deshalb problematisch, weil die Variabilität innerhalb der Gesichtsklasse sehr groß ist. Diese ist bedingt durch verschiedene Blickwinkel, Gesichtsausdrücke, Verdeckungen, Beleuchtungsbedingungen und dem Vorhandensein oder Nicht-Vorhandensein von Strukturelementen wie Bärten oder Brillen. Die Gesichtsdetektion ist ein sehr intensiv beforschtes Gebiet in der Bildverarbeitung. Im Rahmen dieser Arbeit wurden verschiedene aus der Literatur bekannte Verfahren implementiert und vergleichend untersucht, um letztlich ein geeignetes Verfahren für das Gesamtsystem auswählen zu können. Die Theorie der untersuchten Verfahren und die Ergebnisse der Vergleiche werden in diesem Kapitel vorgestellt. Zunächst soll aber versucht werden, eine Taxonomie verschiedener Verfahren zur Gesichtsdetektion zu erstellen.

3.2.2 Literatur

Abbildung 3.2 zeigt eine Taxonomie von Gesichtsdetektionsverfahren. Die einzelnen Gruppen können wie folgt charakterisiert werden:

Merkmalsbasierte Verfahren Zu den merkmalsbasierten Verfahren zählen solche, die kein Gesichtsmodell einsetzen, sondern für Gesichter typische Merkmale, wie z.B. Hautfarbe [Yang and Waibel, 1996]. Solche Verfahren können dann eingesetzt werden, wenn bestimmte Randbedingungen erfüllt sind, z.B. wenn sich nur eine Person in einem bestimmten Abstand vor der Kamera aufhält und der Hintergrund keine Objekte mit ähnlichen Merkmalen enthält. Werden diese Randbedingungen nicht eingehalten, erweisen sich diese Verfahren als zu unspezifisch.

Modellbasierte Verfahren Modellbasierte Verfahren für die Gesichtsdetektion besitzen eine „Erwartung“ darüber, was sie sehen sollten, d.h. eine wie auch immer geartete Beschrei-

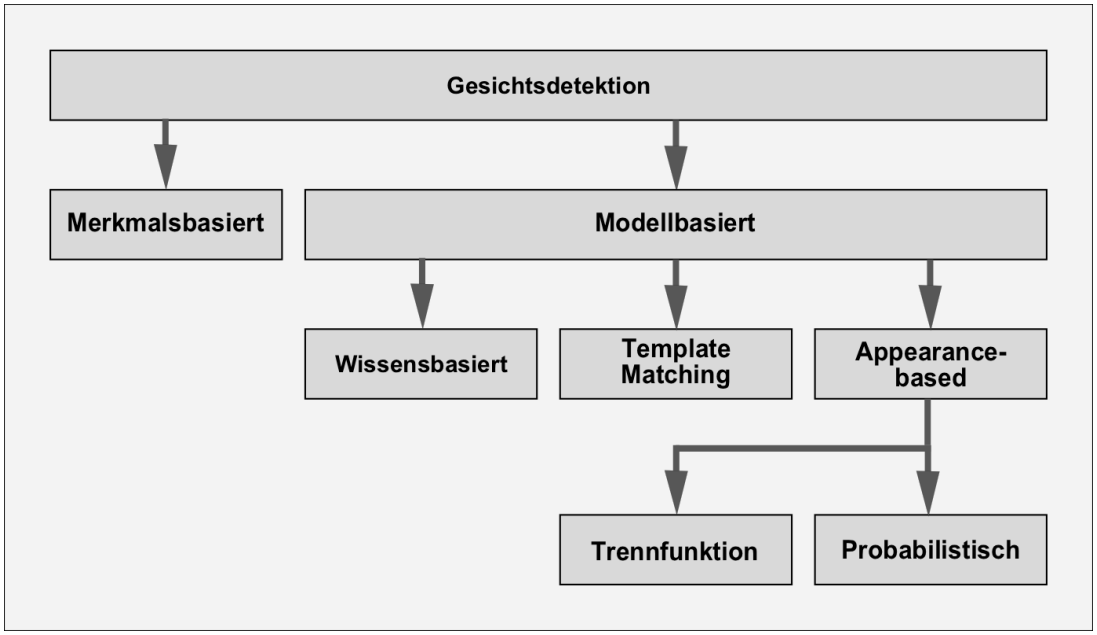


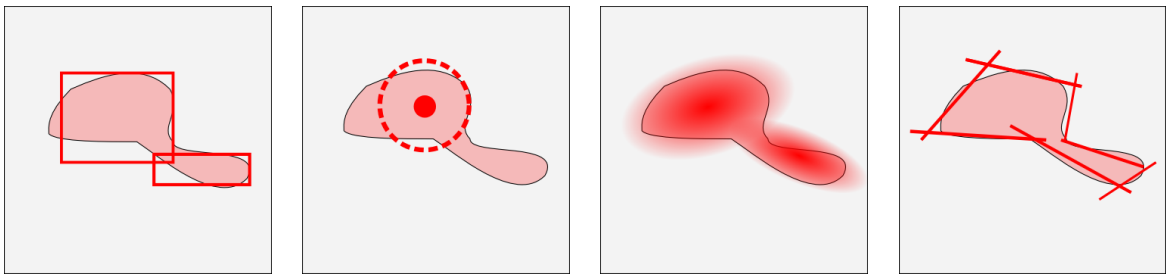
Abbildung 3.2: *Taxonomie von Gesichtsdetektionsverfahren.*

bung eines Gesichts. Die modellbasierten Verfahren können in folgende Gruppen eingeteilt werden:

Wissensbasierte Verfahren Die hier eingesetzten Modelle werden sozusagen von Hand erstellt, d.h. es geht das Wissen eines Designers ein, der entscheidet, welche Merkmale in einem Gesicht zu finden sein sollten und in welcher räumlichen Beziehung diese zueinander stehen. Solche Modelle können starr sein oder eine gewisse Variabilität in der Ausprägung und Anordnung der Merkmale zulassen, wobei auch diese Variabilität vom Designer definiert werden muss. Ein Vorteil dieser Vorgehensweise besteht darin, dass keine repräsentative Trainingsmenge vorliegen muss. Ein Vertreter für ein wissensbasiertes Verfahren ist [Kotropoulos and Pitas, 1997].

Template-Matching-Verfahren Diese Modelle verwenden ein repräsentatives Beispiel oder einige wenige Beispiele für die Beschreibung von Gesichtern. Für die Detektion wird eine Korrelation von entsprechend vorverarbeiteten Ausschnitten aus dem Eingabebild mit dem Template berechnet. Die Modelle werden zwar aus Trainingsdaten erstellt, beschreiben aber nicht die Klasse der Gesichter oder Nicht-Gesichter an sich, sondern nur typische Vertreter. Auch diese Modelle können starr oder flexibel sein. Ein Verfahren mit starren Templates ist das Edge-Orientation-Matching [Fröba and C., 2001], und ein Beispiel für ein Verfahren mit einem flexiblem Template ist das Elastic-Graph-Matching [Wiskott et al., 1997a].

Appearance-basierte Verfahren Hier werden die Modelle aus einer Menge von Trainingsbildern erstellt, die die Variabilität der Gesichts- bzw. Nicht-Gesichtsklasse abdecken sollte. Entsprechend der eingesetzten Trainingsmethoden kann hier nochmals zwischen probabilistischen und neuronalen Verfahren unterschieden werden. Vertreter dieser Gruppe sind Eigenfaces [Turk and Pentland, 1991], die Modellie-



(a) (b) (c) (d)

Abbildung 3.3: *Veranschaulichung der Verfahren zur Gesichtsdetektion. Die Rechtecke repräsentieren jeweils den so genannten Image-Space, den Raum aller möglichen Bilder. Der eingezeichnete Unterraum stellt den Face-Space dar. Dieser Raum wird je nach verwendeter Vorverarbeitung und Merkmalsextraktion unterschiedlich definiert und kann dementsprechend unterschiedliche Dimensionen haben. (a) Bei wissensbasierten Verfahren wird das Modell vom Entwickler entworfen. Von dessen Geschick beim Entwurf hängt es ab, wie gut das Modell den Face-Space abdeckt. (b) Template-Matching Verfahren repräsentieren die Menge der Gesichter durch ein oder mehrere repräsentative Beispiele. Ein zu testendes Muster wird anhand des Abstandes von diesem Template klassifiziert. (c) Bei probabilistischen Verfahren wird jedem Punkt im Image-Space eine Wahrscheinlichkeit für die Zugehörigkeit zu einer der beiden Klassen zugewiesen. (d) Bei den Verfahren mit Trennfunktion werden die beiden Klassen z.B. durch Hyperebenen getrennt.*

rung der Verteilungsfunktion der Gesichtsklasse mit multidimensionalen Gaußfunktionen [Sung and Poggio, 1998]. Beispiele neuronaler Verfahren sind [Yang et al., 2000] [Féraud et al., 2001] und [Rowley et al., 1998] und die kaskadierten Klassifikatoren nach [Viola and Jones, 2004].

Abbildung 3.3 veranschaulicht die prinzipielle Funktionsweise der einzelnen Verfahren.

3.2.3 Untersuchte Verfahren

Im Folgenden werden drei Verfahren für die Gesichtsdetektion vorgestellt, deren Implementierung im Rahmen dieser Arbeit erfolgte. Dabei handelt es sich mit dem Edge-Orientation-Matching [Fröba and C., 2001] um ein Template-Matching-Verfahren und mit dem Detektor nach Rowley [Rowley et al., 1998] und dem AdaBoost-Detektor [Viola and Jones, 2004] um zwei appearance-basierte Verfahren.

3.2.3.1 Edge-Orientation-Matching

Dieses Verfahren wurde am Fraunhofer Institut für integrierte Schaltungen in Erlangen entwickelt [Fröba and C., 2001]. Es gehört zur Gruppe der Template-Matching Verfahren, wobei das verwendete Template auf Kantenstärken und -orientierungen beruht. Ein Vorteil des Verfahrens besteht darin, dass für die Erstellung des Modells keine negativen Trainingsbeispiele benötigt werden.

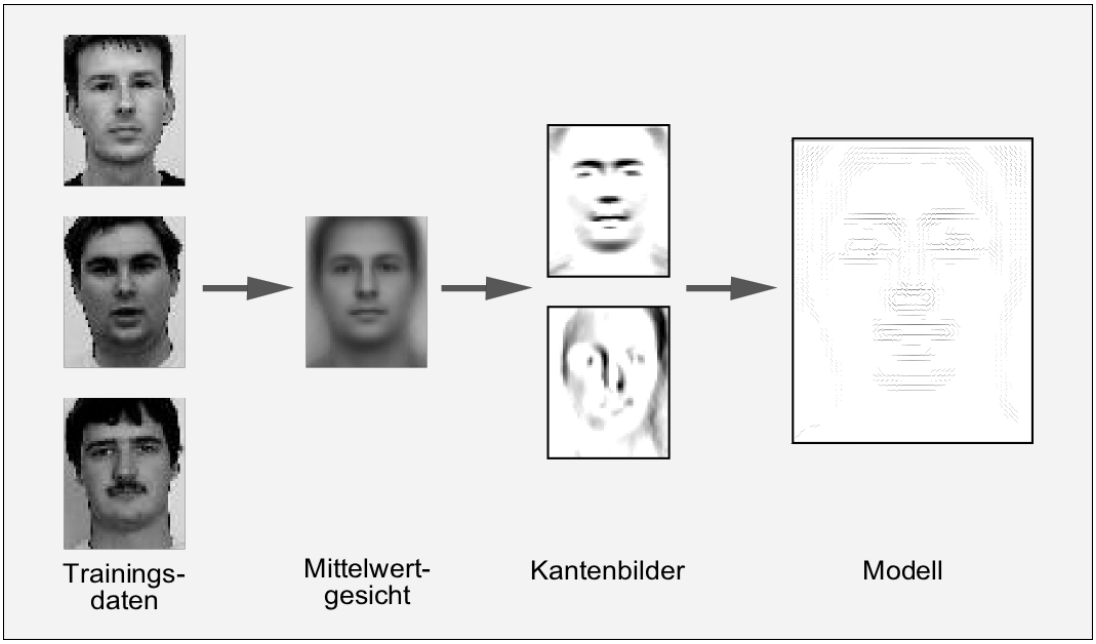


Abbildung 3.4: *Modellerstellung beim Edge Orientation Matching. Berechnung des Mittelwertgesichtes, Kantendetektion, Berechnung der Kantenstärke und -orientierung. Im Modell wird die Richtung einer Kante durch die Richtung des Vektors dargestellt und die Stärke durch dessen Länge.*

Modellerstellung Abbildung 3.4 zeigt den Ablauf der Erstellung des Templates.

Mittelwertgesicht: Aus einer Menge von Gesichtsbildern wird ein Mittelwertgesicht I_{mean} berechnet. Die Bilder müssen alle die gleiche Größe haben und das Gesicht muss sich immer an der gleichen Stelle befinden.

Kantendetektion: Auf dem Mittelwertgesicht wird ein Histogrammausgleich durchgeführt (siehe Anhang A.3.1) und ein Kantendetektor angewandt, der für jeden Pixel die Stärke und Orientierung der Grauwertstruktur in einer lokalen Umgebung um den Pixel ermittelt. Bei den in [Fröba and C., 2001] und auch in dieser Arbeit verwendeten Kantendetektoren handelt es sich um einen horizontalen und einem vertikalen Sobelfilter:

$$\mathbf{K}_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \qquad \mathbf{K}_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{3.1}$$

Die Faltung des Mittelwertbildes mit den Sobel-Filtern ergibt:

$$\mathbf{G}_x(x,y) = \mathbf{K}_x * \mathbf{I}_{mean}(x,y), \tag{3.2}$$

$$\mathbf{G}_y(x,y) = \mathbf{K}_y * \mathbf{I}_{mean}(x,y). \tag{3.3}$$

Kantenstärke und -orientierung: Die Stärke $\mathbf{S}$ und die Orientierung ϕ der Kanten kann nach Gleichung 3.4 und 3.5 berechnet werden. Da die Kanteninformation in homogenen Be-

reichen des Bildes hauptsächlich aus Bildrauschen resultiert, wird ein Schwellwertoperator auf die Kantenstärke angewendet, siehe Gleichung 3.6. Da nur die Orientierung der Kanten verwendet wird und nicht auch deren Richtung (keine Unterscheidung zwischen Übergängen von hell nach dunkel bzw. dunkel nach hell), wird der Winkel ϕ entsprechend Gleichung 3.7 so angepasst, dass er Werte zwischen 0 und π annimmt. Die Kantenstärke und -orientierung an jedem Punkt des Bildes kann schließlich als komplexe Zahl mit Betrag und Winkel dargestellt werden, siehe Gleichung 3.8.

$$\mathbf{S}(x,y) = \sqrt{\mathbf{G}_x(x,y)^2 + \mathbf{G}_y(x,y)^2} \tag{3.4}$$

$$\phi(x,y) = \arctan\left(\frac{\mathbf{G}_y(x,y)}{\mathbf{G}_x(x,y)}\right) + \frac{\pi}{2} \tag{3.5}$$

$$\mathbf{S}_T = \begin{cases} S(x,y) & ,falls \quad \mathbf{S}(x,y) < T_s \\ 0 & ,sonst \end{cases} \tag{3.6}$$

$$\phi' = \begin{cases} \phi(x,y) & ,falls \quad 0 \leq \phi(x,y) < \pi \\ \phi(x,y) - \pi & ,falls \quad \pi \leq \phi(x,y) < 2\pi \end{cases} \tag{3.7}$$

$$\mathbf{V} = \mathbf{S}_T \exp(j\phi'). \tag{3.8}$$

Modellanwendung Der Ablauf einer Gesichtsdetektion mit dem Edge-Orientation-Matching ist in Abbildung 3.5 dargestellt.

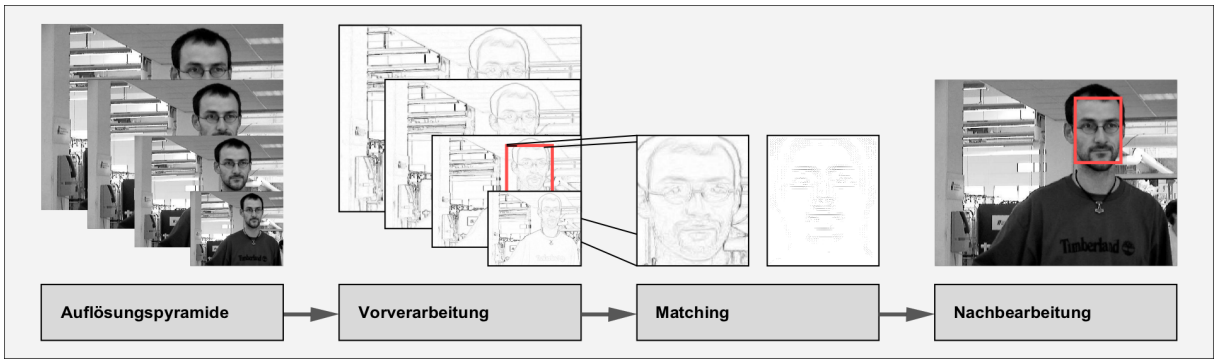


Abbildung 3.5: *Der Ablauf einer Gesichtsdetektion mit dem Edge-Orientation-Modell. Um Gesichter in verschiedenen Größen detektieren zu können, wird auf einer Auflösungspyramide gearbeitet. In der Vorverarbeitung wird ein Kantenorientierungsfeld für die Bilder aller Auflösungsstufen berechnet. Beim Matching wird die Ähnlichkeit aller Bildausschnitte mit dem Template berechnet und schließlich wird in der Nachbearbeitung eine Auswahl möglicher Gesichtspositionen getroffen.*

Auflösungspyramide: Um Gesichter verschiedener Größe mit einem Template fester Größe detektieren zu können, wird auf mehreren Auflösungsstufen des Eingabebildes gearbeitet.

Vorverarbeitung: Analog zur Erstellung des Templates wird für das Eingabebild entsprechend der Gleichungen 3.2 bis 3.8 ein Kantenorientierungsfeld berechnet.

Matching: Der Abstand zwischen dem Kantenorientierungsfeld $\mathbf{V}_I$ des Bildes mit dem Kantenorientierungsfeld $\mathbf{V}_M$ des Templates wird an jeder Bildposition (x, y) wie folgt berechnet:

$$\begin{aligned} \mathbf{C}(x,y) &= \sum_{n=-\frac{N}{2}}^{\frac{N}{2}} \sum_{m=-\frac{M}{2}}^{\frac{M}{2}} \left(dist(\mathbf{V}_M(m,n), \mathbf{V}_I(x+m, y+n))\right), \qquad (3.9) \\ dist(\bullet) &= \begin{cases} sin(|arg(V_I) - arg(V_M)|) & , wenn \quad \|V_B\|, \|V_I\| > 0 \\ 1 & , sonst \end{cases}, \end{aligned}$$

wobei $M \times N$ die Größe des Templates ist.

Nachbearbeitung Die lokalen Minima von $\mathbf{C}(x, y)$ stellen mögliche Gesichtspositionen dar. Die n besten Positionen, die außerdem einen bestimmten Schwellwert unterschreiten, werden als mögliche Gesichtspositionen ausgegeben.

3.2.3.2 Gesichtsdetektion nach Rowley

Dieses Verfahren wurde an der Carnegie Mellon University entwickelt [Rowley et al., 1998]. Es handelt sich um ein modellbasiertes neuronales Verfahren, siehe Abbildung 3.2. Das Gesamtsystem hat drei Stufen. Die erste besteht aus einer Vorverarbeitung der Bilddaten, die zweite aus einem neuronalen Netz, das als Klassifikator fungiert und die dritte Stufe aus einer Nachbearbeitung, welche Mehrfachdetektionen eliminiert. Der neuronale Klassifikator arbeitet auf Bildausschnitten der Größe 20×20 Pixel und besteht aus einem zweischichtigen Multilayer-Perceptron. Die Neuronen der Hidden-Schicht haben unterschiedliche rezeptive Felder, deren Formen der Detektionsaufgabe angepasst sind. Im Fall der Gesichtsdetektion gibt es drei verschiedene Arten von rezeptiven Feldern: vier Hidden-Neuronen mit 10×10 Pixeln, 16 mit 5×5 Pixeln und sechs mit überlappenden rezeptiven Feldern von 20×5 Pixeln, siehe Abbildung 3.6.

Modellerstellung

Vorverarbeitung: In der Vorverarbeitung des betrachteten Bildausschnitts wird ein Intensitätsausgleich (siehe Anhang A.3.2) und ein Histogrammausgleich (siehe Anhang A.3.1) durchgeführt. Verfälschungen durch den Hintergrund werden vermindert, indem die Vorverarbeitungsschritte nur auf eine ovale Region in der Mitte des Bildausschnittes angewandt werden. Diese Region würde, falls das Bild ein Gesicht enthält, dem Gesicht entsprechen, während alles außerhalb Hintergrund ist.

Training Um das Netzwerk zu trainieren, müssen sowohl positive als auch negative Trainingsbeispiele vorhanden sein. Während die Bereitstellung einer Datenbank mit Gesichtsbildern kein Problem darstellt, ist die Auswahl repräsentativer Nicht-Gesichter schwierig, da die Trainingsbeispiele alle möglichen Situationen enthalten sollten. Um dies zu realisieren, arbeitet der Rowley-Detektor mit einem Bootstrap-Algorithmus, der die negativen Trainings-

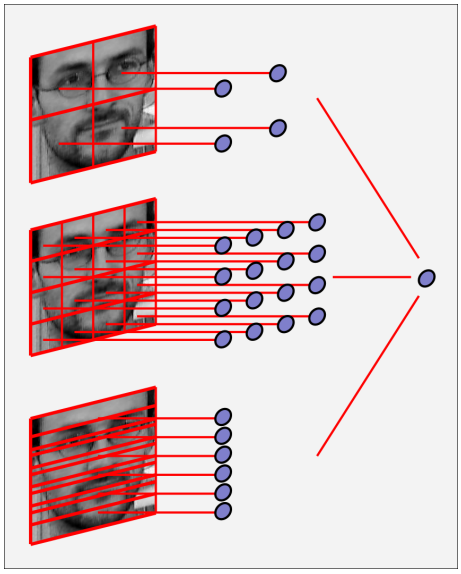


Abbildung 3.6: *Der Aufbau des neuronalen Klassifikators. Die rezeptiven Felder sind durch rote Rechtecke dargestellt. Die horizontalen Streifen im untersten Bild überlappen sich um jeweils einen Pixel und detektieren horizontale Strukturen wie die Augenbrauen, die Augen oder die Mundpartie, während die Neuronen mit quadratischen rezeptiven Feldern einzelne Merkmale finden sollen, wie z.B. einzelne Augen, die Nase oder einen Mundwinkel.*

beispiele während des Lernprozesses generiert. Das Netz wird zunächst mit zufälligen Gewichten initialisiert. Beim Trainingsalgorithmus handelt es sich um Error-Backpropagation in Verbindung mit einer Cross-Validation-Technik, mit der Besonderheit, dass wie bereits erwähnt, die negativen Trainingsbeispiele während des Trainings erzeugt und ausgewählt werden. Dieser Bootstrap-Algorithmus funktioniert wie folgt:

1 Aus einem Satz von Bildern, die keine Gesichter enthalten, wird zufällig eine Anzahl von Nicht-Gesichtbeispielen generiert und auf diese die Vorverarbeitungsschritte angewandt.

2 Das neuronale Netz wird mittels Error-Backpropagation trainiert.

3 Der Detektor wird auf ein Bild angewandt, das keine Gesichter enthält, wobei alle Ausschnitte gesammelt werden, die fälschlicherweise als Gesichter identifiziert wurden.

4 Aus diesen Ausschnitten erfolgt eine zufällige Auswahl, auf die die Vorverarbeitungsschritte angewandt wird. Diese Ausschnitte werden dem Trainingsdatensatz als negative Beispiele hinzugefügt. Weiter bei 2.

In der in dieser Arbeit verwendeten Implementierung wird das Training abgebrochen, wenn alle zur Verfügung stehenden Nicht-Gesichtsbilder verwendet wurden. D.h. in jedem Bootstrap-Zyklus wird aus einem der Nicht-Gesichtsbilder eine festgelegte Anzahl von negativen Trainingsbeispielen generiert und das verwendete Bild anschließend als benutzt gekennzeichnet. Wurden alle Bilder benutzt, wird das Training beendet.

Modellanwendung Der Ablauf einer Gesichtsdetektion mit dem Verfahren nach Rowley ist in Abbildung 3.5 dargestellt.

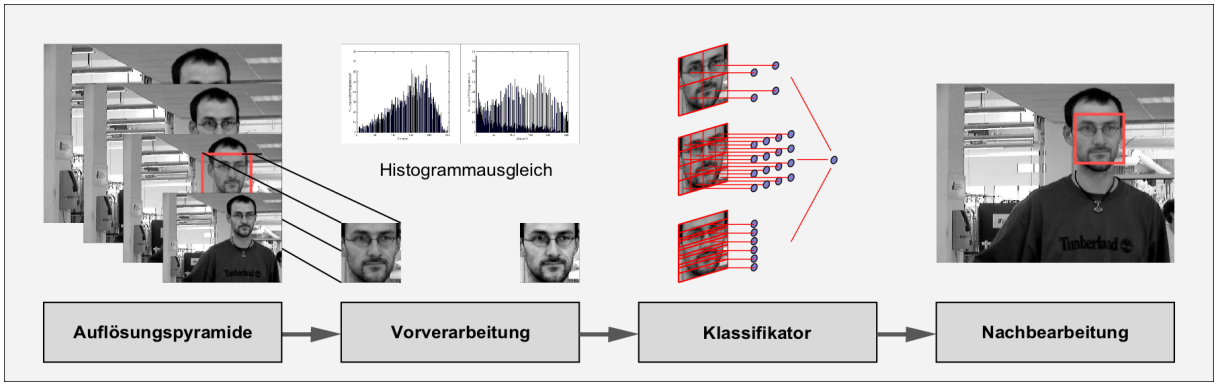


Abbildung 3.7: *Der Ablauf einer Gesichtsdetektion mit dem Rowley-Detektor. Der Klassifikator besteht aus einem neuronalen Netz, dessen Eingangsdaten durch einen Intensitäts- und einen Histogrammausgleich vorverarbeitet werden. Die Ergebnisse des neuronalen Klassifikators werden in einer Nachbearbeitung zusammengefasst.*

Auflösungspyramide: Um Gesichter verschiedener Größe mit einem Detektor fester Größe detektieren zu können, wird auf mehreren Auflösungsstufen des Eingabebildes gearbeitet.

Vorverarbeitung: Jeder Bildausschnitt durchläuft die selbe Vorverarbeitung wie die Bildausschnitte, die für das Training verwendet wurden.

Anwendung des Klassifikators: Der Klassifikator wird an jeder Position des Bildes und auf jeder Stufe der Auflösungspyramide angewandt. Er generiert einen Ergebniswert zwischen -1 und 1.

Nachbearbeitung: Da der neuronale Klassifikator in gewissen Grenzen invariant gegen Größen- und Positionsänderungen der gesuchten Merkmale ist, kommt es in der Regel zu Mehrfachdetektionen. Die Aufgabe der Nachbearbeitung ist es, solche Mehrfachdetektionen zu erkennen und auf eine Gesichtsposition zu reduzieren. Gleichzeitig soll die Anzahl der Fehldetektionen nach Möglichkeit reduziert werden. Dies geschieht mit einem heuristischen Ansatz. Wegen der oben beschriebenen Gründe wird ein Gesicht normalerweise mehrere Male detektiert, während dies bei einer Fehldetektion in der Regel nicht der Fall ist. Es wird die Anzahl aller Detektionen gezählt, die sich innerhalb einer gewissen Nachbarschaft im Bild befinden. Nur solche Positionen, die häufiger als ein Schwellwert als Gesichter klassifiziert wurden, sind mit großer Wahrscheinlichkeit wirklich welche. Wenn eine Position im Bild als Gesicht klassifiziert wurde, so sind alle Detektionen, die diesen Bereich überlappen, mit hoher Wahrscheinlichkeit Fehler und können ebenfalls gelöscht werden. Dies wird realisiert, indem diejenigen Regionen als Gesichter klassifiziert werden, die am häufigsten detektiert wurden, während alle überlappenden Detektionen, die weniger häufig sind, gelöscht werden.

3.2.3.3 AdaBoost und Filterkaskaden

Dieses Verfahren wurde erstmals 2001 in [Viola and Jones, 2001] vorgestellt. Es handelt sich um ein appearance-basiertes Verfahren, siehe Abbildung 3.2. Im Folgenden werden einige Besonderheiten dieses Verfahrens näher erläutert.

Einfache Klassifikatoren: Die verwendeten Filter bestehen aus einfachen Rechtecken und lassen sich mit Hilfe des Integralbildes (siehe Anhang A.4) sehr effizient berechnen, siehe Abbildung 3.8. Durch Anwendung eines Schwellwertes auf den berechneten Filterwert erhält man einen einfachen binären Klassifikator. Außerdem wird dem Schwellwert eine Polarität zugewiesen. Eine Polarität von -1 bedeutet, dass ein Bildausschnitt, dessen Filterwert kleiner als die Schwelle ist, als Hintergrund klassifiziert wird. Umgekehrt bedeutet Polarität 1, dass ein Bildausschnitt, dessen Filterwert kleiner als die Schwelle ist, als Gesicht klassifiziert wird.

Abbildung 3.8: *Für den Aufbau eines Klassifikators verwendete einfache Filter. Der Filterwert wird berechnet, indem die Grauwerte aller Pixel des Bildes, die unter einem weißen Rechteck liegen, von jenen Pixeln abgezogen werden, die unter einem grauen Rechteck liegen.*

Komplexe Klassifikatoren: Ein komplexer Klassifikator wird aus mehreren einfachen Klassifikatoren gebildet. Dazu werden diese in allen möglichen Größen und an allen möglichen Positionen in einem Klassifikatorfenster der Größe 24×24 Pixel positioniert, wobei eine sehr große Anzahl möglicher Kombinationen existiert. Bei AdaBoost handelt es sich um ein Verfahren, mit dem eine optimale Kombination einfacher Klassifikatoren hinsichtlich der Klassifikationsaufgabe gefunden wird.

Klassifikatorkaskaden: Bei Klassifikatorkaskaden handelt es sich um mehrere hintereinander geschaltete komplexe Klassifikatoren mit binärer Ausgabe, die sich besonders für Zweiklassenprobleme eignen, bei denen eine Klasse häufiger auftritt. Die Klassifikation erfolgt dabei durch sukzessives Ausdünnen der häufiger auftretenden Klasse (Nicht-Gesichter). Nach der letzten Stufe bleiben die Daten der seltener auftretenden Klasse übrig (Gesichter), siehe Abbildung 3.9. Das Prinzip dieser Vorgehensweise besteht darin, dass viele der Daten der häufiger auftretenden Klasse sehr leicht erkennbar sind und frühzeitig mit einfachen Klassifikatoren verworfen werden können. Nur schwer zu klassifizierende Daten gelangen in die hinteren Stufen der Kaskade und werden dort mit komplexeren Klassifikatoren untersucht. Während die Rechenkomplexität in der Kaskade also von vorne nach hinten zunimmt, nimmt die Anzahl der zu untersuchenden Bildpunkte ab.

Modellaufbau Bei AdaBoost werden die Trainingsdaten gewichtet. Jedes Trainingsbeispiel besteht aus dem Trainingsbild, einer Klassifikation in Gesicht oder Nicht-Gesicht und einem

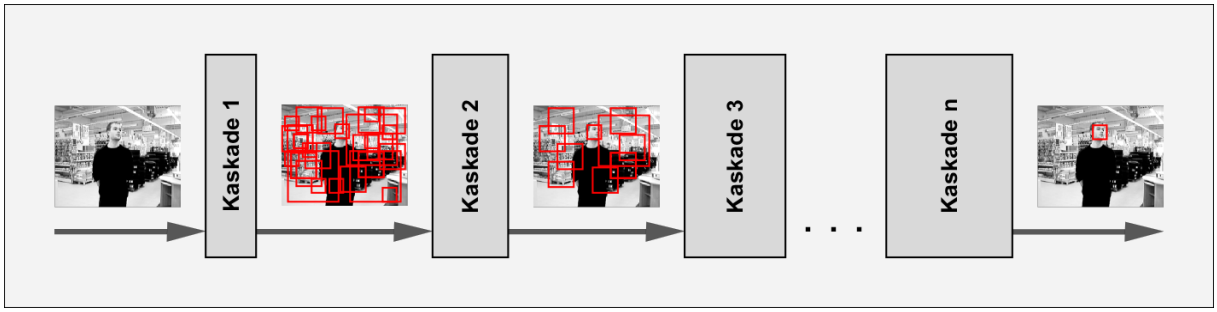


Abbildung 3.9: *Prinzip der Klassifikatorkaskaden. Die vorderen Stufen enthalten einfache Klassifikatoren und verwerfen bereits viele Nicht-Gesichter. Am Ende der Kaskade befinden sich sehr komplexe Klassifikatoren, die aber nur noch auf einen kleinen Teil der Daten angewandt werden müssen.*

Wichtungsfaktor, der anfangs für alle Trainingsbeispiele gleich ist. Dieser Wichtungsfaktor wird nach jedem Durchlauf neu berechnet, so dass die falsch klassifizierten Beispiele stärker betont werden, und es wird nach einem weiteren einfachen Klassifikator gesucht, der das neue Klassifikationsproblem am Besten löst. Dies wird so lange wiederholt, bis eine vorgegebene Anzahl von Klassifikatoren oder eine Fehlerschwelle erreicht ist. Der entstehende komplexe Klassifikator besteht aus einer gewichteten Kombination aller vorher ermittelten einfachen Klassifikatoren und einer Schwellwertoperation. Er realisiert also eine gewichtete Mehrheitsentscheidung. Die Aufgabe des eigentlichen Trainingsalgorithmus besteht darin, aus der Menge von einfachen Klassifikatoren in jedem Trainingslauf denjenigen auszuwählen, welcher die beste Trennung zwischen positiven und negativen Trainingsbeispielen realisiert, sowie die Wichtung der einzelnen schwachen Klassifikatoren und den Schwellwert für die Filterantwort festzulegen. Dieser Schwellwert wird entweder so gewählt, dass die Gesamtzahl der Fehlklassifikationen minimal wird, oder dass die Falsch-Negativ-Rate Null beträgt, also keines der positiven Trainingsbeispiele aussortiert wird.

Modellanwendung Da die Berechnungskomplexität der primitiven Filter auf dem Integralbild unabhängig von der Größe des Filters ist, muss keine Auflösungspyramide verwendet werden, sondern die Klassifikatoren werden mit einem Skalierungsfaktor sukzessive vergrößert. Bei der Anwendung durchläuft jeder Bildausschnitt die Kaskaden von vorne nach hinten, wobei in jeder Kaskadenstufe entschieden wird, ob es sich um ein Gesicht handelt. In diesem Fall wird der Ausschnitt an die folgende Kaskadenstufe weitergegeben und weiter analysiert. Ansonsten wird der Ausschnitt verworfen und nicht weiter betrachtet. Die Ausschnitte, die die gesamte Kaskade durchlaufen haben, sind die gefundenen Gesichter.

3.2.4 Vergleichende Untersuchungen

3.2.4.1 Trainingsdaten

Um eine vergleichende Bewertung der verschiedenen Verfahren vornehmen zu können, müssen für die Modellerstellung vergleichbare Daten verwendet werden. Die hierfür verwendete Datenbank mit positiven Trainingsbeispielen besteht aus 2610 Bildern mit einer Auflösung von 640×480 Pixeln. Die Bilder stammen von 87 verschiedenen Personen beiderlei Geschlechts. Hierbei wurde jeweils nur der Kopf- und Schulterbereich aufgenommen, wobei der Winkel des Gesichts zur Kamera bei jeder Person in gleicher Weise variiert wurde. Es wurden jeweils verschiedene Aufnahmen gemacht, bei denen das Gesicht um $-30°$, $-15°$, $0°$, $15°$ und $30°$ Grad horizontal und um $-20°$, $0°$ und $20°$ Grad vertikal gegenüber der Kameraachse verdreht war. Zusätzlich wurde die Beleuchtung der Szene zwischen natürlichem und künstlichem Licht variiert. Damit ergeben sich 30 Aufnahmen pro Person. Einige Beispiele sind in Abbildung 3.10 zu sehen.

Abbildung 3.10: *Einige Beispiele aus der Gesichtsdatenbank, die zum Generieren der Trainingsdaten verwendet wurde.*

Um die Detektor trainieren zu können, müssen die vorhandenen Bilder an die jeweils erforderliche Größe der Trainingsdaten angepasst werden. Daher ist es nicht möglich, für jedes Verfahren exakt die gleichen Daten zu verwenden. Die negativen Trainingsbeispiele wurden aus einer Sammlung von Bildern der Größe 640×480 generiert, die in einem Baumarkt aufgenommen wurden. Da sowohl beim AdaBoost- als auch beim Rowley-Verfahren die Auswahl der negativen Beispiele während des Trainings zufällig erfolgt, kann nicht gewährleistet werden, dass alle Verfahren identische Trainingsdaten verwenden.

3.2.4.2 Testdaten

Zur Bewertung der Detektionsleistung kam bei allen Verfahren der gleiche Testdatensatz zum Einsatz. Dabei handelt es sich um 118 Bilder der Größe 640×480, die in einem Baumarkt während des normalen Betriebes aufgenommen wurden. Dieser Datensatz enthält insgesamt 115 Gesichter verschiedener Größe und sehr unterschiedlicher Orientierung. Da zudem die Ausleuchtung der Gesichter stark variiert und der Hintergrund auf den meisten Bildern sehr stark strukturiert ist, stellt dieser Testdatensatz eine große Herausforderung für ein Gesichtsdetektionsverfahren dar. Deshalb sind Detektionsraten von über 90%, wie sie in den jeweiligen Originalpublikationen angegeben sind, nicht zu erwarten. Einige Beispiele aus dem Testdatensatz sind in Abbildung 3.11 zu sehen.

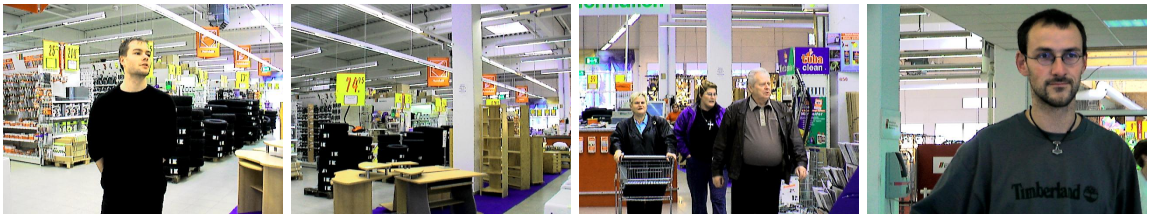


Abbildung 3.11: *Einige Beispielbilder aus dem Testdatensatz für die Gesichtsdetektion. Die Bilder zeichnen sich durch stark strukturierte Hintergründe und unterschiedliche Beleuchtungsverhältnisse aus.*

3.2.4.3 Ergebnisse

Im Testdatensatz existiert zu jedem zu untersuchenden Bild eine Liste von Koordinaten, die für jedes im Bild enthaltene Gesicht zwei Quadrate definieren. Zunächst wird das kleinste mögliche Rechteck berechnet, welches alle vorher gelabelten Merkmale des Gesichtes (die Mittelpunkte der Augen, die Nasenspitze, beide Mundwinkel sowie die Mitte des Mundes) umschließt. Der Mittelpunkt dieses Rechtecks definiert den Mittelpunkt der Kontrollquadrate. Das kleinere Quadrat hat eine Kantenlänge, die der Länge der längeren Seite dieses Rechtecks entspricht. Das größere Quadrat hat die vierfache Kantenlänge. Eine Gesicht gilt dann als korrekt detektiert, wenn die Detektion das kleinere Kontrollquadrat umschließt und vollständig innerhalb des Großen liegt, siehe Abbildung 3.12.

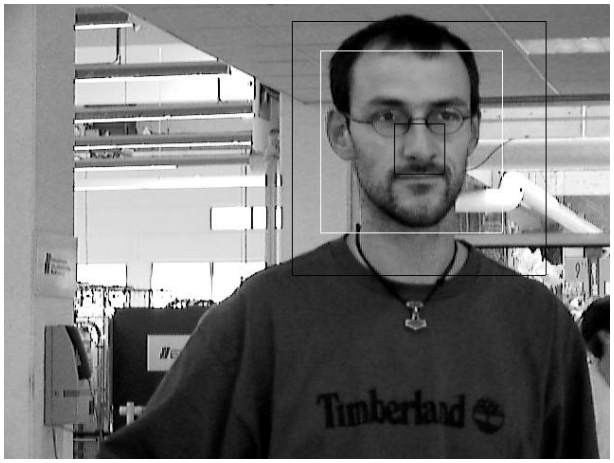


Abbildung 3.12: *Auswertung der Gesichtsdetektion. Die beiden schwarzen Rahmen umschließen den Bereich, in dem eine Gesichtsdetektion als positiv gewertet wird, d.h. der vom Gesichtsdetektor gelieferte weiße Rahmen muss sich wie in diesem Beispiel zwischen den beiden schwarzen Rahmen befinden.*

Mögliche Fehler bei der Gesichtsdetektion sind zum einen Falsch-Negative, bei denen Gesichter nicht detektiert werden und Falsch-Positive, bei denen ein Gesicht in einer Bildregion detektiert wird, wo keines ist. Diese Fehler werden durch die Detektionsrate und die Falsch-Positiv-Rate ausgedrückt. Die Detektionsrate wird definiert als das Verhältnis zwischen der Anzahl korrekt detektierter Gesichter und der Anzahl der von einem Menschen detektierten Gesichter. Die Falsch-Positiv-Rate ist die Anzahl von Falsch-Positiv-Detektionen, bezogen auf die mögliche

Anzahl von Detektionen. Letztere lag beim verwendeten Testdatensatz mit 118 Bildern der Größe 640×480 bei 36249600, wobei unterschiedliche Detektionsgrößen unberücksicht blieben. Bei z.B. 200 Falsch-Positiv-Detektionen auf dem gesamten Datensatz ergibt dies eine Falsch-Positiv-Rate von $0,00055173\%$, also rund $1,7$ Falsch-Positiv-Detektionen pro Bild. Ein guter Gesichtsdetektor zeichnet sich durch eine hohe Detektionsrate bei gleichzeitig möglichst geringer Falsch-Positiv-Rate aus, wobei jedes Verfahren so parametriert werden kann, dass eines der beiden Gütemaße optimiert wird.

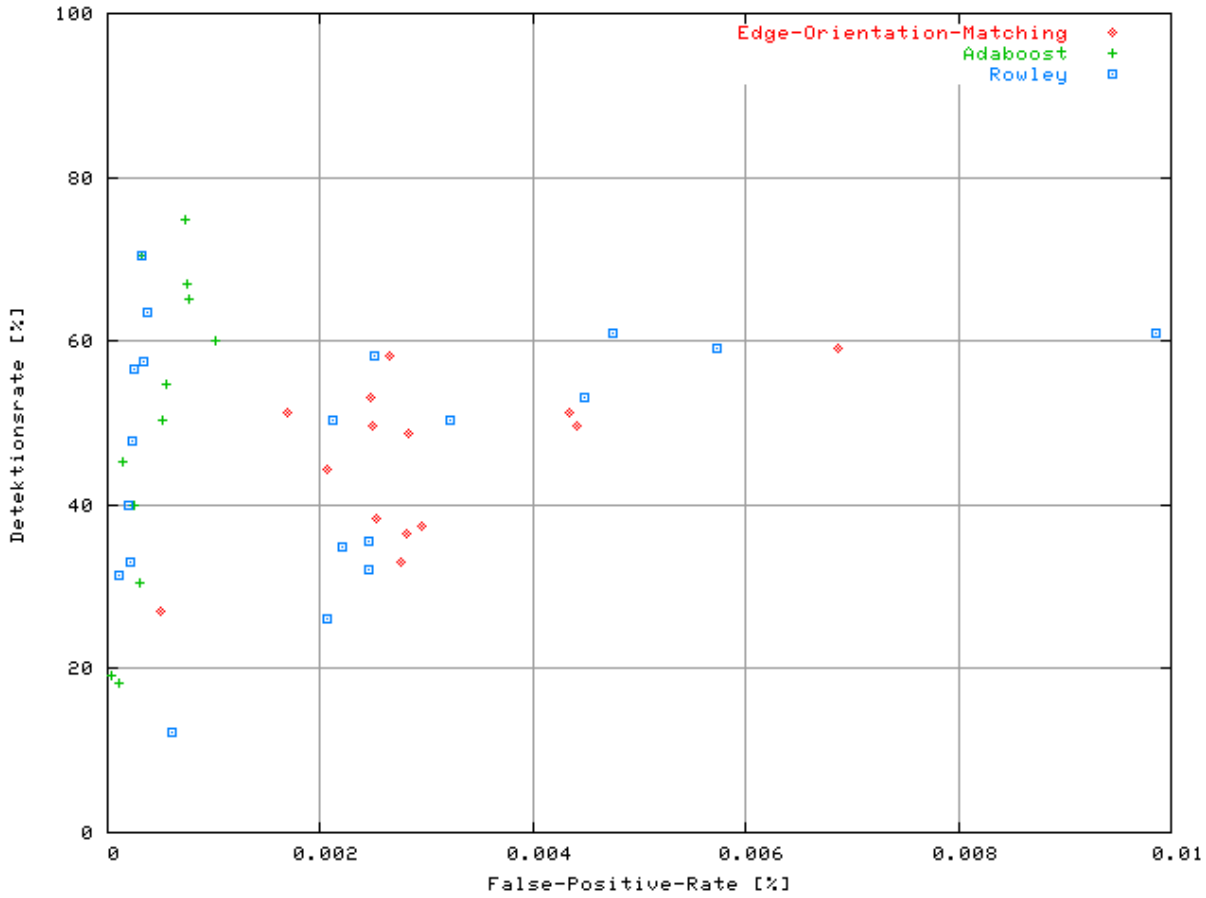


Abbildung 3.13: *Detektionsraten und Falsch-Positiv-Raten der drei Gesichtsdetektoren auf dem Testdatensatz. Die Detektions- und Falsch-Positiv-Raten wurden jeweils für verschiedene Parameterkombinationen ermittelt. Es wird ersichtlich, dass AdaBoost und das Verfahren nach Rowley ähnlich gute Detektionsraten erreichen, wobei die Falsch-Positiv-Raten bei Rowley geringfügig besser sind. Wesentlich schlechter schneidet das Edge-Orientation-Matching ab.*

In verschiedenen Testreihen, bei denen sowohl die Parameter beim Training als auch bei der Anwendung der Gesichtsdetektoren variiert wurden, wurden die in Abbildung 3.13 dargestellten Detektions- und Falsch-Positiv-Raten auf dem Testdatensatz ermittelt. Bei jedem Verfahren wurden neben der Anzahl der Auflösungsstufen und der Skalierung zwischen den Stufen diverse Parameter und Schwellwerte variiert. Dies geschah zunächst so, dass jeweils ein Parameter geändert wurde, während die anderen konstant gehalten wurden. Aus den hieraus gewonnenen Erkenntnissen wurde versucht, eine optimale Parameterkonfiguration zu finden, die die Detektionsleistung maximiert. Abbildung 3.13 zeigt die besten Ergebnisse für die drei Verfahren.

Es wird ersichtlich, dass das Verfahren nach Rowley bei vergleichbaren Detektionsraten geringfügig bessere Falsch-Positiv-Raten erreicht als AdaBoost, während das Edge-Orientation-Matching wesentlich schlechter abschneidet. Dieses Ergebnis steht im Gegensatz zu ersten Untersuchungen, bei denen der Detektor von Rowley sehr schlecht abschnitt, [Wilhelm et al., 2003c].

Das Ergebnis ist nicht verwunderlich, wenn man bedenkt, dass das Trainingsregime für die ersten beiden Verfahren durch Bootstrapping die Anzahl von Falsch-Positiv-Detektionen minimiert. Im Gegensatz dazu wird das Template beim Edge-Orientation-Matching nur aus Positiv-Beispielen erstellt. Außerdem approximieren die beiden appearance-basierten Verfahren den Face-Space mit größerer Genauigkeit als das Template-Matching-Verfahren, das ja nur die größten Ähnlichkeiten zum Template des mittleren Gesichtes bestimmt.

3.2.4.4 Test auf rotierten Gesichtern

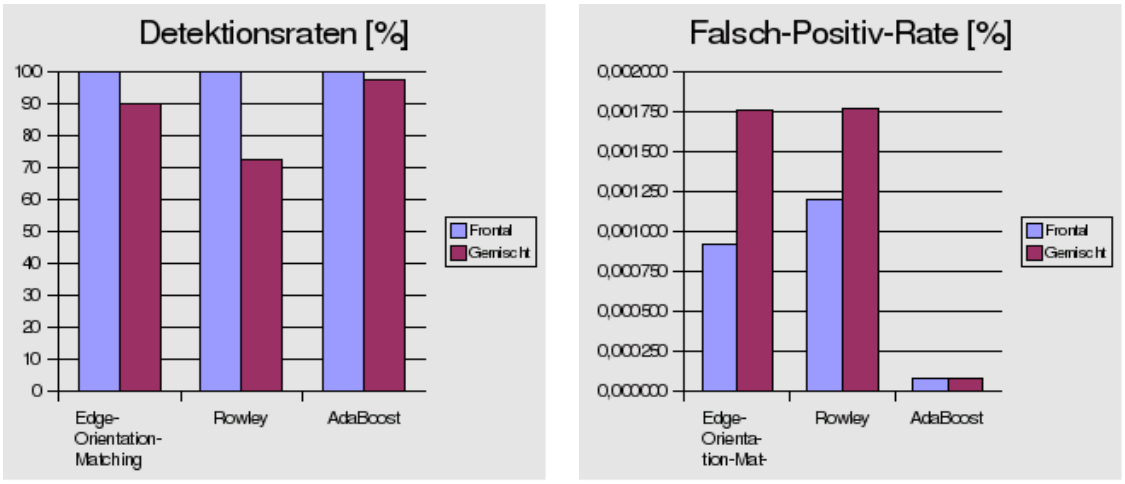
Die verwendeten Testbilder aus dem Baumarkt stellen sehr hohe Ansprüche an die Gesichtsdetektoren, was die verhältnismäßig niedrigen Detektionsraten bei allen Verfahren erklärt. Abschließend soll die grundsätzliche Leistungsfähigkeit anhand von Testreihen unter idealen Bedingungen beurteilt werden. Zu diesem Zweck wurden zwei zusätzliche Testdatensätze zusammengestellt, die aus jeweils 40 Bildern bestehen. Diese entstammen der selben Datenbank, aus der auch die Positiv-Beispiele für das Training entnommen wurden. Sie wurden also unter kontrollierten Bedingungen vor einem homogenen Hintergrund aufgenommen und enthalten jeweils nur ein Gesicht. Datensatz 1 enthält nur frontale, Datensatz 2 unterschiedlich orientierte Gesichter. Es wurden die im vorherigen Test jeweils besten gefundenen Parametersätze verwendet.

Datensatz	1		2	
Detektor	ER (%)	FPR (%)	ER (%)	FPR (%)
Edge-Orientation	100	0.00092	90.0	0.00176
Rowley	100	0.00120	72.5	0.00177
AdaBoost	100	0.00008	97.5	0.00008

Tabelle 3.1: *Performanz auf den beiden unter kontrollierten Bedingungen aufgenommenen Testdatensätzen, einmal mit frontalen Ansichten und einmal mit unterschiedlichen Orientierungen. Dargestellt ist jeweils die Detektionsrate (ER) und die Falsch-Positiv-Rate (FPR).*

Tabelle 3.1 und Abbildung 3.14 zeigt die Detektions- und Falsch-Positiv-Raten der drei untersuchten Detektoren auf diesen Datensätzen. Wie zu erwarten war, ist die Detektionsleistung auf den frontalen Gesichtern signifikant besser als auf den Testdaten aus dem Baumarkt. Jedes Verfahren erreicht hier eine Detektionsrate von 100%. Die schlechteste Falsch-Positiv-Rate liefert das Rowley-Verfahren. In diesem Test schneidet der AdaBoost-Detektor mit einer Falsch-Positiv-Rate von 0.00008 am besten ab.

Beim Test auf unterschiedlich orientierten Gesichtern zeigen sich gravierende Unterschiede zwischen den Verfahren. Bei AdaBoost fällt die Detektionsrate nur geringfügig, die Falsch-Positiv-Rate bleibt bemerkenswerterweise sogar gleich. Etwas schlechter fällt das Ergebnis beim Edge-



(a) (b)

Abbildung 3.14: *Detektions- und Falsch-Positiv-Raten aus Tabelle 3.1.*

Orientation-Matching aus, die erreichten Detektionsraten sind aber immer noch als gut zu bezeichnen. Da die Personen auf den Bildern nicht mehr frontal, sondern seitlich fotografiert wurden, sind auf den Bildern Teile des Haaransatzes, die Ohren usw. zu erkennen. Aufgrund der fehlenden Robustheit des Verfahrens führt dies zu einer signifikant höheren Falsch-positiv-Rate. Beim Rowley-Detektor bricht die Detektionsrate im Vergleich zum vorhergehenden Test am stärksten ein, es zeigt große Probleme bei der Detektion seitlicher Gesichtsansichten. Damit einher geht, ähnlich wie beim Edge-Orientation-Matching, ein starker Anstieg der Falsch-Positiv-Rate.

3.2.5 Fazit

Aufgrund der guten Ergebnisse auf dem Testdatensatz aus dem Baumarkt und den im Vergleich zu den anderen Verfahren sehr guten Ergebnissen auf den Datensätzen mit frontalen und seitlichen Ansichten wurde AdaBoost für den Einsatz im Gesamtsystem ausgewählt. Hinzu kommt die Tatsache, dass AdaBoost aufgrund seiner Kaskadenstruktur der mit Abstand schnellste unter den getesteten Gesichtsdetektoren ist. Da die Implementierungen im Hinblick auf Optimierungen sehr unterschiedlich ausfallen, wurden zeitliche Betrachtungen nicht mit in diese Arbeit aufgenommen. Die allgemeine Beobachtung, dass AdaBoost schneller ist als die anderen getesteten Verfahren, behält aber sicherlich Gültigkeit.

3.3 Detektion von Gesichtsmerkmalen

3.3.1 Aufgabenstellung

Die im fovealen Vision-System eingesetzten Verfahren zur Analyse der Gesichtsbilder benötigen entweder sehr exakt ausgerichtete Gesichtsausschnitte, da sie mit nicht-adaptiven Modellen arbeiten, oder sie passen ein adaptives Modell auf ein Gesicht an. Auch im letzteren Fall ist es aber von großem Vorteil, wenn die detektierten Gesichter zunächst genauer ausgerichtet werden, da dann die Modellanpassung vereinfacht und somit auch beschleunigt wird. Um eine solche exakte Ausrichtung zu erreichen, müssen saliente Strukturen im Gesicht gesucht und das Gesichtsbild anhand dieser mittels einer affinen Transformation ausgerichtet werden. In den nächsten Abschnitten werden hierfür geeignete Strukturen und Verfahren für deren Detektion ausgewählt.

3.3.2 Geeignete Strukturen

Zunächst muss festgelegt werden, wie viele und welche Strukturen im Gesicht detektiert werden sollen. Als Minimum sind dabei zwei Punkte nötig, da es damit möglich ist, verschiedene Gesichtsbilder auf eine einheitliche Größe und in eine einheitliche Lage zu bringen. Hierfür bieten sich die Augen an, da sie eine besonders auffällige Struktur aufweisen und zudem, anders als der Mund, auch relativ forminvariant sind. Abbildung 3.15 zeigt einen schematischen Ablauf der Transformation. Diese läuft in den folgenden Schritten ab:

Translation: Im Originalbild befindet sich der Koordinatenursprung in der linken oberen Ecke des Bildes. Da die Rotation um den Mittelpunkt zwischen den Augen erfolgen soll, wird dieser Punkt in den Koordinatenursprung verschoben.

Rotation: Für die Rotation wird der Winkel zwischen der Verbindungslinie der Augen und der Waagerechten verwendet.

Skalierung: Für die Skalierung mit zwei Punkten wird das Verhältnis zwischen dem Augenabstand im Ziel- und im Originalbild bestimmt. Dieses wird als Skalierungsfaktor in x- und in y-Richtung verwendet. Werden drei Punkte verwendet, wird zusätzlich ein Skalierungsfaktor in y-Richtung ermittelt. Dazu wird der Abstand zwischen dem Mittelpunkt zwischen den Augen und der Nasenspitze im Ziel- und im Originalbild bestimmt. Das Verhältnis zwischen diesen ist der zu verwendende Skalierungsfaktor in y-Richtung.

Translation: Zum Schluss wird der Koordinatenursprung wieder in die linke obere Ecke des Zielbildes verschoben.

Da eine Vorwärtstransformation von Pixeln vom Originalbild zum Zielbild Lücken im Zielbild erzeugen würde, wird eine Rückwärtstransformation der Pixel des Zielbildes durchgeführt und deren Grauwerte aus dem Mittelwert der jeweils vier nächsten Nachbarn im Originalbild gebildet. Dazu wird die berechnete Transformationsmatrix invertiert.

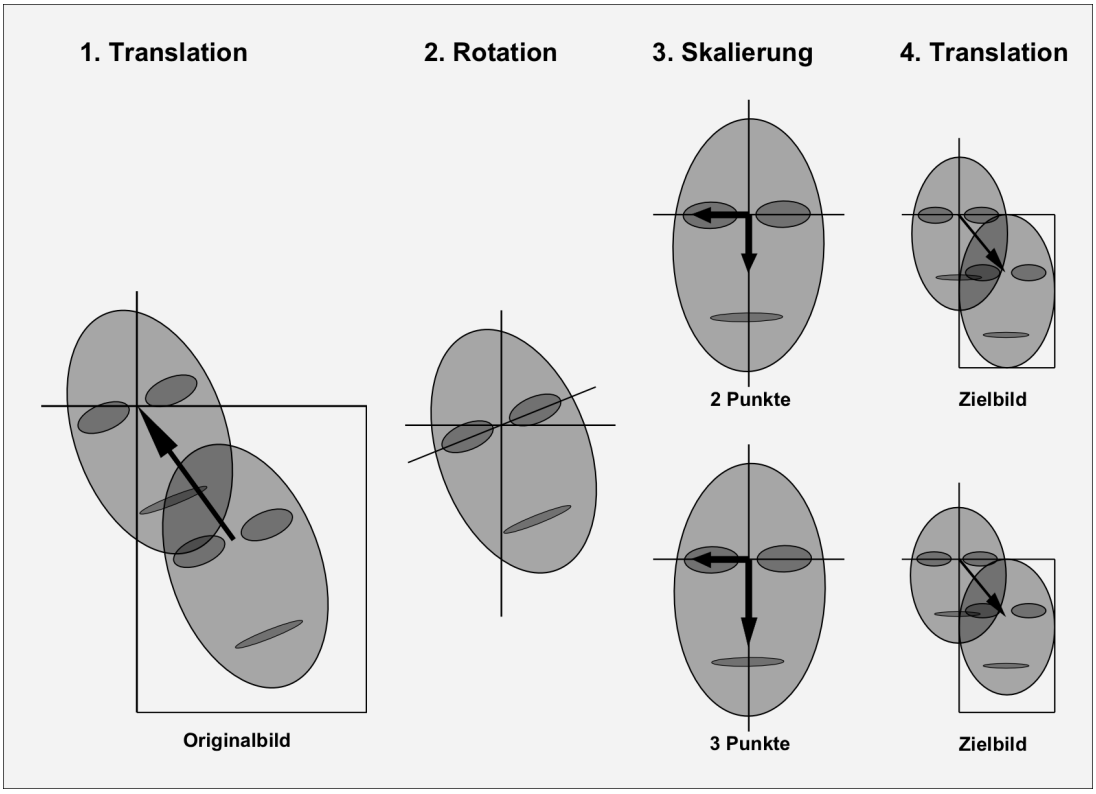


Abbildung 3.15: *Bestimmung einer affinen Transformation. Die Augen sollen im Zielbild an festen Positionen platziert werden. Aus dem Verhältnis zwischen Augenabstand im Zielbild und im Originalbild kann eine Skalierung bestimmt werden, die in x- und in y-Richtung gleich groß ist. Bei der Verwendung von 3 Punkten wird neben den Augen ein dritter Punkt gesucht, der im Zielbild wiederum immer auf einer festen Stelle platziert wird (z.B. die Nasenspitze). Es wird also eine unterschiedliche Skalierung in x- und in y-Richtung bestimmt.*

Durch die Projektion des Kopfes auf die Projektionsfläche eines zweidimensionalen Aufnahmesystems entstehen aufgrund der Drehung des Kopfes unterschiedliche Abbildungen des Gesichtes. Man unterscheidet zwischen in-plane Rotation, bei der die Drehachse senkrecht auf der Abbildungsfläche steht und der out-of-plane Rotation, bei der die Drehachse nicht senkrecht zur Abbildungsfläche steht. Die out-of-plane Rotation führt zur teilweisen Verdeckung, zur ungleichmäßigen Skalierung von Gesichtsmerkmalen.

Es wäre zu vermuten, dass durch die Verwendung eines dritten Punktes und der Bestimmung einer unterschiedlichen Skalierung in x- und in y-Richtung eine exaktere Positionierung des Gesichtes möglich wäre, so dass z.B. auch die Nasenspitze auf eine feste Position platziert werden könnte. Diese Vorgehensweise würde aber mehrere gravierende Nachteile mit sich bringen. Zum einen kommt es mit jedem zusätzlich zu suchenden Merkmal mit größerer Wahrscheinlichkeit auch zu Fehldetektionen, wodurch alle nachfolgenden Verarbeitungsschritte keine brauchbaren Eingabedaten erhalten würden. Zum anderen wird die Form des Gesichtes verändert, d.h. lange schmale Gesichter werden in der Länge gestaucht und breite kurze Gesichter werden gestreckt, siehe Abbildung 3.16. Außerdem ist unklar, welchen Einfluss diese Änderungen auf die Analyse

der Gesichtsbilder haben, da bei einer Stauchung bzw. Streckung des Gesichtes unter Umständen für die Klassifikation relevante Informationen verworfen werden, z.B. bei der Identitätsschätzung.

(a) (b)

Abbildung 3.16: *In (a) wurden die Gesichter anhand von zwei Gesichtsmerkmalen ausgerichtet, d.h. es wurde die selbe Skalierung in x- und y-Richtung verwendet. Es ist deutlich zu erkennen, dass die Positionen der Nasenspitzen und die Münder sehr stark variieren. In (b) wurde anhand der Nasenspitze eine separate Skalierung in y-Richtung bestimmt, so dass die Nasenspitze immer auf die selbe Position im Zielbild platziert werden kann. Da der Abstand vom Mittelpunkt zwischen den Augen zur Nasenspitze von Person zu Person sehr unterschiedlich ist, führt dies zu einer nicht erwünschten Stauchung bzw. Streckung des Gesichtes und zu einer sehr großen Variation in der Position des Mundes.*

Weiterhin wirkt sich, wie in [Lyons et al., 2000] beschrieben, eine vertikale Drehung eines Gesichtes sehr stark auf die Erscheinung von Gesichtsausdrücken aus. Durch die konvexe Form in der Mundregion erscheint der Mund bei einem nach oben geneigten Gesicht als traurig (Mundwinkel nach unten) und bei einem nach unten geneigten Gesicht als fröhlich (Mundwinkel nach oben). Im japanischen Theater wird dieser Effekt ausgenutzt, um mit statischen Masken Gesichtsausdrücke und somit Gefühle ausdrücken zu können, siehe Abbildung 3.17.

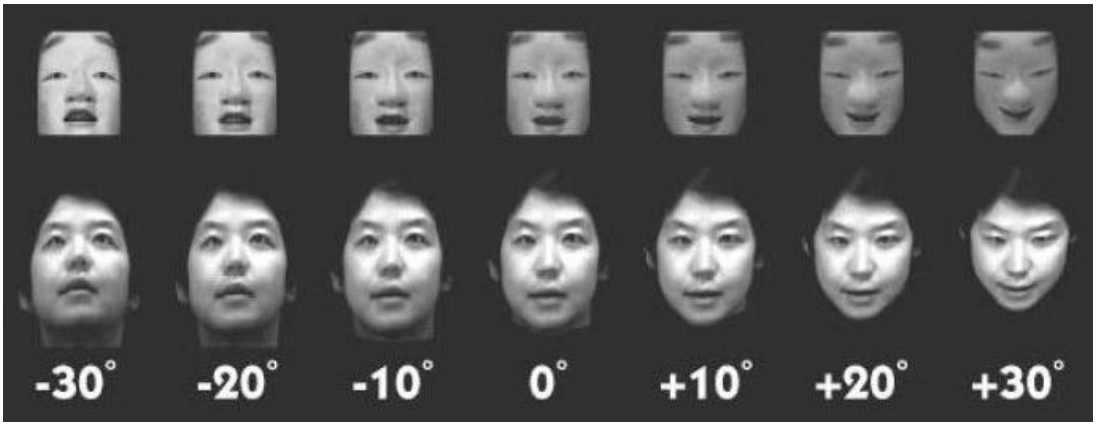


Abbildung 3.17: *Der Noh-Mask-Effekt. Zu sehen ist die Wirkung einer vertikalen Drehung eines Gesichtes auf die Wahrnehmung von Gesichtsausdrücken. Ein neutraler Gesichtsausdruck (Mitte) erscheint, wenn das Gesicht von oben gesehen wird, als freundlich (rechts) bzw. von unten als traurig (links). Durch die besondere Hervorhebung von Konturen bei der japanischen Noh-Maske wird dieser Effekt noch verstärkt, er ist aber auch bei natürlichen Gesichtern zu beobachten. Abbildung übernommen aus [Lyons et al., 2000].*

Um eine möglichst exakte Ausrichtung des Gesichtes unter Umgehung der eben aufgezählten Nachteile zu erreichen, wäre ein $3D$-Modell der Gesichtsstrukturen notwendig. Für die verwendeten Verfahren ist es nicht sinnvoll, einen dritten Punkt im Gesicht zu suchen und separate Skalierungen in x- und y-Richtung zu bestimmen. Die Gesichtsnormalisierung wird also anhand der Augenpunkte durchgeführt werden.

In [Lien et al., 1998] werden 3 Merkmalspunkte für die Normalisierung der Gesichter verwendet, zwei befinden sich an der Innenseite der Augen und einer an der Unterkante der Nase. Diese Art der Normalisierung ist notwendig, da es sich um ein bewegungsbasiertes Verfahren zur Klassifikation von Gesichtsausdrücken handelt, dass für die Flussschätzung eine sehr exakte Normalisierung aufeinander folgender Bilder voraussetzt. Allerdings handelt es sich um ein semi-automatisches System, bei dem die entsprechenden Labelpunkte von Hand gesetzt werden müssen.

3.3.3 Untersuchte Verfahren

Für die Augendetektion können prinzipiell die selben Verfahren eingesetzt werden wie für die Gesichtsdetektion. Aufgrund der speziell auf Gesichter angepassten Struktur der rezeptiven Felder beim Verfahren von Rowley ist dieses ohne entsprechende Modifikationen nicht geeignet. Es werden also das Edge-Orientation-Matching und der AdaBoost-Detektor vergleichend untersucht, wobei hier aufgrund der Aufgabenstellung, anders als bei der Gesichtsdetektion, die Positioniergenauigkeit im Vordergrund steht.

3.3.4 Vergleichende Untersuchungen

3.3.4.1 Genauigkeit des Gesichtsdetektors

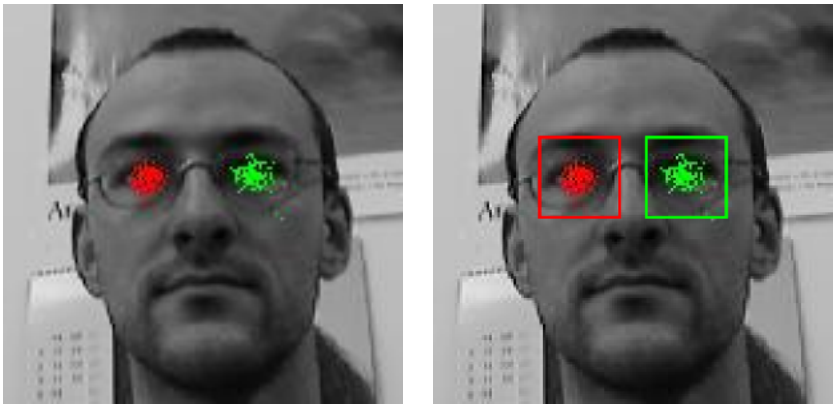
Als erstes wurde ermittelt, wie genau die Positionierung der Augen bei ausschließlicher Verwendung des Gesichtsdetektors ist. Dazu wurde über einem Satz von 226 vom Gesichtsdetektor gelieferten Bildern die mittlere Position der Augen, die Varianz und die Maximal- und Minimalwerte der Positionen bestimmt. Die Ergebnisse sind in Tabelle 3.2 zu sehen. Abbildung 3.18(a) zeigt die Verteilung dieser Positionen auf einem der Bilder aus dem Datensatz.

links	min	max	μ	σ^2
x	47	61	53.7	8.4
y	58	73	65.1	8.2

rechts	min	max	μ	σ^2
x	85	105	93.4	11.7
y	56	80	63.8	9.8

Tabelle 3.2: *Es wurden für das linke und das rechte Auge jeweils die minimale und die maximale Position, der Mittelwert und die Varianz in x- und in y-Richtung bestimmt.*

Um möglichst optimale Bedingungen für die Algorithmen zur Analyse der Gesichtsstruktur zu schaffen, soll untersucht werden, ob durch eine Detektion der Augen und eine anschließende affine Transformation eine genauere Positionierung des Gesichtes erreicht werden kann. Die Be-



(a) (b)

Abbildung 3.18: *(a) Es wurden die Augenpositionen für einen Datensatz mit 226 Bildern ermittelt und beispielhaft in ein Bild dieses Datensatzes eingezeichnet. (b) Aufgrund der maximalen und minimalen Augenpositionen definierter Suchbereich für das linke und für das rechte Auge.*

reiche, in denen nach Augen gesucht werden muss, können dabei anhand der Werte in Tabelle 3.2 eingeschränkt werden. Neben einer Reduzierung der Rechenzeit führt dies auch zu einer Verringerung der Anzahl an Fehldetektionen. Tabelle 3.3 und Abbildung 3.18(b) zeigen die definierten Suchbereiche.

links	min	max
x	40	70
y	50	80

rechts	min	max
x	80	110
y	50	80

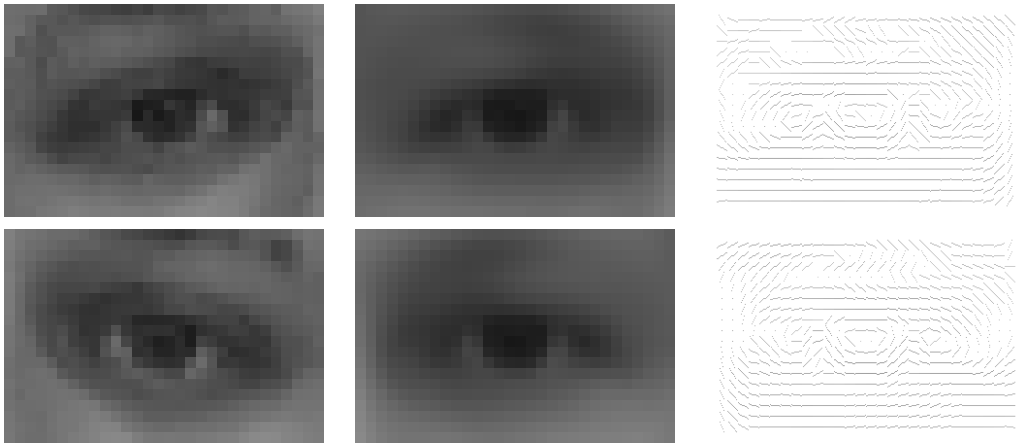
Tabelle 3.3: *Die Suchbereiche für das linke und das rechte Auge überdecken jeweils 30 Pixel in x- und in y-Richtung. Das sind 12.5% des gesamten Bildes.*

3.3.4.2 Trainingsdaten

Die Trainingsdaten für die Modellerstellung wurden aus der im nächsten Abschnitt beschriebenen NIFace2-Datenbank extrahiert. Dazu wurden die Augen anhand der Labelpunkte aus allen Bildern der Datenbank ausgeschnitten und auf eine Größe von 30×20 Pixeln skaliert, wodurch jeweils 700 Positivbeispiele für das linke und das rechte Auge zur Verfügung standen. Als Negativbeispiele, die für das Training des AdaBoost-Detektors notwendig sind, wurden die Testdaten für das Training des Gesichtsdetektors verwendet, wobei aber die Augen ausgeblendet wurden. Zum Test wurden die Detektoren auf den selben Datensatz angewandt, auf dem die Genauigkeit des Gesichtsdetektors bestimmt wurde, wobei die Abweichungen der detektierten Positionen von den von Hand markierten Positionen ermittelt wurden. Zum einen soll der Mittelwert hier möglichst Null sein, das Merkmal also an der richtigen Position gefunden werden und zum anderen sollen die Varianzen kleiner sein als die aus Tabelle 3.2. Ein Auge gilt als nicht detektiert, wenn es außerhalb eines Radius von 10 Pixeln um die wirkliche Position gefunden wurde.

3.3.4.3 Edge-Orientation-Matching

Aus den Trainingsdaten wird, wie schon beim Gesichtsdetektor beschrieben, zunächst ein Mittelwertbild berechnet, aus dem dann ein Kantenorientierungsmodell berechnet wird. Abbildung 3.19 veranschaulicht dies für das linke und das rechte Auge.



(a) (b) (c)

Abbildung 3.19: *Erstellung des Templates für die Augendetektion. (a) Beispiele aus den Trainingsdaten. (b) linkes und rechtes Mittelwertauge. (c) Kantenorientierungs-Templates für das linke und das rechte Auge.*

Auf dem Testdatensatz wurden sowohl das linke als auch das rechte Auge jeweils 5 mal nicht gefunden. Die Positionierungsgenauigkeit ist in Tabelle 3.4 dargestellt. Im Vergleich zu Tabelle 3.2 kann eine deutliche Verbesserung festgestellt werden.

links	μ	σ^2
x	1.26	4.96
y	-0.01	2.88

rechts	μ	σ^2
x	-0.18	8.57
y	-0.50	4.15

Tabelle 3.4: *Mittelwerte und Varianzen der Abweichung der Augenpositionen von den wahren Positionen bei Verwendung des Edge-Orientation-Matching.*

3.3.4.4 AdaBoost

Beim Vergleich der Gesichtsdetektoren wurde AdaBoost als deutlich überlegenes Verfahren identifiziert und schließlich für die Anwendung im Gesamtsystem ausgewählt. Daher war es nahe liegend, auch für die Augendetektion auf AdaBoost zurückzugreifen. Für das Training dieses Detektors wurden die Positivbeispiele auf eine Auflösung von 15×10 reduziert. Auf dem Testdatensatz wurde das linke Auge 4 mal und das rechte Auge 5 mal nicht gefunden. Die Positionierungsgenauigkeit ist in Tabelle 3.5 dargestellt. Im Vergleich zu Tabelle 3.2 kann eine deutliche Verbesserung festgestellt werden.

links	μ	σ^2	**rechts**	μ	σ^2
x	1.27	7.57	**x**	-0.35	5.37
y	-0.83	6.37	**y**	-0.27	3.26

Tabelle 3.5: *Mittelwerte und Varianzen der Abweichung der Augenpositionen von den wahren Positionen bei Verwendung des AdaBoost-Augendetektors.*

3.3.4.5 Fazit

Der AdaBoost-Detektor schneidet bei der Lokalisierung der Augen nicht so eindeutig besser ab wie dies bei der Gesichtsdetektion der Fall war. Beim linken Auge erwies sich das Edge-Orientation-Matching als besser, beim rechten Auge dagegen AdaBoost. Die beiden Verfahren sind bei der Positioniergenauigkeit in etwa gleichwertig, wobei AdaBoost eine Fehldetektion weniger liefert. Abbildung 3.20 zeigt einige Detektionsergebnisse der beiden Verfahren. Für das Gesamtsystem wird auch hier, insbesondere wegen seines geringeren Rechenzeitbedarfs, auf den AdaBoost-Detektor zurückgegriffen.

(a) (b) (c) (d)

Abbildung 3.20: *Einige Beispiele für detektierte Augen vom Edge-Orientation-Matching in der oberen Reihe und von AdaBoost in der unteren Reihe. Die gelabelten Positionen sind schwarz und die detektierten weiß eingezeichnet. Das Gesicht selbst ist leicht aufgehellt, damit die Detektionen besser sichtbar sind.*

3.4 Ansteuerung der PTU

Nachdem das foveale Vision-System eine Person detektiert hat, übernimmt es die Ansteuerung der PTU, da so eine wesentlich genauere Positionierung erreicht werden kann. Das Ziel der Regelung besteht darin, das detektierte Gesicht in der Mitte des von der Frontalkamera aufgenommenen Bildes zu platzieren, siehe Abbildung 3.21.

Die Werte d_x und d_y werden mit einem Faktor k multipliziert und dazu verwendet, den Schwenk-

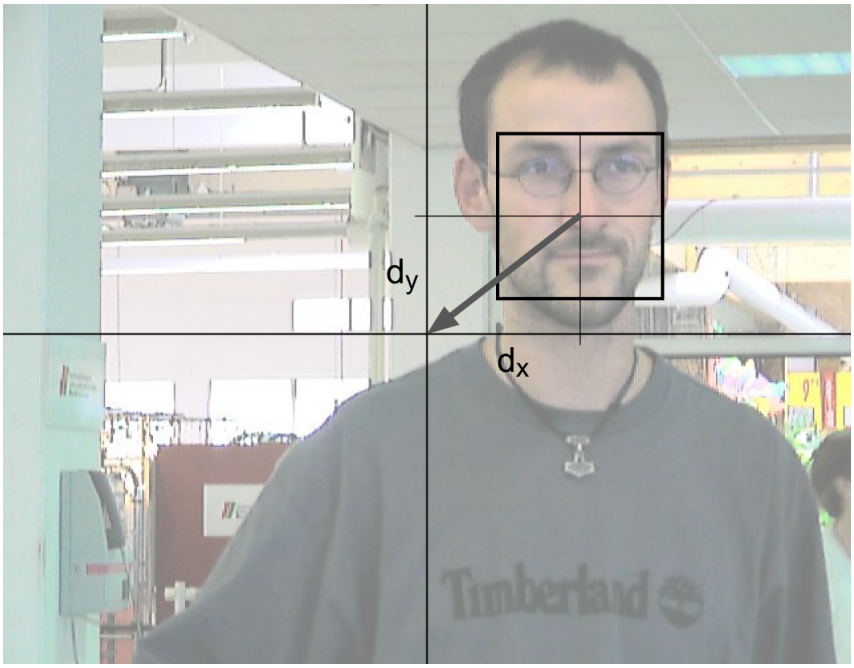


Abbildung 3.21: *Ansteuerung der PTU durch das foveale Vision-System. Das Ziel besteht darin, das Gesicht des Nutzers in der Mitte des Bildes zu platzieren. Dazu wird der Abstand der Gesichtsdetektion von der Bildmitte bestimmt.*

bzw. Neigewinkel der PTU anzupassen. Dazu wird kd_x vom aktuellen Schwenk- und kd_y vom aktuellen Neigewinkel subtrahiert. Der Faktor k bestimmt die Geschwindigkeit, mit der der Kopf des Roboters nachgeführt wird. Durch diese PTU-Regelung wird eine einmal detektierte Person kontinuierlich angeschaut. Dies ist besonders während der Interaktion wichtig, damit der Roboter nicht von anderen vom peripheren Vision-Systems ermittelten auffälligen Bildregionen abgelenkt wird. Mit der PTU-Ansteuerung des fovealen Vision-Systems kann ein Gesicht getrackt werden, solange die Positionsänderung von einem Bild zum nächsten nicht so groß ist, dass das Gesicht nicht mehr im Bild erscheint. Sobald das foveale Vision-System kein Gesicht mehr detektiert, wird die Kontrolle der PTU an das periphere Vision-System zurückgegeben. Dadurch kann der Interaktionspartner wieder gefunden werden, wenn er sich zu weit bewegt hat, bzw. beginnt der Roboter damit, nach einem anderen Interaktionspartner zu suchen. Um einzelne Ausfälle des Gesichtsdetektors zu kompensieren, erfolgt die Rückgabe der Kontrolle der PTU an das periphere Vision-System erst, nachdem für drei Bilder in Folge kein Gesicht detektiert wurde.

Kapitel 4

Nutzeranalyse

4.1 Aufgabenstellung

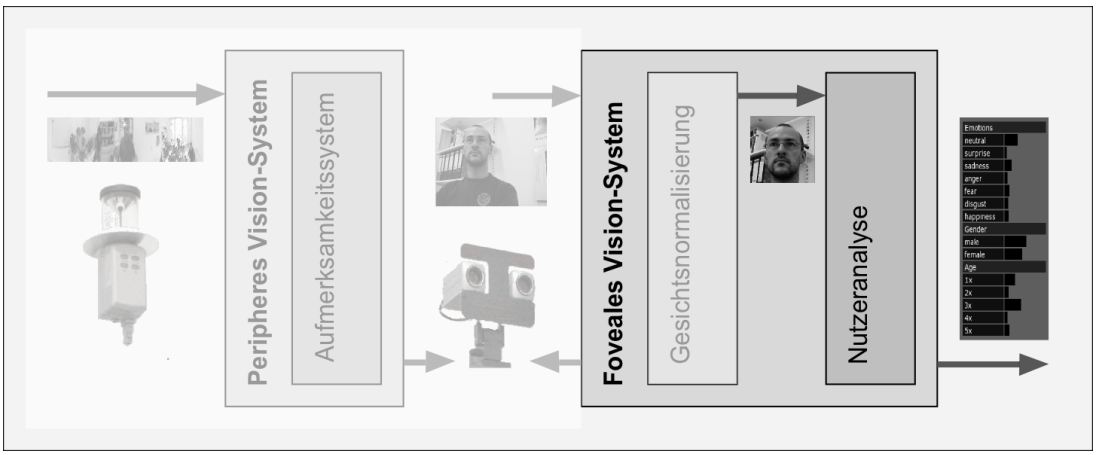


Abbildung 4.1: *Einordnung der Nutzeranalyse in die Systemarchitektur. Die Nutzeranalyse ist die zweite Komponente des fovealen Vision-Systems. Hier wird die normalisierte Darstellung des Gesichtes hinsichtlich Mimik, Alter, Geschlecht und Identität der Person analysiert.*

Abbildung 4.1 zeigt eine Einordnung der Nutzeranalyse in das Systemarchitektur. Nachdem das Gesicht im hochaufgelösten Bild der Frontalkamera gefunden und in eine normalisierte Darstellung gebracht wurde, sollen detaillierte Informationen über den Nutzer wie Alter, Geschlecht, Gesichtsausdruck und Identität, ermittelt werden. Dazu werden aus der normalisierten Darstellung des Gesichtes Merkmale extrahiert und hinsichtlich dieser Kategorien klassifiziert.

4.2 Literatur

Die Anwendung von Methoden der Bildverarbeitung für die Analyse von Gesichtern hat in den letzten Jahren zunehmend an Bedeutung gewonnen. Dies rührt sicherlich von den vielfältigen Anwendungen auf diesem Gebiet, von denen im Folgenden einige genannt werden sollen.

4.2.1 Identität

Eine Anwendung, die bereits große praktische Bedeutung erlangt hat, ist die Identifikation von Personen anhand von Videodaten. Sie spielt vor allem im Sicherheitsbereich eine große Rolle. Hierbei besteht die Motivation darin, technische Lesegeräte für Chipkarten oder Code-Eingaben durch automatische Zugangskontrollen zu ergänzen oder zu ersetzen. Bei der Gesichtserkennung gibt es mehrere Spielarten. Ein 1-zu-1 Vergleich (Verifikation, Authentifizierung) findet dann statt, wenn ein Modell einer Person existiert, z.B. auf einem maschinenlesbaren Ausweis, und festgestellt werden soll, ob die Person, die sich mit der Karte ausweisen will, auch tatsächlich der Karteninhaber ist. Ebenfalls um einen 1-zu-1 Vergleich handelt es sich bei der Anwendung im Baumarkt, bei der der Roboter ein Modell seines aktuellen Kunden erstellt. Nach Verlust des Kontakts soll festgestellt werden, ob es sich bei einem neuen Interaktionspartner um diesen Kunden handelt oder nicht. In einem anderen Szenario existiert eine Datenbank von Modellen für Personen und es soll für eine Person entschieden werden, ob sie in der Datenbank gespeichert ist oder nicht. Hierbei handelt es sich um einen 1-zu-n Vergleich (Identifikation).

Die Gesichtserkennung unterscheidet sich von anderen Mustererkennungsaufgaben, bei denen in der Regel nur wenige Klassen unterschieden werden und für jede Klasse viele Beispiele existieren. In diesem Fall können unbekannte Beispiele klassifiziert werden, indem zwischen den Trainingsbeispielen interpoliert wird. Bei der Gesichtserkennung existieren dagegen sehr viele Klassen, aber in der Regel nur wenige Beispiele pro Klasse, so dass unbekannte Beispiele durch eine Extrapolation der Trainingsbeispiele klassifiziert werden müssen.

Frühe Arbeiten auf dem Gebiet der Gesichtserkennung verwendeten Verhältnisse der Abstände zwischen Feature-Punkten wie der Nasenspitze und den Eckpunkten von Augen und Mund [Kanade, 1977]. Eine Erweiterung dieses Ansatzes verwendet so genannte deformierbare Templates, die parametrierbare Modelle des Gesichtes und seiner Merkmale darstellen [Yuille, 1991]. Beim Elastic-Graph-Matching erfolgt die Merkmalsextraktion mit Hilfe von Gabor-Wavelets. Es wird ein deformierbares Modell auf ein Gesicht angepasst und entsprechend der extrahierten Filterantworten klassifiziert [Wiskott et al., 1997a]. Holistische Methoden betrachten nicht einzelne Gesichtsmerkmale, sondern versuchen, die Erscheinung von Gesichtern als ganzes zu beschreiben. Turk et al. verwenden hierfür einen Eigenface-Ansatz, der auch schon bei der Gesichtsdetektion erwähnt wurde [Turk and Pentland, 1991]. [Phillips, 1999] verwenden Support Vector Machines für die Identifikation und die Verifikation von Personen. Dazu werden zwei Klassen definiert, eine beschreibt die Unterschiede zwischen den Bildern jeweils der selben Person, die andere beschreibt die Unterschiede zwischen den Bildern verschiedener Personen. Bei der Erkennung entscheidet die SVM für zwei Bilder, ob diese zur selben Klasse gehören (intrapersonal) oder zu unterschiedlichen Klassen (extrapersonal). Für Training und Test wurden die Augenmittelpunkte manuell gelabelt. Liu et al. verwenden so genannte „Independent Gabor Features“ für die Gesichtserkennung [Liu and Wechsler, 2003]. Dabei handelt es sich um Gabor-Merkmalsvektoren, deren Dimension zunächst mittels einer PCA reduziert wird und deren Redundanz anschließend durch eine ICA minimiert wird.

4.2.2 Geschlecht

Für die Klassifikation des Geschlechts wird in [Golomb et al., 1991] ein vollverschaltetes zweischichtiges neuronales Netz verwendet. Die Bildgröße beträgt 30×30. In [Wiskott et al., 1995] und [Bendlin, 2004] wurde das Elastic-Graph-Matching für die Geschlechtsschätzung eingesetzt. In [Moghaddam and Yang, 2000] werden verschiedene Klassifikatoren wie Nearest Neighbor, Radial Basis Function Netzwerke und SVMs verglichen, um normalisierte Grauwertgesichter der Größe 80×40 bzw. 21×12 zu klassifizieren. Dabei wurden mit SVMs die niedrigsten Fehlerraten erreicht. [Shakhnarovich et al., 2002] verwenden AdaBoost für die Klassifikation des Geschlechts und der ethnischen Zugehörigkeit (asiatisch/nicht-asiatisch). Die Gesichtsausschnitte werden mit dem im Abschnitt 3.2.3.3 beschriebenen Verfahren detektiert und nicht weiter normalisiert.

4.2.3 Alter

Obwohl das Alter einer Person für die Gestaltung der Interaktion besonders wichtig zu sein scheint, steckt die Forschung auf dem Gebiet der automatischen Altersschätzung noch in den Kinderschuhen. Es existieren Ansätze für die Modellierung bzw. die Simulation von Alterungsprozessen [Burt and Perrett, 1995], aber nur eine bekannte Arbeit zur automatischen Altersschätzung [Lanitis et al., 2004], in der verschiedene Klassifikatoren für die automatische Altersschätzung verglichen werden. Für die Merkmalsextraktion werden Farb-Active-Appearance-Models verwendet, die auf 400 Farbbildern von 40 Personen trainiert wurden, wobei lediglich 22 Parameter für die Beschreibung eines Gesichtes verwendet wurden. Das abgedeckte Altersspektrum umfasste den Bereich von Neugeborenen bis zu einem Alter von 35 Jahren. Sehr gute Ergebnisse wurden mit Multilayer-Perceptrons erreicht.

4.2.4 Gesichtsausdruck

Ebenfalls von großem praktischen Interesse ist die automatische Extraktion von Gesichtsausdrücken aus Videodaten. Anwendungen finden sich z.B. in der Psychologie, wo Videosequenzen von Interviews hinsichtlich der gezeigten Gesichtsausdrücke ausgewertet werden müssen. Eine automatische Analyse könnte hier den sehr aufwendigen Prozess der manuellen Annotation ersetzen.

Ein Pionier in der Analyse des Zusammenhangs zwischen menschlicher Gemütsregung und dem Gesichtsausdruck ist Charles Robert Darwin (1809 - 1882). In seinem Buch *The Expression of the Emotion in Man and Animals* [Darwin, 1872] beschreibt er die Universalität von Gesichtsausdrücken als Resultat einer evolutionären Entwicklung bei Mensch und Tier. Damit stand er zu seiner Zeit in starkem Widerspruch zu anderen Wissenschaftlern und Gelehrten, die den menschlichen Gesichtsausdruck als von der umgebenden Kultur bestimmt ansahen. Die Deutungen eines Gesichtsausdrucks in verschiedenen Kulturen sollte sich nach deren Auffassung so sehr unterscheiden wie andere kulturelle Elemente wie Sprache, Haltungen oder Wertvorstellungen.

Auch in der Mensch-Maschine-Interaktion spielt die Mimikanalyse bereits eine Rolle. Hier be-

steht die Zielstellung darin, die Ergonomie von Mensch-Maschine-Schnittstellen zu verbessern, indem Stress oder Unkonzentriertheit des Nutzers erkannt und geeignete Maßnahmen ergriffen werden. Laut [Zimmermann et al., 2003] sollte eine anpassungsfähige Mensch-Maschine-Schnittstelle versuchen, den Nutzer in einer positiven Stimmung zu halten und negativen Stimmungen entgegenwirken. Dabei kommen in der Regel Sensoren zum Einsatz, die am Körper des Nutzers angebracht werden, die in den meisten Szenarien aber eine Belästigung für den Anwender darstellen und deshalb durch berührungslose Verfahren ersetzt werden sollten. Beispiele für die Anwendung einer Mimikanalyse auf mobilen Servicerobotern werden in [Bartlett et al., 2003] und [Shiomi et al., 2004] beschrieben.

Liefert das Mimikanalysesystem ein Ergebnis nach einem Kodiersystem wie dem Facial Action Coding System (FACS), siehe Abschnitt 4.3.2.1, kann dieses anstatt der reinen Bildinformation gespeichert oder übertragen und später wieder in ein ausdruckhaftes Gesicht synthetisiert werden. Anwendungen wie Videokonferenzen, Videokompressionsverfahren oder die Charakteranimation in der Filmindustrie können davon profitieren.

Die Verfahren können grob in zwei Klassen unterschieden werden, bei denen entweder die Deformation von Gesichtsstrukturen oder deren Bewegung analysiert wird. Ein Überblick über Methoden zur Analyse von Gesichtsausdrücken wird in [Fasel and Luettin, 2003] gegeben. Im Folgenden werden einige Beispiele für jede Klasse benannt.

4.2.4.1 Deformation

Padgett und Cotrell [Padgett and Cottrell, 1997] verglichen die Leistung von Mimikanalysesystemen, die als Merkmalsextraktion entweder eine globale PCA (Eigenfaces), eine lokale PCA im Bereich der Augen und des Mundes (Eigenfeatures) oder eine PCA für zufällig ausgewählte Bildausschnitte (Eigenvektoren) verwenden. Die Idee, die hinter der Verwendung ausgewählter Bildbereiche liegt, ist, dass strukturelle Unterschiede zwischen Individuen, die bei einer globalen Betrachtung des Gesichtes berücksichtigt werden, für eine Identifikation zwar wichtig, für eine Klassifikation von Gesichtsausdrücken aber eher hinderlich sind. Die Klassifikation erfolgte mit einem zweischichtigen Feed-Forward-Netz mit 10 Hidden-Neuronen. Dabei erreichen sie mit der PCA für zufällig ausgewählte Bildausschnitte eine Erkennungsrate von 86% auf unbekannten Gesichtern. [Lisetti and Rumelhart, 1998] nutzen das Grauwertbild ohne Merkmalsextraktion als Eingabe für ein dreischichtiges Feed-Forward Netz. Dabei fanden sie heraus, dass man eine bessere Erkennungsrate erhält, wenn man das Gesicht in drei Regionen aufteilt und diese einzeln verarbeitet. [Fellenz and Taylor, 1999] verwenden neben dem Grauwertbild und einer PCA-Repräsentation auch eine durch Gaborfilterung entstandene Repräsentation des Bildes. Dabei wurden 6 verschieden orientierte Gaborfilter verwendet und das manuell normierte Gesichtsbild mit diesen im Ortsbereich gefaltet. Dabei wurden nur die Klassen Freude, Wut, Trauer und Neutral betrachtet. Als Klassifikator kam ein Feed-Forward Netzwerk zum Einsatz. Die Erkennungsraten auf einem Testdatensatz lagen bei lediglich 60%, was die Autoren auf die geringe Auflösung von 35×37 Pixeln und auf die unzureichende Normalisierung der Gesichtsbilder zurückführen. [Lyons and Budynek, 1999] verwenden für die Merkmalsextraktion eine Variante

des Elastic-Graph-Matching, bei der ein regelmäßiges Gitter durch Anpassung seines Schwerpunktes und seiner Skalierung in x- und y-Richtung auf ein Gesicht angepasst wird. Die extrahierten Filterantworten werden durch eine PCA in ihrer Dimension reduziert und schließlich mittels einer Linear-Discriminant-Analysis hinsichtlich Gesichtsausdruck, Geschlecht und ethnischer Zugehörigkeit (asiatisch/andere) klassifiziert. Auf einer Datenbank mit 193 Bildern von neun japanischen Frauen wird eine Erkennungsrate von 75% erreicht. In [Dailey and Cottrell, 1999] wird eine Repräsentation von Gesichtern durch Gabor-Jets und eine Repräsentation durch Hauptkomponenten, die durch eine PCA auf zufällig gewählten Bildausschnitten durchgeführt wurde, gegenübergestellt. Als Klassifikatoren werden neuronale Netze verwendet. Die Autoren kommen zu dem Schluss, dass beide Repräsentationsformen im Kontext der Mimikanalyse nahezu gleichwertig sind. In [Dailey et al., 2002] wird ein neuronales Netzwerk für die Mimikanalyse vorgestellt. Als Eingabe dienen durch eine PCA komprimierte Gaborfilterantworten, die auf einem regelmäßigen Raster der Größe 29×35 berechnet wurden. Mit diesem Modell können die Autoren eine Reihe von psychologischen Phänomenen erklären, die mit der Wahrnehmung von Gesichtsausdrücken zusammenhängen. In [Lanitis et al., 1995] werden Active Appearance Modelle für die Analyse von Gesichtsausdrücken verwendet. Die Klassifikation erfolgt, indem für jede Basisemotion die Verteilung der Appearance-Parameter bestimmt wird und ein neues Beispiel anhand der Mahalanobis-Distanz zu einer Basisemotion zugeordnet wird. Es werden Erkennungsrate von 74% erreicht. [Littlewort et al., 2003] extrahieren Gabor-Merkmale aus Grauwertbildern der Größe 48×48 und verwenden AdaBoost für die Auswahl einer relativ kleinen Anzahl von Merkmalen, die für die Klassifikation von Gesichtsausdrücken ausreichend sind. In der Vorverarbeitung erfolgt neben einer Gesichtsdetektion keine weitere Normalisierung der Darstellung des Gesichtes. Die Erkennungsraten liegen über 90%.

4.2.4.2 Bewegung

[Mase and Pentland, 1991] verwenden den optischen Fluss, um die Aktivität von 12 der 44 Gesichtsmuskeln zu schätzen. Dazu wurde für jeden Muskel eine Region im Gesicht bestimmt und die Achse der möglichen Bewegungen dieses Muskels festgelegt. Der dichte optische Fluss wurde in acht Richtungen quantifiziert und als grobe Schätzung der Muskelaktivitäten verwendet. [Yacoob and Black, 1996] [Yacoob and Davis, 1996] stellten ein lokales parametrisches Modell zur exakten Schätzung sowohl der Bewegung von Gesichtsmerkmalen wie Mund, Nase, Augenlidern und Augenbrauen, als auch der Bewegung der Ränder dieser Merkmale vor. Diese Bewegungen werden mit einem kleinen Satz von Parametern dargestellt. Das Modell beachtet dabei nur dauerhafte Gesichtsmerkmale und lässt nicht dauerhafte Merkmale wie Furchen und Falten unberücksichtigt. [Lien et al., 1998] analysieren Action Units im oberen Gesichtsbereich. Dazu bedienten sie sich eines Trackings von Merkmalspunkten, eines durch eine PCA komprimierten dichten optischen Flusses und einer Faltenerkennung basierend auf der Detektion von Grauwertgradienten. Im Vorfeld der Flussschätzung wird die Ansicht des Gesichtes mittels einer affinen Transformation normalisiert. Die raum-zeitliche Klassifikation der Gesichtsausdrücke erfolgt mittels HMMs. Ein Nachteil des Verfahrens liegt in der Notwendigkeit einer sehr exakten Normalisierung der Gesichter im Vorfeld der Flussschätzung, für die drei Punkte im Gesicht

von Hand markiert werden müssen. [Otsuka and Ohya, 1998] schätzen den optischen Fluss im Bereich von Augen und Mund. Die berechneten Verschiebungsvektorfelder werden in x- und in y-Richtung Fourier-transformiert und ein Merkmalsvektor bestehend aus den Fourier-Koeffizienten weiterverarbeitet. Die für die Berechnung des optischen Flusses verwendeten Merkmalspunkte werden automatisch aufgrund von lokalen Extrema und Sattelpunkten der Helligkeitsverteilung des ersten Bildes einer Bildsequenz gewählt. [Yoneyama and Iwano, 1997] teilen das normalisierte Gesichtsbild in acht mal zehn Regionen ein und berechnen in diesen die Verschiebungsvektorfelder, um diese dann in drei Kategorien zu Quantisieren: Bewegung nach oben, Bewegung nach unten und keine Bewegung. Gegenüber Ansätzen mit optischem Fluss werden bei Differenzbildern die Differenzen in der Intensität einzelner Pixel in zwei Bildern, jedoch nicht die Richtung der Intensitätsänderung betrachtet. Zur Erkennung von Gesichtsausdrücken wird häufig von dem Bild einer Person dessen neutrale Referenz abgezogen. Dazu müssen beide Gesichtsbilder gut normiert sein. Ansätze zur Differenzbildanalyse sind z.B. in [Donato and Bartlett, 1999] zu finden. [DeCarlo and Metaxas, 1997] präsentieren eine formale Methodik zur Integration des optischen Flusses und verformbarer 3D Modelle, die sie zur Schätzung der Gesichtsform und des Gesichtsausdrucks verwenden.

Prinzipiel sind sowohl statische als auch dynamische Methoden für die Mimikanalyse geeigent. Ein Vorteil dynamischer Verfahren liegt darin, dass sie die zeitliche Entwicklung von Gesichtsausdrücken, die mitunter wichtig für die Unterscheidung von Mimiken ist, unmittelbar erfassen. Ein Nachteil ist dagegen, dass in der Regel neutrale Referenzbilder vom Anfang einer Mimiksequenz und eine sehr exakte normalisierte Darstellung der einzelnen Bilder benötigt werden. Im Abschnitt 4.3 erfolgen weitere Betrachtungen zur Eignung der verschiedenen Verfahren und Kodierungen von Gesichtsausdrücken.

4.3 Datenbasis

In diesem Abschnitt werden Problemstellungen erläutert, die sich bei der Erstellung einer Datenbasis für die Klassifikation von Geschlecht, Alter, Identität und Gesichtsausdruck ergeben. Dabei sind für die verschiedenen Erkennungsaufgaben unterschiedliche Kriterien einzuhalten. Für die Klassifikation von Alter und Geschlecht muss für eine ausgewogene Verteilung gesorgt werden, d.h. es sollte weder ein Geschlecht, noch eine bestimmte Altersgruppe überwiegen, da das Erkennungssystem keine a priori Annahmen über diese Kategorien treffen soll. Für die Personenidentifikation müssen ausreichend viele Bilder jeder Person vorhanden sein, so dass ein Trainings-, Test- und ein Validierungsdatensatz erstellt werden kann. Denkbar schwieriger gestaltet sich die Aufnahme von Gesichtsausdrücken. Die Fragestellungen betreffen hier die Auswahl der Kodierung der Gesichtsausdrücke und die Gewinnung des Bildmaterials. In den folgenden Abschnitten werden diesbezügliche Vorüberlegungen zusammengetragen.

4.3.1 Basisemotionen und deren Bewertung

Laut [Zimmermann et al., 2003] handelt es sich bei Emotionen um Reaktionen auf einen Reiz oder auf einen Gedanken. Es sind bewusst wahrgenommene Ereignisse mit einer Dauer von Sekunden bis Minuten. Um zu ermitteln, wie viele unterscheidbare grundlegende Emotionen das Gesicht übermitteln kann, wurden Fotografien verschiedener Gesichtsausdrücke ausgewertet. Die Betrachter sollten jedem Foto ein Emotionswort zuordnen, welches dabei aus einer Reihe von Wörtern ausgewählt werden konnte. Anschließend wurde die Übereinstimmung in der Zuordnung von Emotion und Gesichtsausdruck analysiert und Kenntnisse über die durch das Gesicht vermittelbaren Emotionen abgeleitet. In [Ekman and Friesen, 1975b] sind sechs Emotionen beschrieben, die in der Psychologie am etabliertesten sind. Dabei handelt es sich um Überraschung, Trauer, Wut, Angst, Ekel und Freude. Studien belegen, dass diese Gesichtsausdrücke von Menschen universell ausgedrückt und wahrgenommen werden [Ekman, 1989], d.h. unabhängig von der ethnischen Zugehörigkeit oder dem Alter. Neben dem Gesichtsausdruck liefern physiologische Signale wie Hautleitwert, Herzfrequenz, Blutdruck, Atmung, Pupillendilatation, Elektroencephalogramm oder Aktionspotentiale von Muskeln Informationen über Stimmungen und Gefühle einer Person.

Das Gesicht kann laut Ekman [Ekman and Friesen, 1975b] in drei Regionen aufgeteilt werden. Diese Regionen sind mehr oder weniger unabhängig in ihrer Erscheinung. Es sind die Stirn/Brauenregion, die Augen/Nasenregion und die Mundregion. Sie können, müssen sich aber nicht bei einem Gesichtsausdruck vom neutralen Bild unterscheiden. Dabei hat die Darstellung einer Basisemotion z.B. in zwei Regionen in Verbindung mit dem Neutralbleiben der Dritten eine ganz bestimmte Bedeutung. So kann Überraschung durch zwei Regionen angezeigt werden, ohne dass die Dritte mit einbezogen wird. Diese Zwei-Regionen-Überraschung kann verschiedene Bedeutungen haben [Ekman and Friesen, 1975b]. Die erste Möglichkeit ist die fragende Überraschung. Hier verbleibt die Mundregion neutral, während die beiden anderen Regionen Überraschung anzeigen. Eine zweite Form der Überraschung ist die erstaunte Überraschung,

hier verbleibt die Stirn/Brauenregion neutral. Die dritte Form der Überraschung ist die betäubte oder wenig interessierte Überraschung, wenn eine Person zum Beispiel erschöpft ist oder unter Medikamenten steht. Hier bleibt die Augen-/Nasenregion neutral. Die vierte Form ist schließlich der volle Ausdruck von Überraschung in allen 3 Regionen.

4.3.2 Kodierung von Gesichtsausdrücken

Ein Gesichtsausdruck entsteht durch Kontraktion/Relaxation der Gesichtsmuskeln, was zu einer Formänderung von Gesichtsmerkmalen wie Augenlidern, Augenbrauen, Nase, Lippen und Mund führt. Typischerweise sind die Muskelaktivitäten kurz, die meisten haben eine Dauer von 250 Millisekunden bis zu 5 Sekunden. Bei der Beschreibung von Gesichtsausdrücken spielen der Ort des Auftretens einer Bewegung und deren Intensität und Dynamik eine Rolle. Die Intensität eines Gesichtsausdrucks kann am Verformungsgrad der betrachteten Gesichtsmerkmale oder an der Dichte der Falten in bestimmten Gesichtsregionen gemessen werden. So kann die Intensität eines Ausdruckes von Ekel daran abgelesen werden, wie stark die Nase gerümpft ist, und wie dicht Falten an den Nasenseiten auftreten. Jedoch variiert die Intensität eines Gesichtsmerkmales von Person zu Person. Daher ist es schwer, eine absolute Intensität eines Gesichtsausdrucks anzugeben. Da oft mit der Erkennung von Gesichtsausdrücken auch deren Interpretation bezüglich des emotionalen Zustandes verbunden ist, lässt sich somit auch nur schwer eine Aussage über den Grad der Emotion der betreffenden Person machen. Auch gibt es eine Schwierigkeit bei der Erkennung schneller, spontaner Ausdrücke, die in ihrer Intensität weit weniger stark sind als dargestellte Gesichtsausdrücke, die meist sehr intensiv ausgeprägt sind. Neben der reinen Deformation von Gesichtsmerkmalen trägt auch die zeitliche Entwicklung von Mimiken Informationen zu deren Unterscheidung. Statische Bilder können nichts über das zeitliche Verhalten von Gesichtsausdrücken aussagen. Daher wird es meist als wichtig angesehen, neben der Deformation auch die Dynamik eines Gesichtsausdrucks zu messen. Der Verlauf eines Gesichtsausdrucks wird laut [Fasel and Luettin, 2003] in drei zeitliche Phasen eingeteilt: *Aufsetzen* (onset), *Halten* (apex) und *Absetzen* (offset).

Trotzdem enthält die in dieser Arbeit verwendete Datenbank aus drei Gründen nur statische Gesichtsausdrücke. Zum Ersten sollen für die Mimikanalyse die selben Methoden der Merkmalsextraktion eingesetzt werden wie für die Analyse von Geschlecht, Alter und Identität, so dass im Gesamtsystem mit einer Merkmalsextraktion sämtliche Klassifikationsaufgaben gelöst werden können. Zum Zweiten benötigen die Verfahren zur Extraktion dynamischer Merkmale, wie z.B. der optische Fluss, in der Regel Referenzbilder vom Anfang einer Sequenz, z.B. ein neutrales Bild wie in der Cohn-Kanade-Datenbank. Ein solches Bild ist in einer realen Anwendung in der Regel nicht verfügbar. Zum Dritten können auch mit einer Extraktion statischer Merkmale dynamische Aspekte von Gesichtsausdrücken erfasst werden, indem diese hinreichend schnell ermittelt und dynamisch klassifiziert werden.

Es existieren zwei Ansätze zur Kodierung von Gesichtsausdrücken:

Visuelle Klassen: Bewegungen oder Verformungen im Gesicht können in verschiedene visuelle Klassen eingeteilt werden. Hier wird sozusagen ein Katalog aller möglichen Veränderungen im Gesicht erstellt, ohne eine Interpretation bezüglich des mentalen Zustandes der Person vorzunehmen.

Emotionale Klassen: Dieser Ansatz orientiert sich daran, welche Nachricht durch einen Gesichtsausdruck übermittelt werden soll. Dabei wird das Vorhandensein einer bestimmten Zahl von Basisemotionen oder mentaler Prototypen angenommen, die über das Gesicht übermittelt werden können. Dabei handelt es sich in der Regel um die bereits erwähnten Basisemotionen Überraschung, Trauer, Wut, Angst, Ekel und Freude.

4.3.2.1 Visuelle Klassen

Der Ansatz, Gesichtsbewegungen und -deformationen in eine Menge visueller Klassen einzuteilen, wird häufig auch als zeichen-basierter Ansatz bezeichnet. Das Ziel ist, Veränderungen im Gesicht in ihrem Ort des Auftretens und ihrer Intensität zu beschreiben. Das bekannteste Verfahren, Gesichtsausdrücke auf diese Art und Weise zu kodieren, ist das Facial Action Coding System (FACS). Es wurde durch P. Ekman und W.V. Friesen Ende der 70er Jahre entwickelt [Ekman and Friesen, 1978]. Das FACS ist ein Wertungssystem, welches für menschliche Beobachter entwickelt wurde und hauptsächlich in der Verhaltensforschung zum Einsatz kommt. Das FACS basiert auf der Erscheinung des Gesichts und enthält keine Interpretation des Gesichtsausdrucks bezüglich der mentalen Zustände der Person. Es verwendet 46 so genannte Action Units (AU), um Deformationen im Gesicht zu beschreiben, sowohl nach ihrem Ort als auch nach der Intensität. Zusätzlich zu den 46 AUs definieren 20 weitere Einheiten Kopf- und Augenbewegungen. Von den Erfindern des *Facial Action Coding System* wurde auch eine Datenbank angelegt, die die Bewertung eines Gesichtsausdrucks mit Hilfe von *Action Units* in eine Affektbedeutung übersetzt, die so genannte FACSAID (Facial Action Coding System Affect Interpretation Database). Abbildung 4.2 zeigt eine Auswahl einiger Action Units und deren zugeordnete Basisemotionen.

Basisemotion	**Action Units (AU)**
Überraschung	1+2+5+26
Trauer	1+4+15
Wut	4+5+7+10+25,26
Angst	1+2+4+5+7+20+25,26
Ekel	4+9+17
Freude	6+12 (+26)

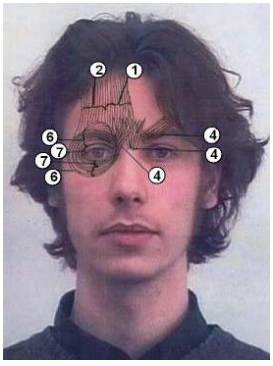


Abbildung 4.2: *Basisemotionen und zugeordnete Action Units, entnommen aus [Kobayashi and Hara, 1997]. Am Beispiel von Überraschung haben die AUs folgende Bedeutung. 1: Hoch heben der inneren Augenbraue (inner brow raiser) 2: Hoch heben der äußeren Augenbraue (outer brow raiser) 5: Heben der Oberlider (upper lid raiser) und 26: Unterkiefer fallen lassen (jaw drop). Das Bild zeigt die Lage einiger dieser AUs.*

Vorteile

detaillierte Repräsentation: Ein Kodiersystem wie das Facial Action Coding System bietet aufgrund seiner vielen Freiheitsgrade (ca. 7000 mögliche Kombinationen der 46 Action Units) eine detaillierte und akkurate Repräsentationsform für die Kodierung von Gesichtsausdrücken. Nur wenige Informationen über den eigentlichen Gesichtsausdruck gehen verloren (z.B. Informationen zum zeitlichen Verhalten).

leichte Synthetisierbarkeit: Die Kodierung in einem System wie dem FACS ist direkt mit der Darstellung von Gesichtsausdrücken verbunden. Die AUs beschreiben, wie die verschiedenen Merkmale eines Gesichts bewegt oder verformt werden müssen, um einen Gesichtsausdruck zu synthetisieren.

Objektivität: Da die einzelnen AUs klar umrissene Veränderungen im Gesicht darstellen, ist es leicht, ein Klassifikationsergebnis zu überprüfen. Man muss hier nichts über den Gemütszustand der Person wissen, da das Kodiersystem lediglich das Offensichtliche in Klassen einteilt und keine Interpretation durchgeführt wird.

Nachteile

große Klassenanzahl: Ein Problem bei umfangreichen Kodiersystemen wie dem FACS ist die große Anzahl von Klassen, die sich aus den vielen möglichen Kombinationen von AUs ergibt. Viele Ansätze arbeiten mit einer Untermenge von AUs, um die Klassenanzahl zu begrenzen.

keine Interpretation: Verfolgt man die Absicht, den Gemütszustand einer Person zu schätzen, muss nach der Klassifikation in visuelle Klassen mit einem Satz von Regeln der emotionale Zustand der Person abgeleitet werden.

4.3.2.2 Emotionale Klassen

Bei diesem Ansatz wird versucht, einem Gesichtsausdruck unmittelbar eine Interpretation hinsichtlich eines emotionalen Zustandes zuzuordnen.

Vorteile

kleine Klassenzahl: Die von Psychologen [Ekman and Friesen, 1975b] gefundenen Emotionen, die durch das Gesicht dargestellt werden können, bilden einen kompakten Satz an Klassen. Dabei handelt es sich um die sechs Basisemotionen Überraschung, Trauer, Wut, Angst, Ekel und Freude und um den neutralen Gesichtsausdruck.

Interpretation enthalten: Im Gegensatz zur Verwendung von visuellen Klassen trifft ein Klassifikator, der bzgl. emotionaler Klassen klassifiziert, bereits eine Annahme über den aktuellen Gemütszustand einer Person.

Nachteile

starker Informationsverlust: Unterteilt man das Gesehene bereits sofort in emotionale Prototypen, verliert man nahezu alle Information über die Erscheinung des Gesichtsausdrucks, er wird nur noch durch sehr abstrakte Emotionswörter repräsentiert.

erschwerte Synthetisierbarkeit: Einen Gesichtsausdruck lediglich aufgrund des Wissens um den emotionalen Zustand einer Person zu synthetisieren, kann sich als recht schwierig herausstellen. Man könnte lediglich einzelne typische Gesichtsausdrücke reproduzieren, jedoch keinerlei Nuancen in der Erscheinung einzelner Merkmale wiedergeben.

schlechte Überprüfbarkeit: Wird ein emotionaler Zustand einer Person geschätzt, so muss er zur Überprüfung der Leistung des Analysesystems auch mit dem tatsächlichen emotionalen Zustand der Person verglichen werden können. Die Gesichtsdaten müssen sozusagen emotional attributiert sein, d.h. es muss Zusatzwissen vorhanden sein, welches nicht direkt aus der Erscheinung des Gesichts abgeleitet werden kann.

4.3.2.3 Fazit

Die visuellen Klassen stellen sicherlich die sinnvollere Kodierung dar, da hier eine objektive Zuordnung von Gesichtsausdrücken auf FACS-Codes erfolgt und durch entsprechende Regelwerke eine ebenso objektive Interpretation hinsichtlich emotionaler Klassen möglich ist. Das Problem ist jedoch, dass für die FACS-Kodierung von Videodaten nicht nur eigens geschultes und zertifiziertes Personal nötig ist, sondern diese Kodierung auch sehr zeitaufwendig ist, weshalb in dieser Arbeit auf die Klassifikation nach emotionalen Klassen zurückgegriffen werden musste.

4.3.3 Aufnahme von Gesichtsausdrücken

Bei der Erstellung einer Datenbank für Gesichtsausdrücke stellt sich die Frage, wie diese glaubhaft erzeugt werden können. Eine korrekte Zuordnung zwischen Emotionen und Gesichtsausdrücken kann eigentlich nur dann erreicht werden, wenn ein Bild der Person in der entsprechenden Gemütsverfassung aufgenommen wird. Dabei entsteht das Problem, bei der Aufnahme der Daten, also unter Laborbedingungen, die gewünschten Emotionen zu provozieren. Die nahe liegendste und am häufigsten angewandte Methode besteht in der Präsentation von Videos mit entsprechendem Inhalt. In [Gross and Levenson, 1995] wurden 494 Personen 78 Videoclips gezeigt und die zwei effektivsten Videoclips für jede Emotion wurden ausgewählt. Bei den betrachteten Emotionen handelte es sich um Neutral, Überraschung, Trauer, Wut, Angst, Ekel, Vergnügen und Zufriedenheit. Die Auswertung wurde dabei aufgrund von Hautleitwert, Puls und Atemfrequenz durchgeführt. Mit diesen 16 Filmen konnten die gewünschten Emotionen zwar erfolgreich provoziert werden. Die Autoren stellten allerdings fest, dass bestimmte Emotionen in der Regel als Mischformen auftreten. Dies gilt vor allem für Wut, die in der Regel zusammen mit Ekel auftritt oder Angst, die zusammen mit Anspannung und Interesse beobachtet wurde. Am erfolgreichsten waren die Videos, mit denen Freude, Ekel und Trauer hervorgerufen wurde.

Eine andere Möglichkeit für die Erzeugung von Trainingsdaten besteht darin, dass die Probanden die gewünschten Gesichtsausdrücke darstellen, ohne dass dabei die entsprechenden Gefühle zu Grunde liegen. Das Problem hierbei ist, dass Menschen häufig nicht in der Lage sind, bestimmte Gesichtsausdrücke glaubhaft darzustellen. Oft können gestellte von realen Gesichtsausdrücken dadurch unterschieden werden, dass bestimmte Muskeln, die bei einem natürlichen Gesichtsausdruck beteiligt werden, beim Posieren inaktiv bleiben. Wie mit diesem Problem umgegangen werden kann, wird im Abschnitt 4.3.5 beschrieben.

4.3.4 Existierende Datenbanken

Die Extraktion von Informationen wie Geschlecht, Alter, Mimik und Identität aus Bilddaten stellt eine extrem komplexe Aufgabe dar, die durch auftretende Bildvariationen wie verschiedene Posen, Beleuchtungen, oder Gesichtsmerkmale wie Bärte oder Brillen erschwert wird. Diese Faktoren sind beim Aufbau einer Datenbank zu berücksichtigen. Aus diesem Grund sollen zunächst frei verfügbare Datenbanken auf ihre Verwendbarkeit für die hier gestellten Aufgaben untersucht werden.

4.3.4.1 Pictures of Facial Affect (POFA)

Die POFA-Datenbank [Ekman and Friesen, 1975a] enthält 110 Fotos von 14 Personen, die die sechs Basisemotionen darstellen. Die Bilder wurden von unbefangenen Beobachtern klassifiziert. Nur bei 70% Übereinstimmung der Klassifikation wurde ein Bild in die Datenbank aufgenommen. Für die Anwendung in dieser Arbeit, besonders im Hinblick auf die Klassifikation von Alter und Geschlecht, wird der Umfang der POFA-Datenbank als zu klein eingeschätzt.

4.3.4.2 Cohn-Kanade-Datenbank

Diese Datenbank enthält eine Reihe von Sequenzen verschiedener Gesichtsausdrücke von 100 Personen im Alter zwischen 18 und 30 Jahren [Kanade et al., 2000]. 65% der Personen sind weiblich. Jede Sequenz zeigt eine Entwicklung des Gesichtsausdrucks vom neutralen Gesicht zur vollen Ausprägung des Gesichtsausdrucks, also dem Aufsetzen (Onset). Die Daten sind nach dem FACS kodiert, wobei einzelne Sequenzen emotionalen Klassen entsprechen. Dabei werden aber einige der Basisemotionen nicht unbedingt realistisch dargestellt, siehe Abbildung 4.3. Weitere Mängel sind eine häufige Überbelichtung der Bilder und der teilweise ins Gesicht hineinreichende Zeitstempel. Der abgedeckte Altersbereich ist für eine Altersschätzung nicht ausreichend.

4.3.5 NIFace2

Aufgrund der Defizite der eben beschriebenen Datenbanken war es notwendig, für die Erstellung und den Test der Analysesysteme eine eigene Datenbank anzulegen. Diese enthält Bilder von 136 Personen, wobei jede Person unter verschiedenen Beleuchtungsbedingungen und mit verschie-

Abbildung 4.3: *Beispielsequenz für die Basisemotion Freude aus der Cohn-Kanade-Datenbank. Es fällt auf, dass die Augenpartie nahezu unverändert bleibt, was typisch für dargestellte Freude ist. Bei einem realen Gesichtsausdruck werden die Augen leicht zusammengekniffen.*

denen Blickrichtungen aufgenommen wurde. Zu jedem Bild sind Identität, Alter, Geschlecht, Blickrichtung und dargestellter Gesichtsausdruck in einer XML-Datei hinterlegt.

Datensatz für Geschlecht, Alter und Identität Es wurde darauf geachtet, dass der Datensatz hinsichtlich der Kategorien Alter und Geschlecht gleichverteilt ist. Dazu wurden 5 jeweils 10 Jahre abdeckende Altersgruppen von 10 bis 60 Jahren gebildet. Insgesamt wurden in diesem Datensatz 70 Personen verwendet, so dass jede Altersgruppe 7 Männer und 7 Frauen enthält. Es wurden Gesichter mit frontaler und in horizontaler Richtung um 5° gedrehter Ansicht, neutralem Gesichtsausdruck und verschiedenen Beleuchtungen verwendet. Damit umfasst die Datenbank für die Klassifikation von Alter, Geschlecht und Identität 490 Bilder.

Datensatz für Mimik Da keine FACS-Kodierung der Bilder möglich war, bestand die einzig praktikable Lösung darin, die sieben Basisemotionen von den Probanden darstellen zu lassen. Um Bilder mit unglaubhaft dargestellten Gesichtsausdrücken auszuschließen, erfolgte nach der Aufnahme eine Zuordnung zu einer der Basisemotionen durch 10 Befundungspersonen. Falls ein Gesichtsausdruck überhaupt nicht eingeordnet werden konnte, oder die Aufnahme von schlechter Qualität war, konnte die Klasse „unbekannt“ ausgewählt werden. Nur wenn eine Übereinstimmung von mindestens sieben Personen zu verzeichnen war, wurde das betreffende Bild in den Mimikdatensatz aufgenommen. Tabelle 4.1 zeigt einige Beispiele klassifizierter Gesichtsausdrücke. Dieses Auswahlkriterium wurde auf die gesamte NIFace2-Datenbank angewendet. Abbildung 4.4 zeigt die Übereinstimmung der menschlichen Klassifizierer. Es ist ersichtlich, dass besonders die Gesichtsausdrücke Angst, Trauer und Ekel durch die Probanden nur schlecht dargestellt bzw. durch die Klassifizierer nur schlecht erkannt werden konnten. Nach der eben beschriebenen Vorklassifikation wurden aus der gesamten Datenbank 30 Personen ausgewählt, die alle 7 Gesichtsausdrücke glaubhaft wiedergeben konnten, wobei auch hier eine Gleichverteilung über die Altersgruppen und das Geschlecht eingehalten wurde. Damit umfasst die Datenbank für die Klassifikation von Gesichtsausdrücken 210 Bilder.

Markierung der Gesichtsmerkmale Die für die Gesichtsanalyse eingesetzten Verfahren benötigen für die Modellerstellung die Positionen markanter Merkmale im Gesicht. Die Gesamtzahl von 120 Labelpunkten ergibt sich dabei aus der Vereinigung der 38 Labelpunkte für das Elastic-Graph-Matching (Abschnitt 4.4), der 2 Labelpunkte für die Indpendent-Component-

			
Neutral	0%	0%	0%
Überraschung	0%	0%	10%
Trauer	0%	30%	0%
Wut	0%	70%	0%
Angst	0%	0%	30%
Ekel	0%	0%	50%
Freude	100%	0%	0%
unbekannt	0%	0%	10%

Tabelle 4.1: *Beispiele für die manuelle Klassifikation von drei Bildern aus der NIFace2-Datenbank. Jedes Bild wurde von 10 Personen klassifiziert und entsprechend der prozentualen Angaben den angegebenen Mimikklassen zugeordnet. Nach dem Auswahlkriterium von 70% konnten die ersten beiden Bilder in den Mimikdatensatz übernommen werden, während das letzte verworfen wurde.*

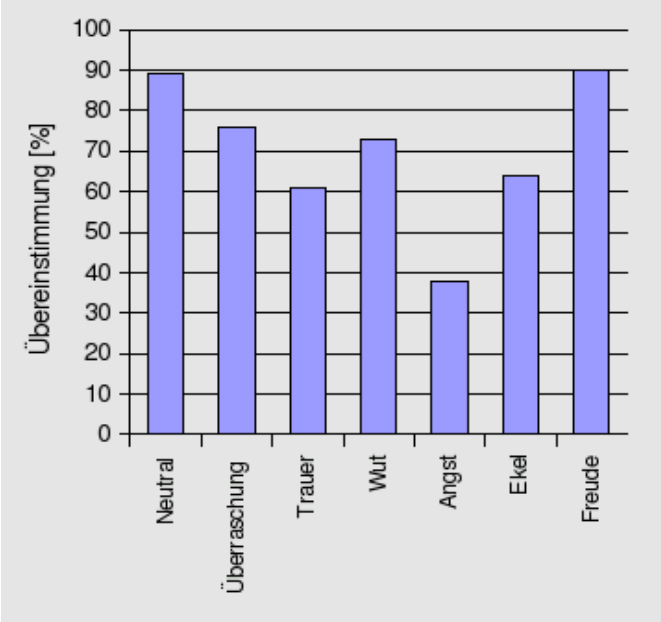
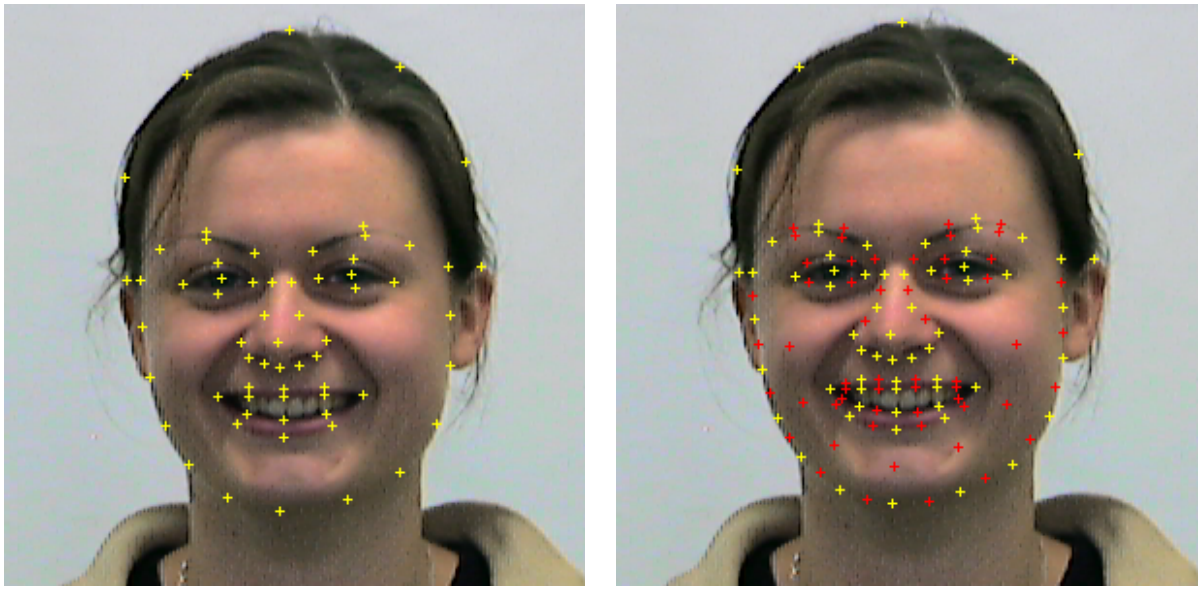


Abbildung 4.4: *Übereinstimmung der manuell klassifizierten Gesichtsausdrücke mit den Klassen, unter denen die Bilder aufgenommen wurden. Besonders gut dargestellt bzw. erkannt wurden die Mimiken Neutral und Freude, besonders schwierig waren Angst, Trauer und Ekel.*

Analysis (Abschnitt 4.5) und der 107 Labelpunkte für die Active-Appearance-Models (Abschnitt 4.6). Um den Aufwand für das Labeln zu minimieren, wurden nur 64 Labelpunkte von Hand gesetzt und die restlichen durch geometrische Beziehungen zwischen diesen automatisch hinzugefügt. Abbildung 4.5 zeigt die manuell und die automatisch gesetzten Labelpunkte.



(a) (b)

Abbildung 4.5: *Darstellung der Label-Punkte. (a) Alle manuell gesetzten und (b) zusätzlich automatisch aufgefüllten Labelpunkte.*

In den folgenden Abschnitten werden die drei für die Nutzeranalyse untersuchten Verfahren vorgestellt. Dabei wird nicht nur die Theorie der Verfahren eingeführt, sondern am Ende jedes Abschnitts werden bereits Untersuchungen zur Leistungsfähigkeit der Verfahren vorgestellt, bevor schließlich in Abschnitt 4.7 eine direkte Gegenüberstellung der drei Verfahren erfolgt.

4.4 Elastic-Graph-Matching

Beim Elastic-Graph-Matching handelt es sich um ein graphenbasiertes Verfahren, das an der Ruhr-Universität Bochum entwickelt wurde [Wiskott et al., 1997a].

4.4.1 Literatur

Die Arbeiten zum Elastic-Graph-Matching basieren auf der stark biologisch motivierten Dynamic-Link-Architecture (DLA) [Lades et al., 1993]. Ein Bild wird in dieser Architektur durch einen attributierten regelmäßigen Graphen repräsentiert, in dem sich Neuronen aufgrund ähnlicher Aktivitäten zu Teilgraphen gruppieren und so eine Objektsegmentierung erfolgt. Für die Objekterkennung wird ein solcher Teilgraph mit einer Sammlung von Modellen verglichen, bei denen es sich um Kopien solcher aus Bildern extrahierter Teilgraphen handelt. Bei diesem Vergleich werden lokale Deformationen der Modellgraphen zugelassen, so dass eine optimale Korrespondenz zwischen Knoten im Modell- und im Bildgraphen erreicht wird. Die einzelnen Knotenpunkte wurden für diese lokale Anpassung zufällig verschoben.

In [Lades et al., 1993] wurde mit einem regelmäßigen rechteckigen Graphen gearbeitet, der neben dem Objekt auch immer einen Teil des Hintergrunds abdeckte. Da beim Elastic-Graph-Matching aber anders als in der Dynamic-Link-Architecture keine Objektsegmentierung erfolgt, waren die Objekthintergründe stets einfarbig. In späteren Arbeiten wurde mit einem kleineren regelmäßigen Graphen gearbeitet, der sich nur innerhalb des Gesichtes befand bzw. mit unregelmäßigen, an des Objekt angepassten Graphen [Wiskott et al., 1995].

Die ersten Arbeiten zum Elastic-Graph-Matching [Wiskott et al., 1997a] waren im Umfeld der Personenidentifikation angesiedelt. Obwohl dies bis heute das Hauptanwendungsfeld ist, wurden im Laufe der Zeit auch Anwendungen für die Geschlechtsschätzung [Wiskott et al., 1995] [Shligerskiy, 2002] [Bendlin, 2004] und für die Mimikanalyse [Hong et al., 1998] realisiert.

4.4.2 Modellerstellung

Für die Erstellung des Modells werden Beispielbilder mit Gabor-Wavelets unterschiedlicher Frequenz und Orientierung gefaltet und anschließend an markanten Punkten die Filterantworten zu Jets zusammengefasst. Mehrere Jets zusammen bilden einen *Face-Graph*, der eine Beschreibung für jeweils ein Gesicht darstellt. Eine genügend große und vielfältige Auswahl an so erstellten Graphen bildet die so genannte *General-Face-Knowledge*. Der *Average-Graph*, eine Art Durchschnittsgesicht, wird erstellt, indem über alle Beispielgraphen in der General-Face-Knowledge gemittelt wird, sowohl über die Positionen der Knotenpunkte, als auch über die Jets selbst. Abbildung 4.6 zeigt den Ablauf bei der Modellerstellung im Überblick. Die einzelnen Schritte werden im Folgenden beschrieben.

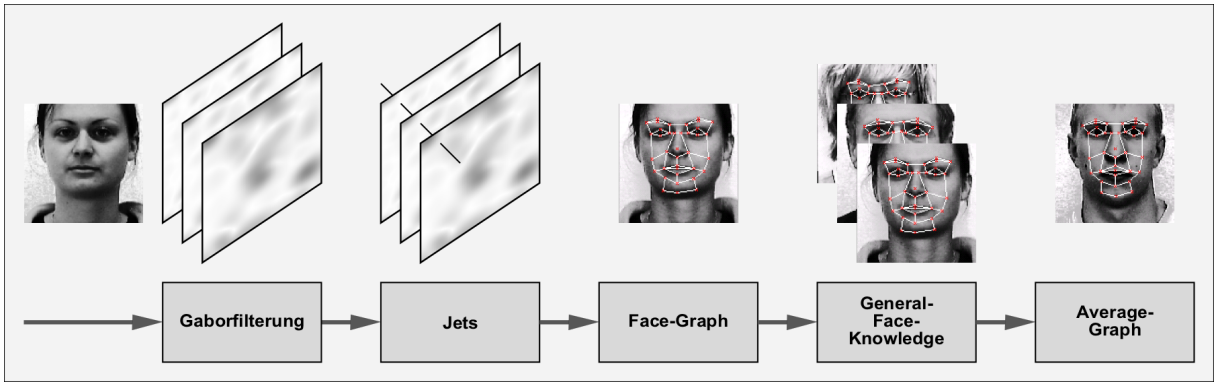


Abbildung 4.6: *Jedes Eingangsbild wird zunächst einer Faltung mit Gabor-Wavelets unterzogen. Die Filterantworten an gelabelten Punkten werden zu Jets zusammengefasst, aus denen dann ein Face-Graph erstellt wird. Die Face-Graphen mehrerer Eingangsbilder bilden die General-Face-Knowledge, aus der schließlich der Average-Graph erstellt wird.*

4.4.2.1 Gaborfilterung

Gabor-Wavelets stellen biologisch motivierte Faltungskerne dar, die ein ähnliches Verhalten zeigen wie die Simple-Cells im visuellen Cortex [Daugman, 1985]. Hierbei können zwei Typen unterschieden werden, die entweder auf Signale mit gerader oder mit ungerader Symmetrie ansprechen. Dieses Verhalten wird durch die Gabor-Wavelets mit einer komplexen Exponentialfunktion, bestehend aus einer Sinus- und einer Kosinuswelle, nachgebildet. Durch eine multiplikative Überlagerung dieser Exponentialfunktion mit einer Gaußfunktion erfolgt eine räumliche Begrenzung der Filterkerne.

Das zweidimensionale Gabor-Wavelet $\psi(\mathbf{x})$ (Mutter-Wavelet) kann folgendermaßen beschrieben werden:

$$\Re\{\psi(\mathbf{x})\} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \cdot \cos(2\pi f(x\cos\theta + y\sin\theta)) \tag{4.1}$$

$$\Im\{\psi(\mathbf{x})\} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \cdot \sin(2\pi f(x\cos\theta + y\sin\theta)) \tag{4.2}$$

oder in komplexer Schreibweise:

$$\begin{aligned} \psi(\mathbf{x}) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{\mathbf{x}^2}{2\sigma^2}\right)} \cdot e^{\imath 2\pi \mathbf{f}^T\mathbf{x}} \\ \mathbf{f}^T\mathbf{x} &= f(x\cos\theta + y\sin\theta) \end{aligned} \tag{4.3}$$

Hierbei sind σ die Standardabweichung der Gaußfunktion (Fensterbreite des Filterkerns), f die Ortsfrequenz der Oberflächenwelle und θ die Orientierung (Wellennormale) des Wavelets. Transformiert man das Gabor-Wavelet aus dem Ortsbereich in den Frequenzbereich, lässt sich dessen richtungsselektive Bandpasswirkung erkennen, siehe Abbildung 4.7.

Um unterschiedlich breite und orientierte Kanten detektieren zu können, werden aus dem Mutter-Wavelet, Gleichung 4.1 und 4.2, mehrere Filterkerne mit unterschiedlicher Orientierung

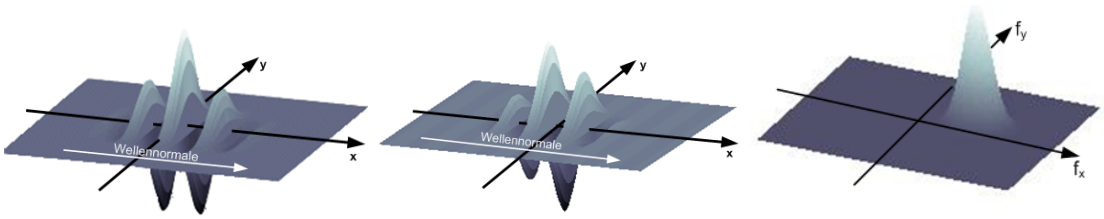


Abbildung 4.7: *Zweidimensionales Gabor-Wavelet. (a) Realteil im Ortsbereich (b) Imaginärteil im Ortsbereich (c) Gabor-Wavelet im Ortsfrequenzbereich.*

und Frequenz erzeugt. Hierbei muss gewährleistet werden, dass unabhängig von der gewählten Frequenz immer gleich viele exzitatorische und inhibitorische Bereiche im Wavelet existieren. Dies kann mittels folgender Gleichung erreicht werden, indem σ in Abhängigkeit der Wellenlänge berechnet wird:

$$\sigma(\lambda) \quad = \quad \frac{\lambda}{\pi}\sqrt{\frac{ln2}{2}} \cdot \frac{2^b+1}{2^b-1}. \tag{4.4}$$

Dabei ist b die Bandbreite des Filters (in Oktaven), die bestimmt, welcher Frequenzbereich von der Funktion abgedeckt wird. $\lambda = \frac{1}{f}$ ist die Wellenlänge der Grundschwingung. Somit können mit folgenden Parametern alle gewünschten Gabor-Wavelets erzeugt werden:

b ... Bandbreite der Einhüllenden
λ ... Wellenlänge der Grundschwingung
θ ... Orientierung des Wavelets.

Die Faltung eines Eingangsbildes $\mathbf{I}(\mathbf{x})$ mit einem Gabor-Wavelet $\psi_n(\mathbf{x})$ im Ortsraum bzw. deren Multiplikation im Frequenzraum ergibt eine komplexe Filterantwort:

$$\mathbf{J}_n(\mathbf{x}) \quad = \quad \mathbf{I}(\mathbf{x}) * \psi_n(\mathbf{x}). \tag{4.5}$$

Die Filterantwort des Gabor-Wavelets n an einem Bildpunkt kann wie folgt durch ihren Betrag und ihre Phase beschrieben werden:

$$j_n \quad = \quad a_n e^{\imath\phi_n}. \tag{4.6}$$

4.4.2.2 Jets

Ein Jet ist die Zusammenfassung der Antworten aller Gabor-Wavelets an einem Bildpunkt zu einem Vektor:

$$\mathbf{j} \quad = \quad \{j_0, ..., j_N\} \tag{4.7}$$

Abbildung 4.8: *Trainingsbild mit eingezeichnetem Face-Graph. Die Positionen der Knotenpunkte wurden so definiert, dass sie im Wesentlichen auf markanten Merkmalen liegen, zum Teil aber auch auf homogenen Flächen. Im Gegensatz zu früheren Arbeiten wurden keine Knotenpunkte auf der Außenkontur verwendet, da diese bei der Anpassung des Graphen aufgrund unterschiedlicher Hintergründe oft nicht gut platziert werden konnten.*

Dabei ist N die Anzahl an Filtern und j_i steht für die Antwort eines bestimmten Gabor-Wavelets. Ein Jet beschreibt somit die Frequenzen und Orientierungen in einer lokalen Umgebung eines Bildpunktes. Bei der Konstruktion der Gabor-Wavelets muss sichergestellt werden, dass der gewünschte Frequenz- und Orientierungsbereich gleichmäßig abgedeckt wird. Dazu werden 8 Orientierungen und 4 Frequenzen verwendet. Wie die Parameter der Gabor-Wavelets ermittelt wurden, wird in Anhang A.5 beschrieben. Die Faltung der 32 Gabor-Wavelets mit einem Eingangsbild ergibt 32 komplexe Filterantworten. Ein Jet wird gebildet, indem diese Filterantworten an jeweils einer Bildposition zusammengefasst werden. Im nächsten Schritt werden die Jets zu einem Face-Graphen zusammengefasst.

4.4.2.3 Face-Graph

Für die Repräsentation von Gesichtern werden so genannte *Face-Graphen* verwendet, siehe Abbildung 4.8. Jedem Knotenpunkt m des Graphen sind ein Jet $\mathbf{j}$ und dessen Koordinaten zugeordnet. In früheren Arbeiten [Bendlin, 2004] wurde ein Graph mit 25 Knoten und 45 Kanten verwendet, wobei die Knotenpositionen so definiert waren, dass sie teilweise im Inneren des Gesichtes, teilweise aber auch auf der Außenkontur des Kopfes lagen. Da bei dieser Anordnung aber besonders bei den niederfrequenten Filtern der Hintergrund sehr stark in des Filterergebnis eingeht, wurde der Face-Graph so verändert, dass er jetzt 38 Knoten und 56 Kanten enthält, die sich alle innerhalb des Gesichtes befinden, siehe Abbildung 4.8. Auf diese Weise werden die markanten Strukturen des Gesichtes durch den Face-Graphen wesentlich besser repräsentiert [Eckardt, 2005]. Die festgelegten Knotenpositionen stimmen gut mit den in [Lyons and Budynek, 1999] als für die Mimikanalyse besonders wichtig eingestuften Gesichtsbereichen überein.

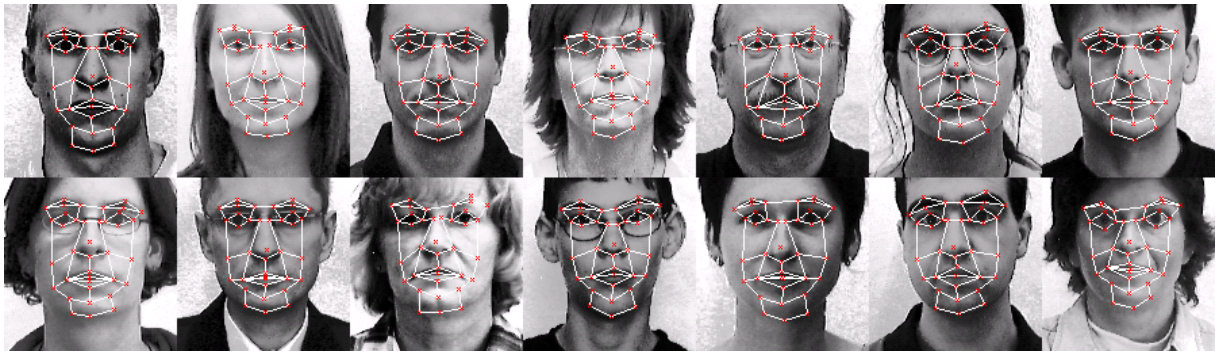


Abbildung 4.9: *Teil einer General-Face-Knowledge. Die GFK setzt sich aus den Gesichtsgraphen mehrerer Personen zusammen, wobei je nach Erkennungsaufgaben Personen unterschiedlichen Alters, unterschiedlichen Geschlechts und mit verschiedenen Gesichtsausdrücken enthalten sein sollten.*

4.4.2.4 General-Face-Knowledge

Bei der General-Face-Knowledge (GFK) handelt es sich um einen Face-Bunch-Graphen. Das bedeutet, dass Face-Graphen derselben Struktur zu einem „Bündel“ zusammengefasst werden. Dadurch existieren für jeden Knotenpunkt verschiedene Ausprägungen von Jets, jede von einem anderen Gesicht. Die General-Face-Knowledge beinhaltet Face-Graphen verschiedener Personen und wird zum Klassifizieren von Gesichtern genutzt. Da die GFK das allgemeine Wissen über Gesichter darstellt, sollte man als Basis möglichst viele voneinander verschiedene Gesichter verwenden, um später unbekannte Personen richtig klassifizieren zu können. Die GFK stellt den Ausgangspunkt für die Klassifikation eines unbekannten Eingangsbildes dar. Ein Teil einer GFK ist in Abbildung 4.9 zu sehen.

4.4.2.5 Average-Graph

Der Average-Graph wird als „Mittel“ aller bekannten Gesichter der General-Face-Knowledge gebildet. Es wird für jeden Knotenpunkt der Mittelwert aller Jets und die mittlere Position aller Knotenpunkte bestimmt. Dieser so gewonnene Graph stellt den Ausgangspunkt für die Anpassung des Modells an ein unbekanntes Eingangsbild dar, siehe Abbildung 4.10.

4.4.3 Modellanwendung

Für die Anpassung des Average-Graphen an ein Gesicht ist es notwendig, die Ähnlichkeit der aus dem Eingabebild extrahierten Jets zu den Jets im Average-Graphen zu bestimmen. Hierfür existieren verschiedene Ähnlichkeitsmaße. Nach der Anpassung eines Face-Graphen kann dieser weiter ausgewertet werden. Die Identifikation einer Person erfolgt durch die Suche nach dem ähnlichsten Graphen in einer Galerie. Für die Bestimmung des Geschlechts, des Alters und des Gesichtsausdrucks einer Person werden die Informationen der General-Face-Knowledge genutzt. Die Entscheidung wird jeweils über Ähnlichkeiten der Knotenpunkte des Bildgraphen mit denen

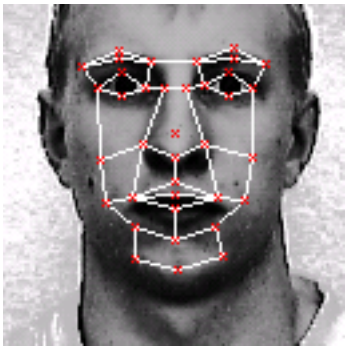


Abbildung 4.10: *Gesicht mit eingezeichnetem Average-Graph.*

der General-Face-Knowledge gefällt. In den folgenden Abschnitten werden die Teilschritte der Modellerstellung erläutert. Abbildung 4.11 zeigt die Anwendung des Elastic-Graph-Matching im Überblick.

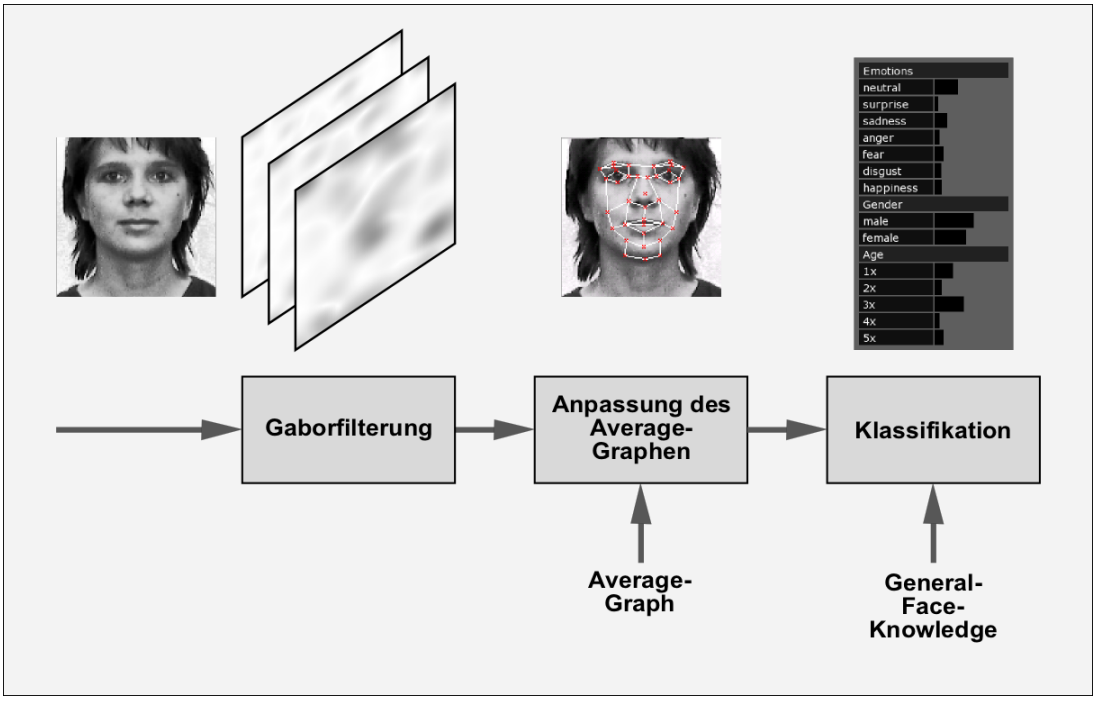


Abbildung 4.11: *Anwendung des Elastic-Graph-Matching. Nach der Gaborfilterung des Eingangsbildes entsprechend Abschnitt 4.4.2.1 wird der Average-Graph auf das Gesicht angepasst. Durch einen Vergleich des so gewonnenen Modellgraphen mit den Graphen der GFK kann eine Klassifikation des Eingangsbildes durchgeführt werden.*

4.4.3.1 Ähnlichkeitsmaße für Jets

Betragsbasiertes Ähnlichkeitsmaß Dieses Ähnlichkeitsmaß nutzt zur Berechnung nur die Beträge der komplexen Filterantworten. Da sich diese in der Umgebung eines Jets nur allmählich ändern, siehe Abbildung 4.12(b), verhält sich das betragsbasierte Ähnlichkeitsmaß entsprechend,

siehe Abbildung 4.13(b). Mit Gleichung 4.8 wird die Ähnlichkeit zwischen den Jets $\mathbf{j}$ und $\mathbf{j}'$ berechnet. Die Komponenten der Jets werden mit a_n und a'_n bezeichnet. Der Term im Nenner dient zur Normierung auf das Intervall $[0, 1]$.

$$\mathcal{S}_{\mathrm{a}}\left(\mathbf{j}, \mathbf{j}'\right) = \frac{\sum_{n=1}^{N} a_n a'_n}{\sqrt{\sum_{n=1}^{N} a_n^2 \sum_{n=1}^{N} a'^2_n}} \tag{4.8}$$

Phasenbasiertes Ähnlichkeitsmaß Für die Einbeziehung der Phaseninformation wird Gleichung 4.8 erweitert:

$$\mathcal{S}_{\phi}\left(\mathbf{j}, \mathbf{j}'\right) = \frac{\sum_{n=1}^{N} a_n a'_n \cos\left(\phi_n - \phi'_n - \mathbf{d}^T \mathbf{k}_n\right)}{\sqrt{\sum_{n=1}^{N} a_n^2 \sum_{n=1}^{N} a'^2_n}} \tag{4.9}$$

Die Symbole ϕ_n und ϕ'_n bezeichnen die Phaseninformationen der zu vergleichenden Jets, siehe Gleichung 4.6. Es wird die Differenz der beiden Phasen gebildet und zusätzlich ein Ausgleichsterm mit einbezogen, der die unterschiedlichen Phasenlagen der Jets ausgleichen soll. Dazu werden der Wellenkoeffizient $\mathbf{k}_n$ und das Displacement $\mathbf{d}$, welches im nachfolgenden Abschnitt erläutert wird, verwendet. Die Kosinusfunktion ist somit ein Maß dafür, wie gut die Phasengleichheit zwischen den Jets erreicht wurde. Als Ergebnis für das Ähnlichkeitsmaß sind Werte im Intervall $[-1, 1]$ möglich. Im Gegensatz zum Betrag der komplexen Jets ändert sich die Phase deutlich schneller, siehe Abbildung 4.12(d), wodurch das phasenbasierte Ähnlichkeitsmaß sensitiver auf Positionsänderungen reagiert. Der Verlauf des phasenbasierten Ähnlichkeitsmaßes kann durch die Wahl des so genannten Fokus beeinflusst werden. Dieser legt fest, welche Frequenzen der Jets für die Berechnung genutzt werden. So steht Fokus 0 nur für die niedrigste Frequenz und 3 für die höchste Frequenz. Durch die Verwendung höherer Frequenzen wird das Ergebnis zunehmend feiner, wodurch eine genauere Positionierung der Jets möglich wird, siehe Abbildung 4.13.

Displacement Wenn man zwei Jets miteinander vergleicht, so ist es möglich, dass sie zueinander phasenverschoben sind. Bei der Schätzung des Displacements wird versucht, diese Phasenverschiebung zu bestimmen [Wiskott et al., 1997b]. Als Ausgangspunkt der Berechnung dient das phasenbasierte Ähnlichkeitsmaß. Dieses wird in seiner Taylor Approximation maximiert, indem nach $\mathbf{d}$ differenziert wird. Das Displacement gibt damit an, um welchen Betrag der Jet $\mathbf{j}'$ zu $\mathbf{j}$ verschoben werden muss, damit das Ergebnis der Approximation maximal wird. Wie auch beim phasenbasierten Ähnlichkeitsmaß kann über den *Fokus* festgelegt werden, welche Frequenzen verwendet werden sollen. Die Herleitung der Berechnung des Displacements wird in Anhang A.6 beschrieben. Man kann das Displacement auch für die Positionierung von Jets verwenden. Dazu wird das Displacement zwischen den Jets $\mathbf{j}$ und $\mathbf{j}'$ mit einem Fokus von 0 berechnet. Anschließend wird an der neuen Position der Jet $\mathbf{j}''$ extrahiert und erneut das Displacement zu $\mathbf{j}$ berechnet,

wobei der Fokus auf 1 erhöht wird. Der Prozess ist beendet, wenn das Displacement mit Fokus 3 bestimmt und der Jet neu positioniert wurde.

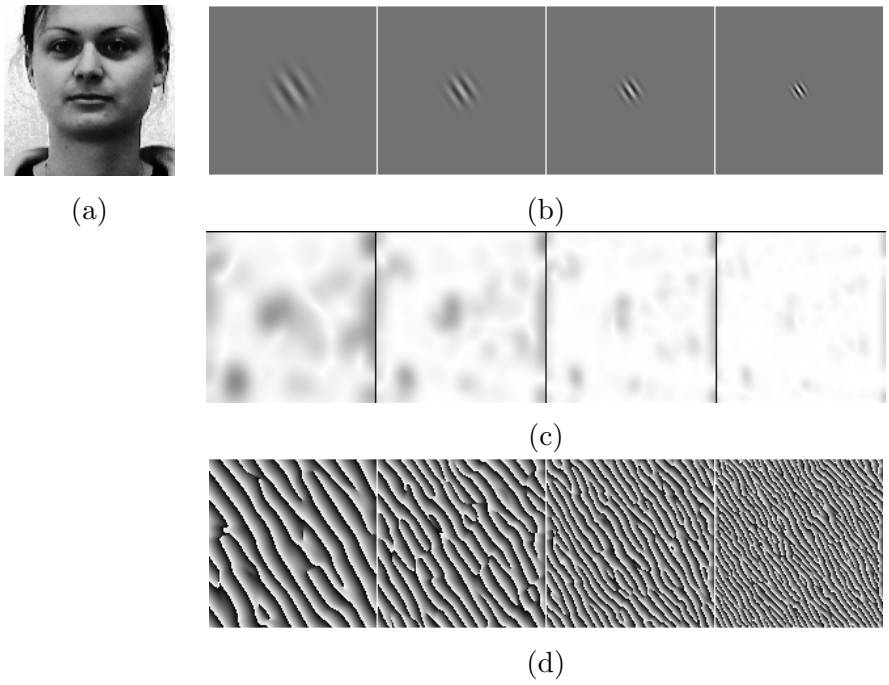


Abbildung 4.12: *Filterantworten für Gaborfilter mit einer bestimmten Orientierung, aber unterschiedlicher Frequenz. (a) Eingangsbild (b) Realteile der Gaborfilter mit von links nach rechts zunehmenden Frequenzen. (c) Betrag der Filterantwort (niedrige Werte sind hell dargestellt und hohe dunkel). Nahe beieinander liegende Jets weisen ähnliche Beträge auf. (d) Phase der Filterantwort. Im Gegensatz zum Betrag ändert sich die Phase der Filterantwort wesentlich schneller.*

Gegenüberstellung der Ähnlichkeitsmaße Das Ziel dieses Abschnittes soll es sein, die Ähnlichkeitsmaße miteinander zu vergleichen. Dazu wurde aus einem Bild ein Jet an einer markanten Position extrahiert und mit den Jets an allen anderen Bildpunkten verglichen. Die ermittelten Ähnlichkeiten werden als Grauwertbilder dargestellt, wobei ähnliche Positionen dunkler und weniger ähnliche heller dargestellt werden. Berechnet wurden das betragsbasierte und das phasenbasierte Ähnlichkeitsmaß und das Displacement. Beim phasenbasierten Ähnlichkeitsmaß und dem dazugehörenden Displacement wurde der Fokus variiert, um dessen Auswirkungen zu verdeutlichen, siehe Abbildung 4.13. Beim Displacement stehen dunkle Grauwerte für negative und helle für positive Displacements. Die unterschiedliche Sensitivität auf Ortsveränderungen kann man dahingehend nutzen, dass man zuerst das betragsbasierte Ähnlichkeitsmaß verwendet, um eine grobe Schätzung der Position durchzuführen und diese dann mit Hilfe des Displacements bzw. mit dem phasenbasierten Ähnlichkeitsmaß zu verfeinern.

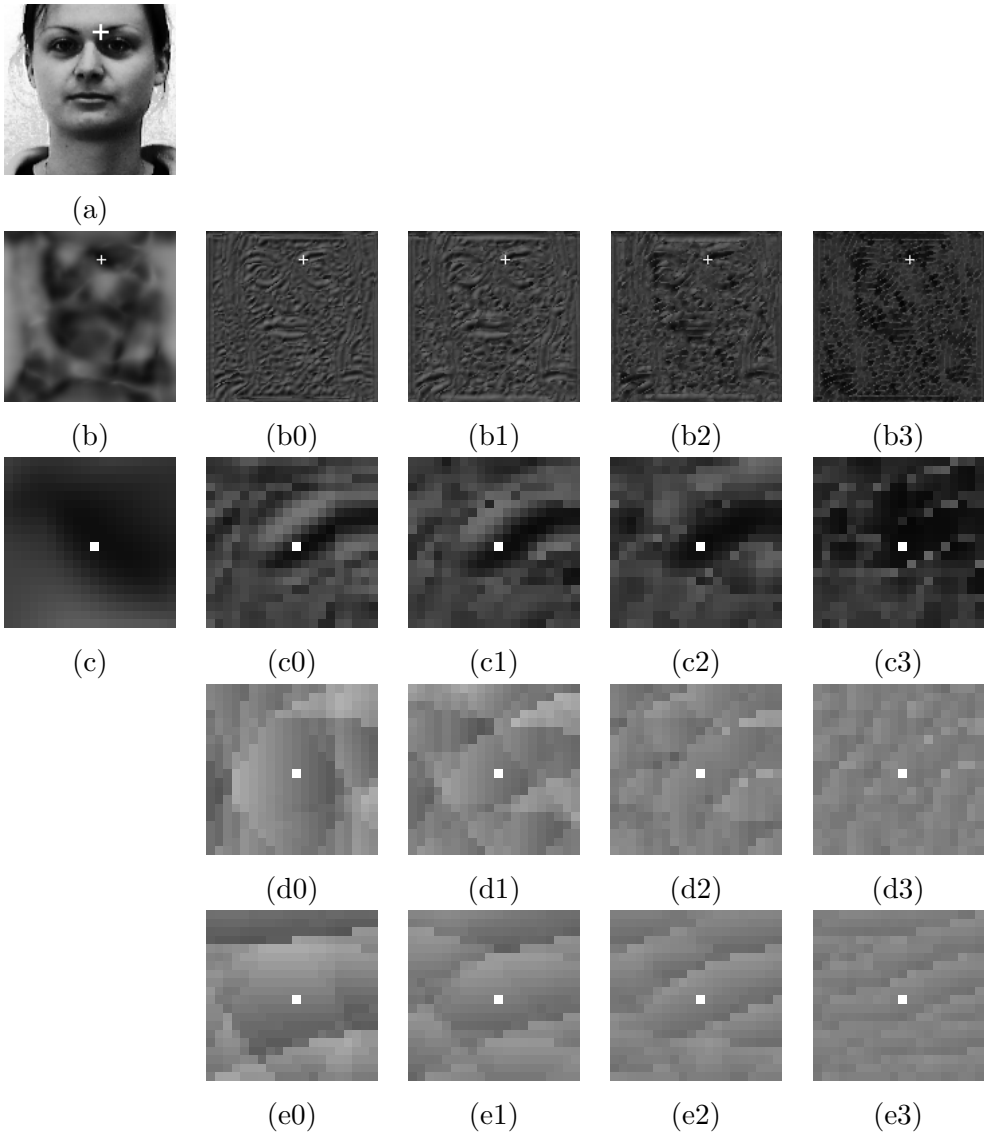


Abbildung 4.13: *Ähnlichkeitsmaße im Überblick. (a) Eingangsbild. Es wurde aus den Gaborfilterantworten an der durch das Kreuz markierten Stelle ein Jet gebildet und die Ähnlichkeitsmaße zu den Jets an allen anderen Positionen im Bild berechnet. (b) Das betragsbasierte Ähnlichkeitsmaß ändert sich nur sehr allmählich und eignet sich deshalb für eine grobe Positionierung der Jets. (b0)-(b3) Das phasenbasierte Ähnlichkeitsmaß mit wachsendem Fokus von links nach rechts. Je größer der Fokus, desto genauer kann ein Jet positioniert werden. (c) Vergrößerte Darstellung des betragsbasierten Ähnlichkeitsmaßes in einem* 20×20 *Pixel großen Ausschnitt um den Jet.(c0-c3) Vergrößerte Darstellung des phasenbasierten Ähnlichkeitsmaßes in einem* 20×20 *Pixel großen Ausschnitts um den Jet. (d0-d3)* 20×20 *Pixel großer Ausschnitt des Displacements in* x*-Richtung. (e0-e3)* 20×20 *Pixel großer Ausschnitt des Displacement in* y*-Richtung. Ein Displacement von* 0 *wird durch einen mittleren Grauwert dargestellt. Positive oder negative Displacements sind entsprechend heller bzw. dunkler. Beim Displacement in* y*-Richtung kodiert ein heller Grauwert somit eine notwendige Verschiebung nach unten und ein dunkler Grauwert eine notwendige Verschiebung des Jets nach oben.*

4.4.3.2 Ähnlichkeitsmaße für Graphen

Für den Vergleich von Graphen können zum einen die Ähnlichkeiten der einzelnen Jets und zum anderen die geometrische Ähnlichkeit zwischen den Graphen $\mathcal{G}$ und $\mathcal{G}'$ herangezogen werden. Für das erste Ähnlichkeitsmaß werden lediglich die betragsbasierten Ähnlichkeitsmaße zwischen den Jets der Graphen verwendet, siehe Gleichung 4.10. Dieses Ähnlichkeitsmaß wird z.B. bei der Personenidentifikation eingesetzt, bei der die Ähnlichkeiten zwischen einem Bildgraphen und den Graphen einer Galerie bestimmt werden:

$$\mathcal{S}_{\mathrm{G}}\left(\mathcal{G},\mathcal{G}'\right) \quad = \quad \frac{1}{M}\sum_{m=1}^{M}\mathcal{S}_{\mathrm{a}}\left(\mathbf{j}_m^{\mathcal{G}},\mathbf{j}_m^{\mathcal{G}'}\right) \tag{4.10}$$

M ist dabei die Anzahl der Jets. Ein anderes Ähnlichkeitsmaß für Graphen verwendet neben dem phasenbasierten Ähnlichkeitsmaß die mittlere quadratische Abweichung der Abstände $\Delta\mathbf{x}_e^{\mathcal{G}}$ zwischen den Positionen der Jets, wobei der Koppelfaktor λ angibt, wie stark die Form das Ergebnis beeinflusst. E ist die Anzahl der Kanten im Graphen.

$$\mathcal{S}_{\mathcal{B}}\left(\mathcal{G},\mathcal{G}'\right) \quad = \quad \frac{1}{M}\sum_{m=1}^{M}\mathcal{S}_{\phi}\left(\mathbf{j}_m^{\mathcal{G}},\mathbf{j}_m^{\mathcal{G}'}\right) - \frac{\lambda}{E}\sum_{e=1}^{E}\frac{\left(\Delta\mathbf{x}_e^{\mathcal{G}}-\Delta\mathbf{x}_e^{\mathcal{G}'}\right)^2}{\left(\Delta\mathbf{x}_e^{\mathcal{G}'}\right)^2} \tag{4.11}$$

4.4.3.3 Anpassung des Graphen

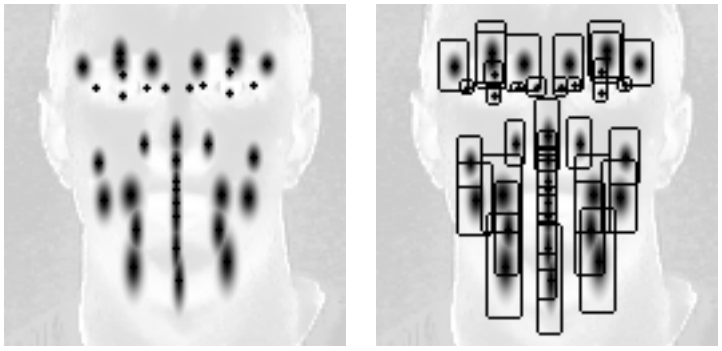
Bevor die Analyse eines Gesichtes erfolgen kann, muss der Average-Graph auf dieses angepasst werden. In der Literatur werden eine Reihe von Schritten beschrieben, mit denen ein Gesicht in einem Bild detektiert bzw. lokalisiert werden kann. Diese wurden in [Bendlin, 2004] implementiert und untersucht. Dabei handelt es sich um:

Global-Move: Der Average-Graph wird in Abständen von vier Pixeln über das Bild geschoben, und die Position mit der maximalen Ähnlichkeit zum jeweiligen Bildgraph wird bestimmt. In einem Fenster um das Maximum wird bei pixelweiser Verschiebung wiederum das Maximum bestimmt.

Skalierung: In diesem Schritt wird die Größe des Gesichtes bestimmt, indem der Average-Graph skaliert und um die im Global-Move gefundene Position wiederum die Ähnlichkeit bestimmt wird. Die Skalierung erfolgt zunächst gleichmäßig und in einem zweiten Schritt getrennt in x- und y-Richtung, so dass auch unterschiedliche Gesichtsformen berücksichtigt werden.

Local-Move: Im Gegensatz zu den vorherigen Schritten werden hier die einzelnen Knoten unter Verwendung des phasenbasierten Ähnlichkeitsmaßes unabhängig voneinander platziert. In einem Fenster von 5×5 Pixeln wird die Ähnlichkeit eines Jets des Average-Graphen zum entsprechenden Jet im Bild berechnet.

Während in [Bendlin, 2004] noch sämtliche Schritte bei der Anpassung des Graphen durchgeführt wurden, ist in der aktuellen Ausbaustufe des Systems aufgrund der Gesichtsnormalisierung aus Kapitel 3 nur noch eine lokale Anpassung des Graphen in Form eines adaptierten Local-Move nötig. Die Anpassung des Graphen wird so wesentlich beschleunigt, da die rechenintensivsten Schritte Global-Move und Skalierung nicht mehr durchgeführt werden müssen. Die Grundidee ist dabei, dass durch die feste Platzierung der Augenpunkte bei der Normalisierung deren räumliche Variation sehr gering ist, während Knotenpunkte im unteren Gesichtsbereich eine wesentlich größere Positionsvarianz aufweisen. Dieser Sachverhalt wird beim adaptierten Lokal-Move berücksichtigt, indem für jeden Knotenpunkt seine spezifische Positionsvarianz in der GFK bestimmt wird. Anhand dieser wird für jeden Jet eine Region-of-Interest bestimmt, innerhalb derer die Ähnlichkeiten zwischen dem Jet aus dem Bildgraphen und dem Average-Graphen ermittelt wird, siehe Abbildung 4.14. Die Größe des Suchfensters in x- bzw. in y-Richtung entspricht jeweils drei mal der Standardabweichung in diese Richtung. Durch diese Vorgehensweise wird der Suchaufwand für die einzelnen Knotenpukte auf das notwendige Maß eingeschränkt. An den Knotenpunkten des angepassten Graphen wird der Bildgraph extrahiert. Dieser dient als Ausgangspunkt für die im folgenden beschriebene Klassifikation.



(a) (b)

Abbildung 4.14: *(a) Knotenvarianzen, die für die GFK ermittelt wurden und für den adaptierten Local-Move verwendet werden. Die Varianzen werden über alle Graphen der GFK berechnet und dienen zur Einschränkung des Suchbereichs beim Local-Move. Es fällt auf, dass die Varianzen in Augennähe sehr klein sind, was auf die Ausrichtung des Gesichtes auf die Augenpunkte bei der Gesichtsnormalisierung zurückzuführen ist. Im unteren Gesichtsbereich sind die Varianzen vor allem in y-Richtung größer, was zum einen durch unterschiedlich lange Gesichter und zum anderen durch geöffnete Münder bei den Mimikdaten hervorgerufen wird. (b) Aus den Knotenvarianzen ermittelte Suchbereiche für den adaptierten Local-Move.*

4.4.4 Klassifikation

Nach der Anpassung des Bildgraphen liegen dessen Jets an ähnlichen Positionen wie die entsprechenden Jets der in der General-Face-Knowledge hinterlegten Modellgraphen. Die Grundidee der Klassifikation besteht darin, dass bei Graphen der selben Klasse die Jets auch ähnliche Ausprägungen aufweisen sollten, d.h. dass bestimmte durch die Jets erfasste Merkmale für die Unterscheidung verschiedener Klassen geeignet sind.

4.4.4.1 Schätzung von Geschlecht, Alter und Gesichtsausdruck

Nachfolgend wird ein Verfahren beschrieben, dass durch [Shligerskiy, 2002] entwickelt und durch [Bendlin, 2004] im Fachgebiet Neuroinformatik und Kognitive Robotik für die Mimikschätzung implementiert und untersucht wurde. Hierbei wird für jeden Jet des Face-Graphen die betragsbasierte Ähnlichkeit zu allen zugehörigen Jets der Graphen in der GFK bestimmt und anschließend die k ähnlichsten für jeden Knotenpunkt extrahiert. Im Anschluss daran erfolgt die Bestimmung der Ähnlichkeit für jeden Knotenpunkt n und die Klasse C mittels:

$$S_n^C = \frac{1}{k} \sum_{i=1}^{k} \mathcal{S}_\mathrm{a} \left(\mathbf{j}_n^{\mathcal{I}}, \mathbf{j}_n^{M_i} \right) \tag{4.12}$$

wobei $\mathbf{j}_n^{M_i}$ der Klasse C angehört. Nachdem die Ähnlichkeiten für jeden Knoten vorliegen, werden die Klassenähnlichkeiten für den gesamten Graphen nach folgender Gleichung berechnet:

$$S^C = \frac{1}{N} \sum_{n=1}^{N} S_n^C \tag{4.13}$$

Es handelt sich dabei also um eine Mehrheitsentscheidung über alle Knoten des Graphen.

4.4.4.2 Schätzung der Identität

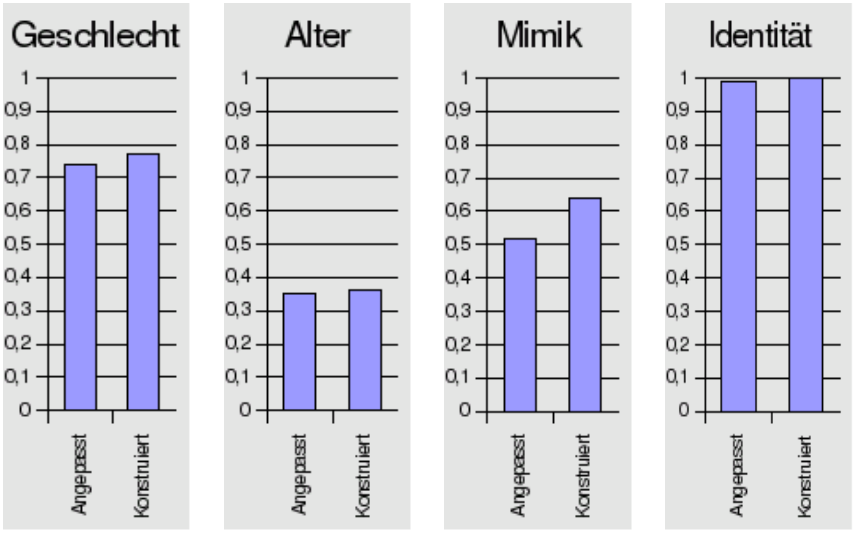
Da hierbei aus einer Gruppe bekannter Personen das dem Eingabebild ähnlichste ermittelt werden soll, erfolgt ein Vergleich des gesamten Bildgraphen mit jeweils einem Modellgraphen:

$$\mathcal{S}_\mathrm{G} \left(G^{\mathcal{I}}, G^M \right) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{S}_\mathrm{a} \left(\mathbf{j}_n^{\mathcal{I}}, \mathbf{j}_n^{M} \right) \tag{4.14}$$

Hierbei wurde für den Jet-Vergleich das betragsbasierte Ähnlichkeitsmaß verwendet, da es, wie bereits in Abschnitt 4.4.3.2 erläutert, relativ robust gegenüber leichten Verschiebungen ist. Dieses Ähnlichkeitsmaß wird jeweils zwischen dem Bildgraphen und allen Modellgraphen ermittelt. Jener Modellgraph, der die größte Ähnlichkeit mit dem Bildgraphen aufweist, kann als die zu erkennende Person angenommen werden. Bewusst wurde bei der Berechnung der Ähnlichkeit der Graphen auf einen Topologievergleich zwischen dem Bild- und Modellgraph verzichtet, da dieser posen- und vor allem mimikabhängig ist. Bei der Identifikation würden jene Modellgraphen bevorzugt werden, welche ähnliche Posen oder Gesichtsausdrücke wie der Bildgraph aufweisen [Bendlin, 2004].

4.4.5 Voruntersuchungen

An dieser Stelle soll untersucht werden, wie gut die automatische Anpassung des Average-Graphen durch den modifizierten Local-Move funktioniert. Dazu werden die Erkennungsraten für die verschiedenen Klassifikationsaufgaben zum Einen bei automatischer Anpassung des Average-Graphen durch den Local-Move und zum Anderen bei manueller Platzierung des Average-Graphen anhand der vorgegebenen Labelpunkte gegenübergestellt, siehe Abbildung 4.15.



(a) (b) (c) (d)

Abbildung 4.15: *Erkennungsraten bei automatischer Anpassung des Average-Graphen durch den modifizierten Local-Move (Angepasst) und bei manueller Platzierung des Graphen anhand der vorgegebenen Labelpunkte (Konstruiert). Die Erkennungsraten sind bei automatischer Anpassung erwartungsgemäß etwas niedriger als bei Platzierung des Graphen. Die geringen Unterschiede in den Erkennungsraten lassen darauf schließen, dass die automatische Platzierung des Graphen sehr gut funktioniert. Nur bei den Mimikdaten ist ein größerer Unterschied festzustellen.*

Abbildung 4.16 zeigt einige Beispiele für die Anpassung des Average-Graphen auf Gesichtern. Aufgrund der relativ starken Positionsabweichungen bei den Mimikdaten brechen die Erkennungsraten bei der Mimikerkennung am stärksten ein, siehe Abbildung 4.15(c). Bei neutralen Gesichtsausdrücken werden die Graphen so gut angepasst, das bei den Erkennungsraten kaum Unterschiede festzustellen sind.

Abbildung 4.16: *Beispiele für die Anpassung des Average-Graphen auf Bildern mit Gesichtsausdrücken. Es ist zu erkennen, dass die Knoten des Graphen um die Augen sehr exakt positioniert werden, wohingegen im Mundbereich, besonders bei weit offenen Mündern starke Abweichungen von der realen Position auftreten können.*

4.5 Independent-Component-Analysis

Die Grundannahme der Independent-Component-Analysis besteht darin, dass Beobachtungen $\mathbf{X}$ als Überlagerungen von unabhängigen Einzelsignalen $\mathbf{s}_i$ aufgefasst werden können. Mit der ICA wird versucht, diese nicht direkt beobachtbaren Einzelsignale zu schätzen. Dazu wird die Entmischmatrix $\hat{\mathbf{W}}$, die Invertierte der Mischmatrix A, geschätzt, mit der das Signalgemisch X in die geschätzten unabhängigen Einzelsignale $\mathbf{s}_i$ zerlegt werden kann, siehe Abbildung 4.17. Im Falle von Gesichtern handelt es sich um in ihrem Auftreten statistisch unabhängige Gesichtsmerkmale. Projiziert man ein Bild auf den von diesen Merkmalen aufgespannten Unterraum, so extrahiert man damit den Grad der Erscheinung der einzelnen Merkmale. Damit kann man von der ICA auch als Merkmalsextraktionsverfahren sprechen, das seine Informationen aus den statistischen Abhängigkeiten innerhalb der Daten selbst bezieht. Die Idee der Analyse von Gesichtern mit der ICA besteht darin, dass bestimmte Komponenten besonders typisch für bestimmte Gesichtsausdrücke, Altersklassen oder typisch für ein bestimmtes Geschlecht sind. Die extrahierten Erscheinungsgrade werden dann hinsichtlich dieser Kategorien klassifiziert.

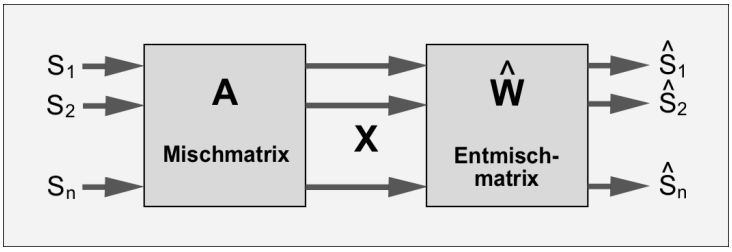
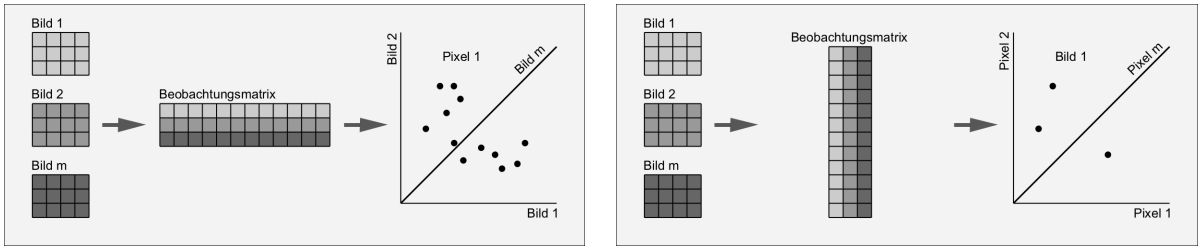


Abbildung 4.17: *Allgemeines ICA-Modell. Jede Zufallsvariable* $\mathbf{X}$ *kann als Mischung mehrerer unabhängiger Komponenten* s_i *angesehen werden. Die ICA versucht eine Entmischmatrix* $\hat{\mathbf{W}} = \mathbf{A}^{-1}$ *zu schätzen, die die Zufallsvariable wieder in ihre unabhängigen Komponenten* $\hat{s}_i$ *zerlegt.*

4.5.1 Literatur

Die Anwendung der Independent-Component-Analysis für die Analyse von Gesichtern stellt eine Weiterentwicklung der Gesichtsdetektion und -erkennung mit Eigenfaces, also Hauptkomponenten, dar [Turk and Pentland, 1991]. Umfangreiche Arbeiten existieren zur Anwendung der ICA bei der Analyse von Gesichtsausdrücken [Bartlett, 2001]. Später wurde diese auch auf die Gesichtserkennung angewendet und deren Vorteile gegenüber der PCA aufgezeigt [Bartlett et al., 2002]. Erste vergleichende Untersuchungen für die Klassifikation von Gesichtsausdrücken und Geschlecht wurden im Rahmen dieser Arbeit in [Wilhelm and Backhaus, 2004] veröffentlicht. Im nächsten Abschnitt wird auf die Vorgehensweise bei der Erstellung des ICA-Modells eingegangen.



(a) (b)

Abbildung 4.18: *Repräsentation von Bildern. (a) Im Bildraum werden die Bilder in den Zeilen der Beobachtungsmatrix angeordnet. (b) Im Pixelraum bildet jedes Bild einen Beobachtungsvektor.*

4.5.2 Modellerstellung

4.5.2.1 Bildkonvertierung in Beobachtungsvektoren

Werden Bilder, also zweidimensionale Signale betrachtet, müssen diese für die Anwendung der ICA zunächst in eine eindimensionale Darstellung überführt werden. Die Bilder werden in einer Beobachtungsmatrix $\mathbf{X}$ angeordnet, wobei jede Spalte von $\mathbf{X}$ als Beobachtungsvektor, also als Resultat eines Zufallsexperiments, aufgefasst wird. Ausgehend von m Bildern der Größe $x \times y$ ergeben sich zwei Möglichkeiten einer Serialisierung.

Bildraum: Hier werden jeweils die i-ten Pixel aller Bilder zu einem Beobachtungsvektor zusammengefasst, ein Bild bildet also eine Zeile der Beobachtungsmatrix, siehe Abbildung 4.18(a). In diesem Fall sind die Bilder Zufallsvariablen und die Pixel sind Realisierungen. Es wird versucht, *unabhängige Bilder* zu erzeugen. Zwei Bilder i und j sind dann unabhängig, wenn für alle Pixel dieser Bilder gilt, dass es nicht möglich ist, einen Pixel im Bild i durch den entsprechenden Pixel im Bild j vorherzusagen. Da dieser Raum von Achsen entsprechend der Bildindizes aufgespannt wird, bezeichnet man diese Darstellung als *Bildraum*.

Pixelraum: Hier werden alle Pixel eines Bildes zu einem Beobachtungsvektor zusammengefasst, ein Bild bildet also eine Spalte der Beobachtungsmatrix, siehe Abbildung 4.18(b). In diesem Fall sind die Pixel Zufallsvariablen und die Bilder sind Realisierungen. Es wird versucht, *unabhängige Pixel* zu erzeugen. Zwei Pixel i und j sind dann unabhängig, wenn für alle Bilder gilt, dass es nicht möglich ist, den Pixel i aufgrund des entsprechenden Pixels j im selben Bild vorherzusagen. Da dieser Raum von Achsen entsprechend der möglichen Pixelpositionen aufgespannt wird, bezeichnet man diese Darstellung als *Pixelraum*.

4.5.2.2 Zentrierung und Sphering

Das Ziel der Zentrierung und des Sphering ist die Bereinigung der Beobachtungsmatrix von den statistischen Abhängigkeiten erster und zweiter Ordnung. Dazu werden zunächst die Zeilen der

Beobachtungsmatrix $\mathbf{X}$ mittelwertfrei gemacht. Daran anschließend wird die Beobachtungsmatrix $\mathbf{X}$ einer Hauptkomponentenanalyse unterzogen, d.h. es werden die Kovarianzmatrix und deren Eigenvektoren und Eigenwerte berechnet:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T, \tag{4.15}$$

$$\mathbf{\Lambda} = Diag(\lambda_i), \tag{4.16}$$

$$\mathbf{T} = [\mathbf{v}^{(0)}, ..., \mathbf{v}^{(m-1)}]. \tag{4.17}$$

Dabei ist $\mathbf{\Lambda}$ eine Diagonalmatrix mit den Eigenwerten von $\mathbf{C}$ absteigend nach der Größe auf der Hauptdiagonalen und $\mathbf{T}$ eine Matrix mit den Eigenvektoren in den Spalten. Im Pixelraum wird aufgrund der hohen Dimension der Spalten der Matrix $\mathbf{X}$ die Kovarianzmatrix sehr groß, was die Berechnung der Eigenwerte aufwendig macht. Dieses Problem kann aber durch einen Trick umgangen werden, bei dem nicht die Kovarianzmatrix im Pixelraum, sondern im Bildraum berechnet wird, siehe Anhang A.7. Im Ergebnis der PCA steht ein durch die Eigenvektoren aufgespanntes Orthonormalsystem zur Verfügung, dessen Kovarianzmatrix der Matrix $\mathbf{\Lambda}$ mit den Eigenwerten entspricht. Das Sphering hat zum Ziel, die Varianz entlang jedes Eigenvektors auf Eins zu normieren. Dazu wird jeder Eigenvektor durch die Wurzel des zugehörigen Eigenwertes dividiert. Schließlich wird die Beobachtungsmatrix $\mathbf{X}$ auf den von der Sphering-Matrix $\mathbf{W}_S$ aufgespannten Raum projiziert:

$$\mathbf{W}_S = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{T}^T, \tag{4.18}$$

$$\mathbf{\Lambda}^{-\frac{1}{2}} = Diag(-\frac{1}{\sqrt{\lambda_i}}), 0 \leq i < m \tag{4.19}$$

$$\mathbf{X}_S = \mathbf{W}_S\mathbf{X} \tag{4.20}$$

Nach dem Sphering gilt für die Matrix $\mathbf{X}_S$ die folgende Bedingung:

$$E\{\mathbf{X}_S\mathbf{X}_S^T\} = \mathbf{I}, \tag{4.21}$$

wobei $\mathbf{I}$ die Einheitsmatrix ist. Während der PCA kann eine Dimensionsreduktion des durch die Beobachtungsmatrix aufgespannten Raumes durchgeführt werden. Hierdurch besteht die Möglichkeit, Störungen im Mischsignal oder ganze latente Quellen zu unterdrücken. Jedoch besteht dabei auch die Gefahr, wichtige Merkmale für die spätere Klassifikation zu entfernen. Um zu bestimmen, wie viele Eigenvektoren für die Unterraumprojektion verwendet werden sollen, werden deren Eigenwerte betrachtet, welche die Varianz der Projektionswerte entlang des entsprechenden Eigenvektors darstellen. Es werden nach folgendem Kriterium die n eigenwertgrößten Eigenvektoren aus $\mathbf{T}$ extrahiert:

$$\frac{\sum_{i=1}^{n-1}\lambda_i}{\sum_{i=1}^{m-1}\lambda_i} \geq r. \tag{4.22}$$

Abbildung 4.19 zeigt ein Beispiel, in dem dieses Maß auf einen Beispieldatensatz angewandt wurde, der einen 490-dimensionalen Raum aufspannt. Es ist zu sehen, dass die ersten 133 eigenwertgrößten Eigenvektoren bereits eine Varianz von 98% abdecken. Somit könnte eine Projektion auf einen 133-dimensionalen Unterraum durchgeführt werden, ohne dass dies eine starke Reduktion der Varianz zur Folge hätte. Beim Sphering würde dann entsprechend mit der kleineren Dimension n gearbeitet, siehe Gleichung 4.19.

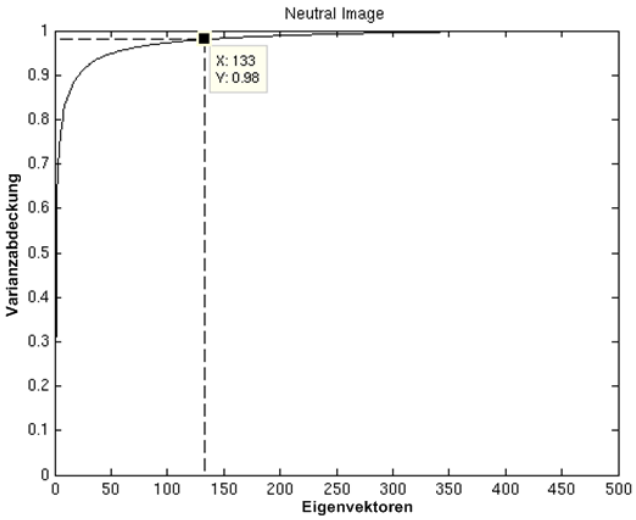


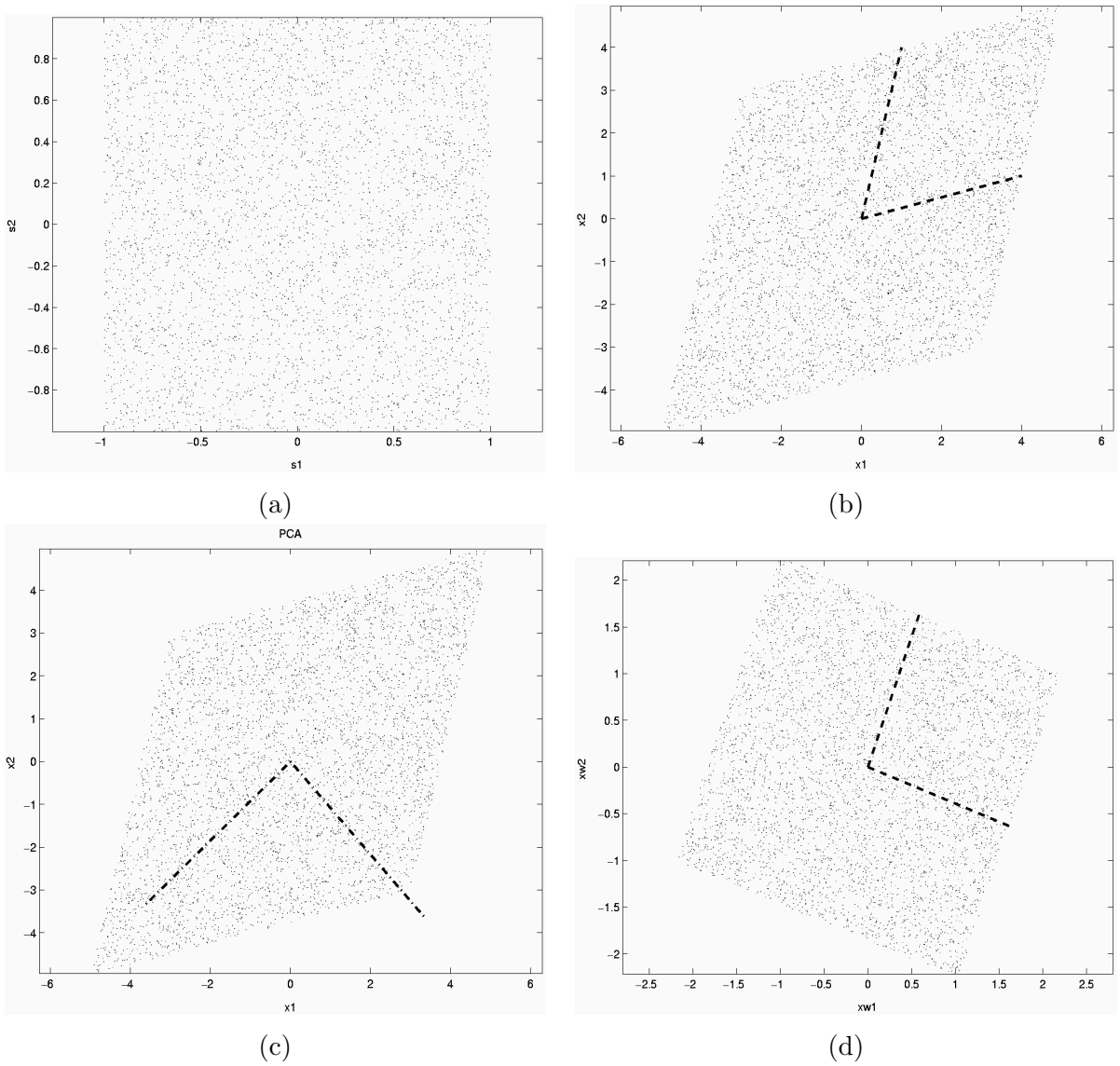
Abbildung 4.19: *Bestimmung der Dimension des PCA-Unterraums beim Sphering. Dargestellt ist ein Datensatz, der einen Unterraum mit* 490 *Dimensionen aufspannt. Im Diagramm wurde die Anzahl der eigenwertgrößten Eigenvektoren markiert, bei der eine Varianzabdeckung von* 98% *erreicht wird. Der Unterraum kann somit ohne eine große Varianzreduktion auf* 133 *Dimensionen verkleinert werden.*

Die Auswirkung des Sphering-Schrittes auf eine Zufallsverteilung wird in Abbildung 4.20 veranschaulicht. Nach dem Sphering muss nur noch eine orthogonale Mischmatrix geschätzt werden, wodurch sich der Freiheitsgrad bei der anschließenden ICA reduziert.

4.5.2.3 Independent-Component-Analysis

Nachdem die beiden Vorverarbeitungsschritte Zentrierung und Sphering durchlaufen wurden, stehen mittelwertfreie, dekorrelierte und normierte Daten zur Verfügung. Im nächsten Schritt werden diese Daten mittels der ICA so transformiert, dass im Ergebnis unabhängige Komponenten entstehen. Hierfür können die unterschiedlichsten Ansätze verfolgt werden. Die Verfahren bestehen jedoch immer aus einer Kontrastfunktion $C(\hat{\mathbf{S}})$, die ein Maß für die Unabhängigkeit der separierten Quellen $\hat{\mathbf{S}}$ berechnet, und einer Optimierungsstrategie zur Adaption der Entmischmatrix $\hat{\mathbf{W}}$. Das Ziel besteht darin, die Entmischmatrix so zu verändern, dass im Ergebnis möglichst statistisch unabhängige Quellen entstehen.

In [Backhaus, 2003] wurden für das Erstellen einer geeigneten Kontrastfunktion verschiedene Prinzipien implementiert und verglichen. Diese können zwei Hauptgruppen zugeordnet werden, den stochastischen und den informationstheoretischen Prinzipien. Der *FastICA*-Algorithmus von



(a) (b)

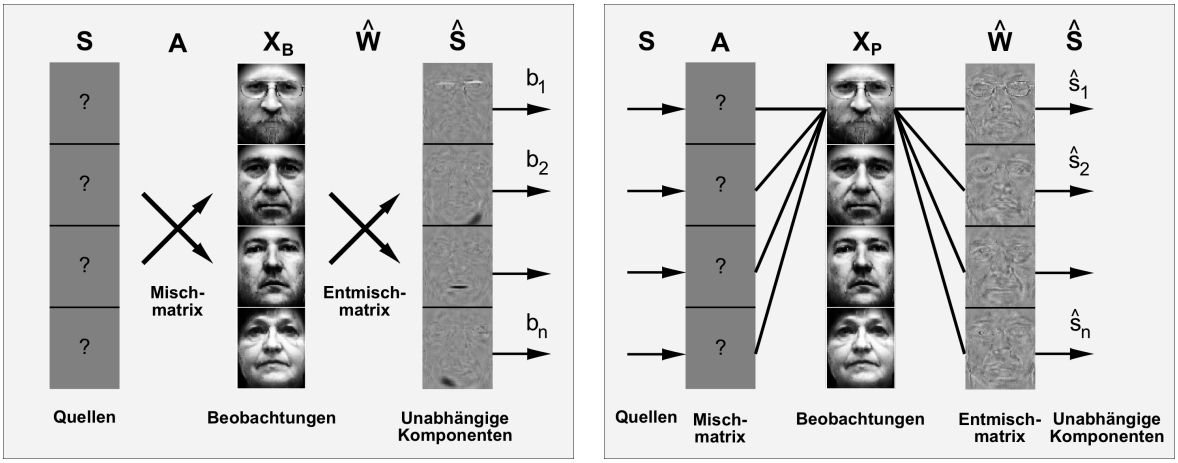
(c) (d)

Abbildung 4.20: *Testdatensatz aus zwei gleichverteilten Zufallskomponenten: (a) Originalkomponenten vor Mischung, (b) gemischte Komponenten vor dem Sphering, (c) gefundene Hauptkomponenten und (d) gemischte Komponenten nach dem Sphering.*

Hyvärinen und Oja [Hyvärinen and Oja, 2000] basiert auf der Maximierung der Nichtgaußhaftigkeit der geschätzten unabhängigen Quellen, der *InfoMax*-Algorithmus von Bell und Sejnowski [Bell and Sejnowski, 1995] auf der Maximierung der Ausgabeentropie, bzw. des Informationsflusses eines neuronalen Netzwerkes mit nichtlinearen Ausgaben. Bei den Untersuchungen in [Backhaus, 2003] stellte sich heraus, dass beide Verfahren eine gute Performanz bei der Separierung der Daten aufweisen. Jedoch müssen bei Verwendung des InfoMax-Algorithmus Nichtlinearitäten (subgauß oder supergauß) als Ausgabefunktion des verwendeten neuronalen Netzes angegeben werden. Stimmen diese nicht mit der tatsächlichen Verteilung der unabhängigen Komponenten überein, kann eine Extraktion der unabhängigen Komponenten nicht garantiert werden. Dagegen können beim FastICA-Algorithmus die Quellen ohne a priori Wissen über deren Verteilung separiert werden. Aus diesem Grund wird in dieser Arbeit ausschließlich der FastICA-

Algorithmus eingesetzt. Eine Beschreibung dieses Algorithmus befindet sich in Anhang A.8.

4.5.2.4 Erzeugung der Basisbilder

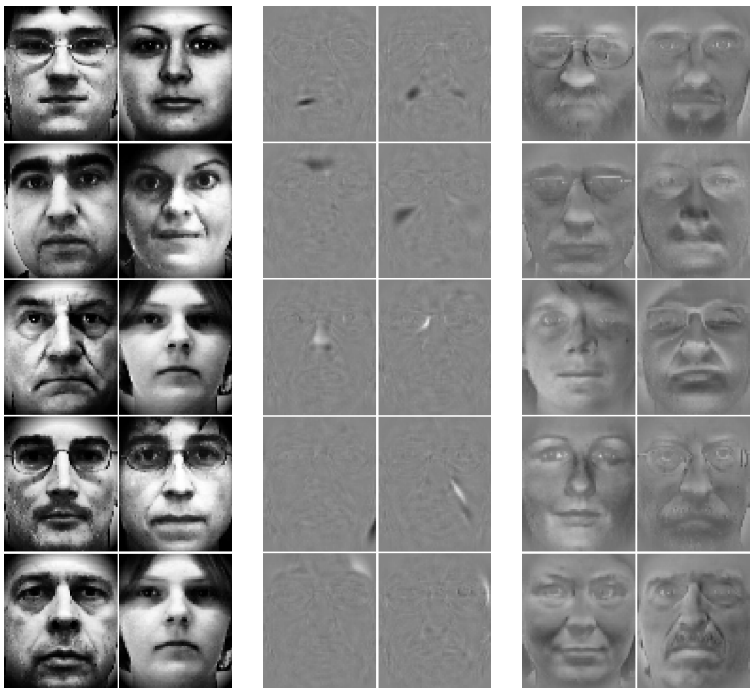


(a) (b)

Abbildung 4.21: *Erzeugung der Basisbilder. (a) Im Bildraum entsprechen die Basisbilder den Zeilen der Matrix* $\hat{\mathbf{S}}$*. (b) Im Pixelraum entsprechen die Basisbilder den Zeilen der Entmischmatrix* $\hat{\mathbf{W}}$*.*

Bildraum: Bei der Darstellung im Bildraum wird versucht, eine Menge statistisch unabhängiger Basisbilder zu erzeugen. Dagegen sind die Koeffizienten b_i, die die einzelnen Gesichter kodieren, nicht notwendigerweise unabhängig. Die ICA schätzt eine Entmischmatrix $\hat{\mathbf{W}}$ so, dass diese eine größtmögliche statistische Unabhängigkeit zwischen den Zeilen von $\hat{\mathbf{S}} = \hat{\mathbf{W}}\mathbf{X}_B$ erzeugt, siehe Abbildung 4.21(a). Die Zeilen der Matrix $\hat{\mathbf{S}}$ lassen sich als Bilder darstellen. Abbildung 4.22(b) zeigt eine Auswahl der gefundenen unabhängigen Komponenten. Wie man sieht, sind diese von lokaler Natur und stellen einzelne Gesichtsmerkmale dar. Sie repräsentieren jeweils die Pixel in den Eingangsdaten, die ein ähnliches Verhalten aufweisen. Das rührt daher, dass statistische Abhängigkeiten in räumlich benachbarten Pixeln bestehen. Die Basisbilder sind außerdem spärlich in dem Sinne, dass viele Pixel Werte nahe Null aufweisen.

Pixelraum: Durch Anwendung der Pixelraumdarstellung erhält man einen spärlichen, statistisch unabhängigen Koeffizientenvektor $\hat{\mathbf{W}}$, auch Faktorcode genannt. Wurde die Entmischmatrix geschätzt, so ergeben sich nach dem inversen ICA-Modell $\hat{\mathbf{S}} = \hat{\mathbf{W}}\mathbf{X}_P$ unabhängige Komponenten in den Zeilen der Matrix $\hat{\mathbf{S}}$. Die Basisbilder für die Projektion sind in der Entmischmatrix $\hat{\mathbf{W}}$ enthalten. Aufgrund des Spärlichkeitskriteriums erhält man eine Darstellung, in der einzelne Basisbilder nur sehr selten aktiviert werden. Dies kann so interpretiert werden, dass jedes Basisbild Merkmale eines bestimmten Bildes der Trainingsmenge enthält, siehe Abbildung 4.22(c)

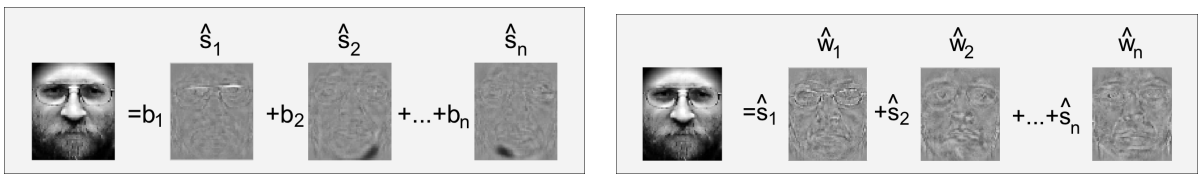


(a) (b) (c)

Abbildung 4.22: *(a) Einige Beispiele für die verwendeten Bilddaten (b) Beispiele für Basisbilder im Bildraum (unabhängige Komponenten). Auffällig ist hier, dass die entstandenen unabhängigen Komponenten aus lokalen Bildstrukturen bestehen, die einfachen Kantenfiltern ähneln. (c) Beispiele für Basisbilder im Pixelraum (Entmischmatrix). Da im Pixelraum ein unabhängiger Faktorkode entsteht, wird jedes Basisbild nur sehr selten aktiviert und enthält somit die typischen Eigenschaften jeweils eines Bildes in der Beobachtungsmatrix.*

4.5.3 Modellanwendung

Bei der Anwendung des ICA-Modells werden die Bilddaten auf den bei der Modellerstellung ermittelten Unterraum projiziert. Die Bilder haben in dieser Arbeit immer eine Größe von 60×70 Pixeln, wobei die Augenmittelpunkte auf eine feste Position im Bild platziert werden. Außerdem erfolgt für jedes Bild ein Histogrammausgleich (siehe Anhang A.3.1) und eine ovale Abschattung der Randbereiche, um den Hintergrund bei der Bildanalyse auszublenden, siehe Abbildung 4.22(a).



(a) (b)

Abbildung 4.23: *Projektion der Bilddaten auf die durch die ICA ermittelten Basisbilder. (a) Im Bildraum sind dies die unabhängigen Komponenten $\hat{\mathbf{s}}_i$. (b) Im Pixelraum handelt es sich um die Zeilen der Entmischmatrix $\hat{\mathbf{w}}_i$.*

Datensatz	Bildraum	Pixelraum
Mimik	130	160
Neutral	140	250

Tabelle 4.2: *Aufgrund des Varianzkriteriums ermittelte Anzahl an Eigenvektoren.*

Bildraum: Die Projektion der Beobachtungsmatrix $\mathbf{X}_B$ erfolgt auf die unabhängigen Basisbilder, wodurch eine Projektionsmatrix $\mathbf{B}$ entsteht, die die Koeffizienten der Bilder enthält, siehe Gleichung 4.23. Abbildung 4.23(a) zeigt die Repräsentation eines Eingabebildes durch einen Satz von Basisbildern (unabhängigen Komponenten) $\hat{\mathbf{S}}$.

$$\mathbf{B} = \hat{\mathbf{S}}\mathbf{X}_B^T \tag{4.23}$$

Pixelraum: Im Pixelraum erfolgt die Projektion der Beobachtungsmatrix $\mathbf{X}_P$ auf die Entmischmatrix $\hat{\mathbf{W}}$, siehe Gleichung 4.24. Abbildung 4.23(b) zeigt die Repräsentation eines Eingabebildes durch einen Satz von Basisbildern.

$$\mathbf{B} = \hat{\mathbf{W}}\mathbf{X}_P \tag{4.24}$$

In beiden Fällen dienen die Fit-Werte $\mathbf{B}$ als Grundlage für die Klassifikation.

4.5.4 Klassifikation

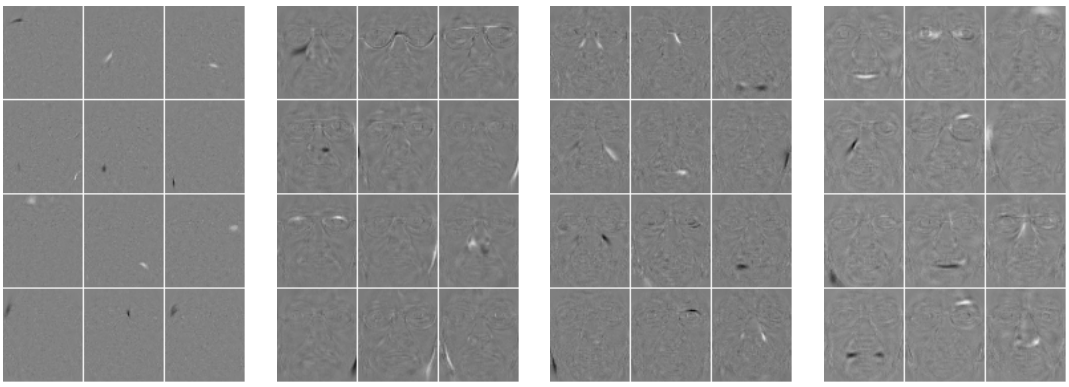
Die bei der Projektion eines Eingabebildes auf die Basisbilder entstehenden Koeffizienten sind ein Maß für die Ausprägung der einzelnen Merkmale in diesem Bild. Für die Klassifikation dieser Koeffizienten hinsichtlich Alter, Geschlecht, Gesichtsausdruck und Identität wurden eine Reihe von Klassifikatoren vergleichend untersucht. Dabei handelt es sich um Nearest-Neighbor-Klassifikatoren, Multilayer-Perceptrons, Radial-Basis-Function-Netze und verschiedene Learning-Vector-Quantifier.

4.5.5 Voruntersuchungen

Unterraumreduzierung Zunächst wurde aus beiden Datensätzen ein unreduzierter Unterraum und einer mit reduzierter Anzahl an Dimensionen sowohl für die Bildraum- als auch für die Pixelraumdarstellung erstellt. Als Anhaltspunkt für die Reduzierung diente das in Abschnitt 4.5.2.2 beschriebene Varianzkriterium mit einer Varianzschwelle von 98% der Gesamtvarianz. Tabelle 4.2 zeigt die Anzahl der beim Sphering-Schritt verwendeten eigenwertgrößten Eigenvektoren.

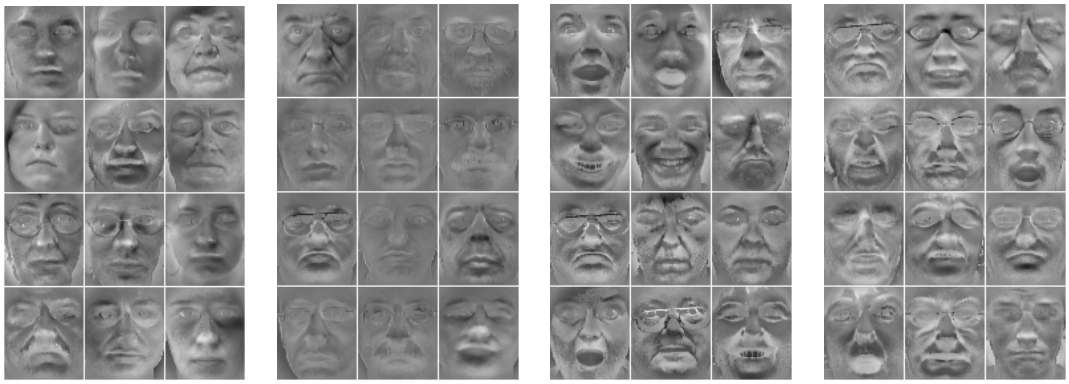
Die durch die ICA ermittelten Basisbilder für den Bildraum sind in Abbildung 4.24 und für den Pixelraum in Abbildung 4.25 dargestellt. Durch die Reduktion im Bildraum entstehen weniger

unabhängige Komponenten, die aber jeweils eine größere zusammenhängende Fläche abdecken. Im Pixelraum beschreibt ein Basisbild die typische Erscheinung eines Eingabebildes, d.h. letztendlich einer Person.



(a) (b) (c) (d)

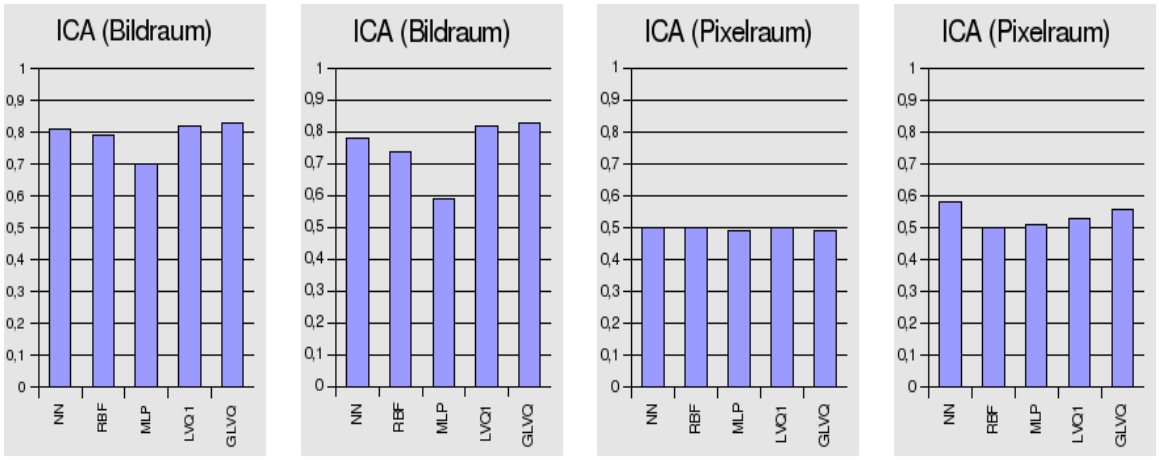
Abbildung 4.24: *Mit der FastICA ermittelte Basisbilder im Bildraum. (a) Einige unabhängige Komponenten der Neutraldaten (unreduziert), (b) einige unabhängige Komponenten der Neutraldaten (reduziert), (c) einige unabhängige Komponenten der Mimikdaten (unreduziert), (d) einige unabhängige Komponenten der Mimikdaten (reduziert). Durch die Reduktion entstehen weniger unabhängige Komponenten, die aber jeweils eine größere zusammenhängende Fläche abdecken.*



(a) (b) (c) (d)

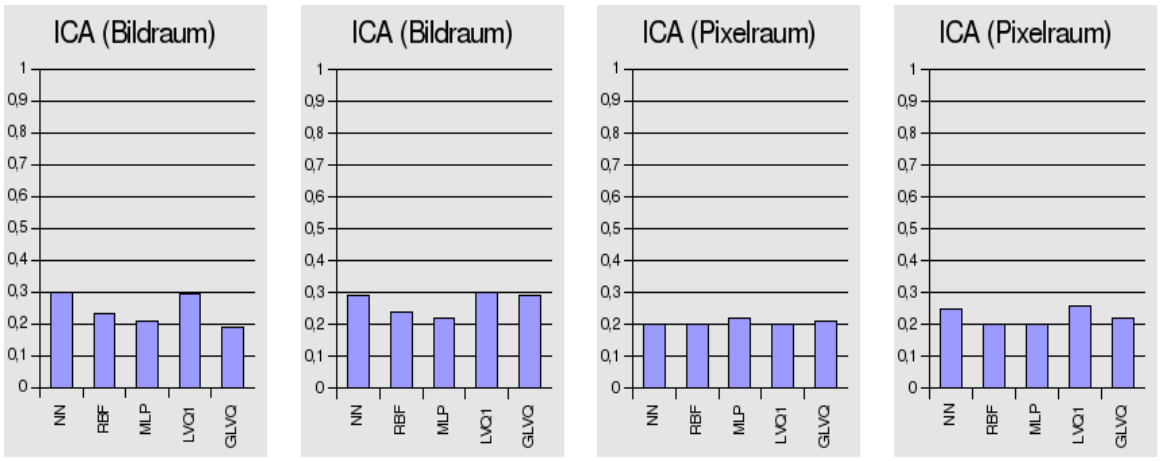
Abbildung 4.25: *Mit der FastICA ermittelte Basisbilder im Pixelraum. Zur besseren Veranschaulichung werden hierbei die Misch- und nicht die Entmischmatrizen dargestellt: (a) Mischmatrix aus den Neutraldaten (unreduziert), (b) Mischmatrix aus den Neutraldaten (reduziert), (c) Mischmatrix aus den Mimikdaten (unreduziert), (d) Mischmatrix aus den Mimikdaten (reduziert).*

Die Erkennungsraten für die Geschlechtsschätzung auf den Neutraldaten sind in Abbildung 4.26 abgebildet. Die besten Erkennungsraten werden mit der ICA im Bildraum ohne Reduktion der Anzahl der unabhängigen Komponenten erreicht. Im Pixelraum sind die Erkennungsraten bei der Geschlechtsschätzung niedriger als im Bildraum.



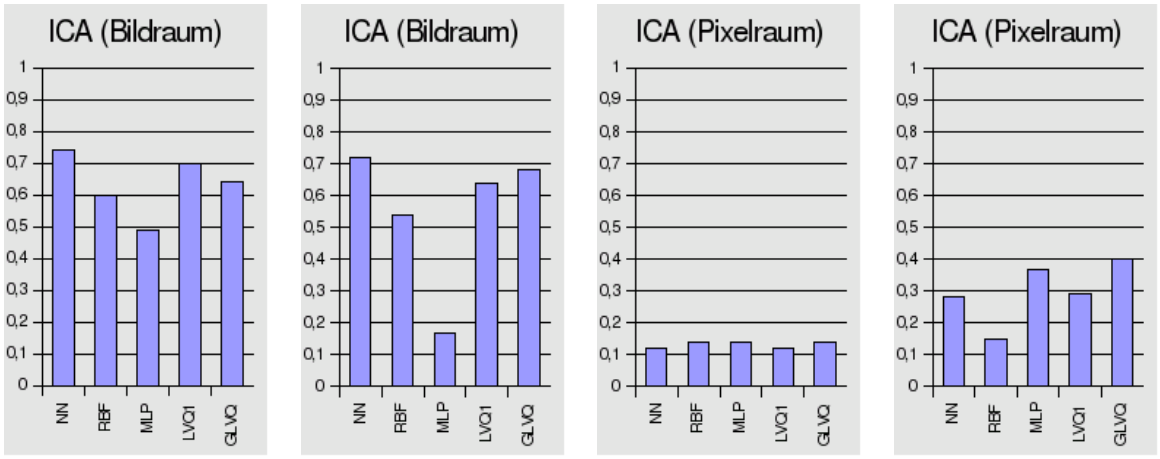
(a) (b) (c) (d)

Abbildung 4.26: *Erkennungsraten der ICA bei der* **Geschlechtsschätzung** *(a) im Bildraum (unreduziert) (b) im Bildraum (reduziert) (c) im Pixelraum (unreduziert) und (d) im Pixelraum (reduziert). Durch die Reduktion im Bildraum steigt die Erkennungsrate bei Verwendung eines MLP, während bei LVQ-Netzwerken keine Verbesserung erreicht werden kann. Die Erkennungsraten im Pixelraum sind durchgängig schlechter als im Bildraum. Ohne Reduktion ist im Pixelraum keine Geschlechtsschätzung möglich.*



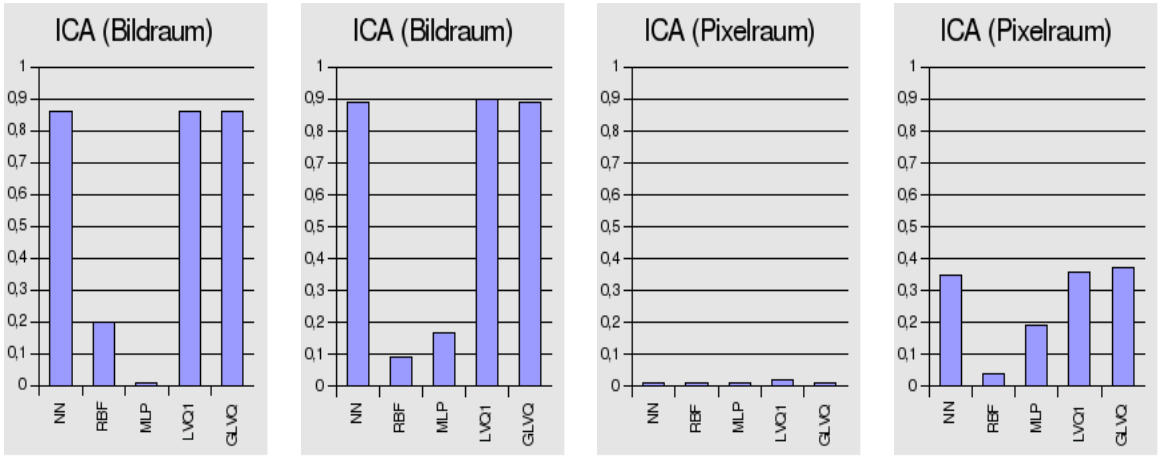
(a) (b) (c) (d)

Abbildung 4.27: *Erkennungsraten der ICA bei der* **Altersschätzung** *(a) im Bildraum (unreduziert) (b) im Bildraum (reduziert) (c) im Pixelraum (unreduziert) und (d) im Pixelraum (reduziert). Durch die Reduktion im Bildraum steigt die Erkennungsrate bei Verwendung eines MLP, während bei LVQ-Netzwerken keine Verbesserung erreicht werden kann. Die Erkennungsraten im Pixelraum sind durchgängig schlechter als im Bildraum. Ohne Reduktion ist im Pixelraum keine Altersschätzung möglich. Die minimale Erkennungsrate bei der Altersschätzung beträgt aufgrund der verwendeten 5 Altersklassen $0,2$. (Dies wäre die Erkennungsrate, wenn das Alter geraten würde).*



(a) (b) (c) (d)

Abbildung 4.28: *Erkennungsraten der ICA bei der* **Mimikschätzung** *(a) im Bildraum (unreduziert) (b) im Bildraum (reduziert) (c) im Pixelraum (unreduziert) und (d) im Pixelraum (reduziert). Die Erkennungsraten im Pixelraum sind durchgängig schlechter als im Bildraum. Ohne Reduktion ist im Pixelraum keine Mimikschätzung möglich.*



(a) (b) (c) (d)

Abbildung 4.29: *Erkennungsraten der ICA bei der* **Identitätsschätzung** *(a) im Bildraum (unreduziert) (b) im Bildraum (reduziert) (c) im Pixelraum (unreduziert) und (d) im Pixelraum (reduziert). Die Erkennungsraten im Pixelraum sind schlechter als im Bildraum. Ohne Reduktion ist im Pixelraum keine Identitätsschätzung möglich.*

4.6 Active-Appearance-Models

Das Aussehen von Objekten hängt unter anderem sehr stark von natürlichen Formvariationen, unterschiedlichen Beleuchtungsbedingungen und der räumlichen Orientierung ab. Active-Appearance-Models (AAMs) sind hinreichend flexibel, um solche Variationen modellieren zu können und gleichzeitig hinreichend spezifisch, um eine bestimmte Klasse von Objekten beschreiben zu können. Bei Active-Appearance-Models handelt es sich um parametrische generative Modelle für die Bildinterpretation. Sie setzen sich zusammen aus Form- und Grauwertmodellen, die durch relativ wenige sogenannte Appearance-Parameter beschrieben werden. Active-Appearance-Models sind in der Lage, sich auf ein unbekanntes Eingabebild anzupassen, wobei die Parameter des Modells entsprechend adaptiert werden. Diese bei diesem Anpassungsprozess erzeugten Parameter können für eine realistische Synthese des Gesichtes verwendet werden und bilden die Grundlage für die Klassifikation.

4.6.1 Literatur

Die Grundlage für die Entwicklung der Active-Appearance-Models bildeten die *Active-Shape-Models*. [Cootes et al., 1992a] [Cootes et al., 1992b] [Cootes and Taylor, 1992] stellten ein parametrisches statistisches Formmodell vor, bei dem eine PCA auf die Abstände zwischen Labelpunkten in Bildern angewendet wurde. Dabei entsteht ein sogenanntes *Point-Distribution-Model*, das zum einen die typische Form der Objektklasse und zum anderen deren Variabilität in den Trainingsdaten beschreibt. In [Cootes et al., 1998] wurden die Active-Appearance-Models als Kombination von Formmodell und Grauwertmodell eingeführt. In [Cootes et al., 2000] wird eine Methode vorgestellt, mit der mit wenigen zweidimensionalen AAMs die Form und die Erscheinung eines Gesichtes aus jedem beliebigen Blickwinkel erfasst werden kann. Eine Anwendung für die Personenidentifikation mit AAMs wird in [Costen et al., 1999] vorgestellt. In [Dornaika and Ahlberg, 2004] konnte gezeigt werden, dass sich AAMs auch zum kontinuierlichen Verfolgen von Gesichtern im Videostrom anwenden lassen. In [Lanitis et al., 1995] werden die Active-Appearance-Modele für der Analyse von Gesichtsausdrücken eingesetzt. Eine Anwendung für die Personenerkennung wird in [Edwards and Cootes, 1998] vorgestellt. In [Wu et al., 2003] wurde gezeigt, dass das Konzept sogar zur Entfernung von Brillengläsern in Gesichtsbildern eingesetzt werden kann. Im technischen Report [Cootes and Taylor, 1999] bieten Cootes und Tayler einen umfassenden Überblick über die Entwicklung und die Anwendung der Active-Appearance-Models.

4.6.2 Modellerstellung

Ein AAM besteht aus einem Form- und einem Grauwertmodell. Die einzelnen Schritte für den Aufbau eines AAMs werden in den folgenden Abschnitten beschrieben.

4.6.2.1 Formmodell

Das Formmodell (Active-Shape-Model) ist eine Weiterentwicklung der *Active-Contour-Models* oder *Snakes* von Kass et al. [Kass et al., 1987]. Der Vorteil von Active-Shape-Models besteht allerdings darin, dass sie spezifisch für die in den Trainingsdaten vorhandenen Formen sind, während eine Snake sich beliebigen Konturen anpassen kann. Die Beschreibung der typischen Formen von Objekten einer Trainingsmenge basiert auf den räumlichen Variationen sogenannter Label-Punkte oder Landmarken. Jeder Label-Punkt hat eine bestimmte Verteilung, die als Point-Distribution-Model (PDM) bezeichnet wird. Zunächst werden die Variationen der Label-Punkte in einer kompakten Form repräsentiert. Wenn eine Form durch n Punkte in d Dimensionen beschrieben wird, kann diese durch einen Vektor mit dn Elementen repräsentiert werden, indem alle Punktkoordinaten hintereinander geschrieben werden. Für die n Labelpunkte (x_i, y_i) eines Grauwertbildes $\mathbf{I}$ entsteht folgender Vektor mit $2n$ Elementen:

$$\mathbf{x} = (x_1, y_1, \ldots, x_n, y_n) \tag{4.25}$$

Abbildung 4.30 zeigt ein Beispiel für ein vollständig gelabeltes Trainingsbild.

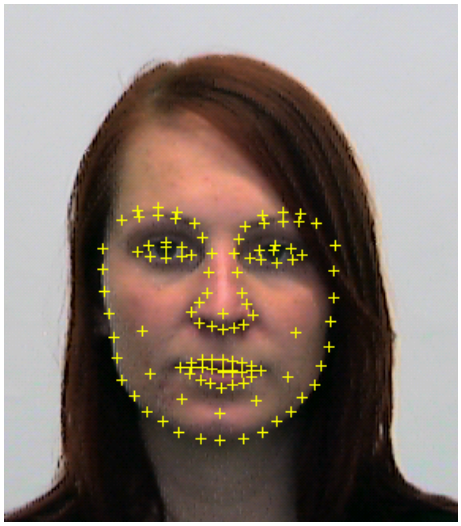


Abbildung 4.30: *Vollständig gelabeltes Trainingsbild für das Formmodell.*

Für s Trainingsbeispiele werden s solcher Vektoren gebildet. Bevor eine statistische Datenanalyse durchgeführt werden kann, müssen die Daten ausgerichtet werden. Dazu werden die einzelnen Formen in den gemeinsamen Schwerpunkt verschoben und so rotiert und skaliert, dass der mittlere quadratische Fehler zur mittleren Form minimal wird. Auf diese Weise werden Variationen aus den Daten entfernt, die auf nicht formverändernde Transformationen, also Rotation, Trans-

lation und Skalierung zurückzuführen sind. Ziel ist es, den quadratischen Fehler aller Vektoren $\mathbf{x}_i$ zum mittleren Vektor $\mathbf{x}'$ zu minimieren, so dass folgende Energie minimal wird:

$$E = \sum_i |\mathbf{x}_i - \mathbf{x}'|^2 \tag{4.26}$$

Die affine Transformation lässt sich durch folgende Matrixmultiplikation darstellen:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{4.27}$$

Hierbei beschreiben t_x und t_y die translatorischen Anteile in x- und y-Richtung, a den Skalierungsfaktor und b den Rotationsfaktor. Da der Formvektor $\mathbf{x}$ unter Verwendung dieser Transformation auf den Bezugsvektor $\mathbf{x}'$ überführt werden kann, ergibt sich folgende Gleichung zur Minimierung der Energie:

$$E(a, b, t_x, t_y) = \sum_{i=1}^{n} (ax_i - by_i + t_x - x'_i)^2 + (bx_i - ay_i + t_y - y'_i)^2 \tag{4.28}$$

Über die Lösung dieser Gleichung erhält man die Parameter der affinen Transformation $\mathbf{T}$. Dabei lässt sich die Berechung entscheidend vereinfachen, wenn man davon ausgeht, dass die Schwerpunkte sämtlicher Formvektoren im Ursprung liegen. Unter Verwendung dieser Annahme bzw. Verschiebung sämtlicher Formvektoren sowie partieller Ableitung und Gleichsetzen mit Null erhält man aus Gleichung 4.28 folgende Berechnungsvorschriften für die Transformationsparameter:

$$t_x = \frac{1}{n} \sum x'_i \tag{4.29}$$

$$t_y = \frac{1}{n} \sum y'_i \tag{4.30}$$

$$a = XX'/|X|^2 \tag{4.31}$$

$$b = \left(\frac{1}{n} \sum x_i y'_i - \frac{1}{n} \sum y_i x'_i \right) |\mathbf{x}|^2 \tag{4.32}$$

Als Bezugsvektor während des Ausrichtens dient der Mittelwertvektor über alle Formdaten. Da sich dieser aufgrund des Anpassungsprozesses ändert, handelt es sich bei der Bestimmung des Bezugsvektors und der Anpassung um einen rekursiven Prozess, welcher solange fortgeführt wird, bis bei der Anpassung der Formvektoren keine Verbesserung mehr erreicht wird. Dies ist dann der Fall, wenn die Änderung vom Mittelwertvektor des vorherigen Schrittes zum aktuellen Mittelwertvektor Null wird, siehe Abbildung 4.31.

Die Datenvektoren bilden nun eine Verteilung im nd-dimensionalen Raum. Wenn diese Verteilung bekannt ist, kann sie dazu verwendet werden, neue Beispiele zu generieren und für gegebene Formen zu entscheiden, ob es sich um plausible Beispiele für diese Objektklasse handelt. Zunächst wird der Mittelwert und die Kovarianzmatrix der Datenverteilung berechnet:

Abbildung 4.31: *Beispiele für die Ausrichtung der Formdaten vor der PCA. (a) Originale Labeldaten. (b) Mittelwertvektor nach der Ausrichtung. (c) Die Formen liegen nach der Ausrichtung in ihrem gemeinsamen Schwerpunkt und sind so rotiert und skaliert, dass der mittlere quadratische Fehler zur mittleren Form minimal wird.*

$$\overline{\mathbf{x}} = \frac{1}{s}\sum_{i=1}^{s}\mathbf{x}_i \tag{4.33}$$

$$\mathbf{C} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{4.34}$$

Im nächsten Schritt werden durch eine PCA die Eigenwerte und die Eigenvektoren der Kovarianzmatrix $\mathbf{C}$ berechnet. Dabei entsteht die Matrix $\mathbf{T}_S$, die die Eigenvektoren in den Spalten enthält und die Matrix λ_S mit den Eigenwerten auf der Hauptdiagonale. Jede in den Trainingsdaten enthaltene Form kann nun als Linearkombination der Hauptkomponenten ausgedrückt werden. Da die ersten Hauptkomponenten den größten Teil der Varianz in den Trainingsdaten beschreiben, kann durch Weglassen der Eigenvektoren mit kleinen Eigenwerten eine spärliche Beschreibung der Formdaten erreicht werden. Wenn $\mathbf{T}_S$ die m eigenwertgrößten Eigenvektoren enthält $\mathbf{T}_S = (\mathbf{t}_1|\mathbf{t}_2|\ldots|\mathbf{t}_m)$, werden die Formen der Trainingsdaten folgendermaßen approximiert:

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{T}_S \mathbf{b}_S \tag{4.35}$$

wobei $\mathbf{b}_S$ eine Menge von Parametern für das deformierbare Formmodel definiert. Diese Parameter können für ein gegebenes $\mathbf{x}$ wie folgt bestimmt werden:

$$\mathbf{b}_S \approx \mathbf{T}_S^T(\mathbf{x} - \overline{\mathbf{x}}) \tag{4.36}$$

Bei Gleichung 4.36 handelt es sich um eine Unterraumprojektion. Durch die Variation der Elemente von $\mathbf{b}_S$ kann die Form $\mathbf{x}$ variiert werden. Die Varianz von $\mathbf{b}_{S,i}$ ist gegeben durch den $i$ten Eigenwert λ_i. Die Anzahl der zu verwendenden Eigenvektoren sollte so gewählt werden, dass das Modell die Varianz der Trainingsdaten bis zu einem gewissen Umfang erklären kann. Jeder Eigenwert der Kovarianzmatrix entspricht der Varianz der Trainingsdaten in der entsprechenden Richtung. Die gesamte Varianz der Trainingsdaten entspricht der Summe aller Eigenwerte $\sum \lambda_i$. Die m größten Eigenvektoren des Formmodells werden dann so gewählt, dass

$$\frac{\sum_{i=1}^{n-1} \lambda_i}{\sum_{i=1}^{m-1} \lambda_i} \quad \geq \quad r. \tag{4.37}$$

wobei r der Anteil der totalen Varianz ist, der durch das Modell erklärt werden soll. Typischerweise wird mit $r = 98\%$ gearbeitet. Um ein Formmodell an einen Bildinhalt anzupassen, existieren Methoden, die auf den Grauwertgradienten oder den Grauwertverlauf in der lokalen Umgebung eines Label-Punktes zurückgreifen [Cootes and Taylor, 1992] [Cootes and Taylor, 1999]. Duch die Verwendung eines Grauwertmodells ergeben sich aber leistungsfähigere Methoden der Modellanpassung. Um ein Grauwertmodell zu erstellen, müssen verschiedene Formen mittels eines Warpings ineinander überführt werden.

4.6.2.2 Warping

Nach Erstellung des Formmodells ist das Grauwertmodell der nächste Schritt bei der Erstellung eines Active-Appearance-Models. Dazu werden alle Gesichter von ihrer ursprünglichen Form in eine einheitliche Form überführt. Diese einheitliche Form ist der Mittelwertvektor des Formmodells. Der erste Schritt beim Warping ist eine Delaunay-Triangulation auf der Markierungspunkteliste. Die erzeugten Dreiecke können nun mittels eines Textur-Mappings von der Quelle auf das Ziel übertragen werden. Hierbei wird zunächst zu einer bestimmten Position im Ausgangsdreieck dessen Position im Zieldreieck ermittelt. Ausgehend von den Eckpunkten $A = (x_a, y_a)$, $B = (x_b, y_b)$ und $C = (x_c, y_c)$ eines Dreiecks lässt sich jeder Punkt innerhalb des Dreiecks durch die Parameter α und β mit $\alpha \geq 0$, $\beta \geq 0$ und $\alpha + \beta \leq 1$ und dem Zusammenhang

$$\begin{pmatrix} x \\ y \end{pmatrix} = \alpha \begin{pmatrix} x_b - x_a \\ y_b - y_a \end{pmatrix} + \beta \begin{pmatrix} x_c - x_a \\ y_c - y_a \end{pmatrix} \tag{4.38}$$

mit

$$\alpha = \frac{x(y_c - y_a) - y(x_c - x_a)}{(y_c - y_a)(x_b - x_a) - (y_b - y_a)(x_c - x_a)} \tag{4.39}$$

$$\beta = \frac{y(x_b - x_a) - x(y_b - y_a)}{(y_c - y_a)(x_b - x_a) - (y_b - y_a)(x_c - x_a)} \tag{4.40}$$

eindeutig beschreiben. Unter Verwendung von α und β sowie der Eckpunkte von Quell- und Zieldreieck lassen sich somit einander entsprechende Koordinaten in zwei unterschiedlichen Dreiecken ermitteln. Nach Ermittlung der Position wird dem Zielpunkt der Grauwert des Quellpunktes zugewiesen, siehe Abbildung 4.32.

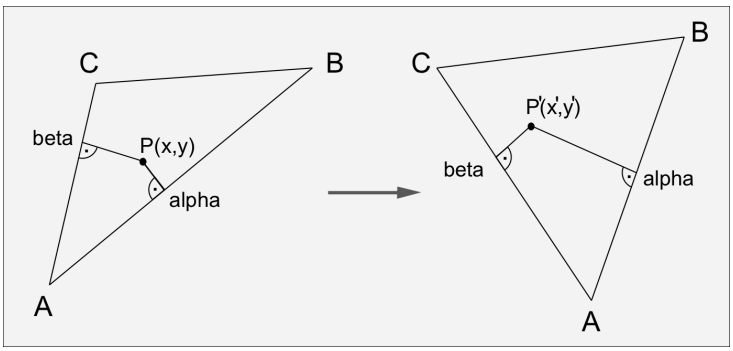


Abbildung 4.32: *Abbildung der Texturdaten aus einem Dreieck des Quellbildes in ein Dreieck des Zielbildes.*

4.6.2.3 Grauwertmodell

Nachdem ein Formmodell für die Gesichter der Trainingsdaten erstellt wurde, wird nun deren Grauwertverteilung modelliert. Dies geschieht auf Grundlage der durch das Warping formnormalisierten Gesichter. Dazu werden alle Gesichter von ihrer ursprünglichen Form auf die bei der Erstellung des Formmodells berechnete mittlere Form gewarpt, siehe Abbildung 4.33. Auf diese Weise entstehen Grauwertgesichter, die in allen Formmerkmalen übereinstimmen und bei denen alle Gesichtsstrukturen an der gleichen Position liegen und die in allen Bildern die gleichen Abmaße haben. Einige Beispiele sind in Abbildung 4.34 dargestellt.

Ähnlich dem Vorgehen beim Formmodell muss auch für das Grauwertmodell ein einheitlicher Grauwertbereich geschaffen werden. Zu diesem Zweck wird wiederum ein Bezugsbild verwendet, an das sämtliche Grauwertgesichter angepasst werden. Hierzu werden ein Skalierungsfaktor α und ein Abstand β zum Mittelwert $\overline{\mathbf{g}}$ bestimmt. Da es sich beim Bezugsbild um das mittlere Grauwertgesicht handelt, handelt es sich auch hier um einen rekursiven Prozess. Der Grauwertvektor $\mathbf{g}'$, der das Gesicht repräsentiert, ergibt sich aus:

$$\mathbf{g}'_i = (\mathbf{g}_i - \beta\mathbf{1})\alpha \tag{4.41}$$

mit dem aktuell betrachteten Grauwertvektor $\mathbf{g}_i$, sowie

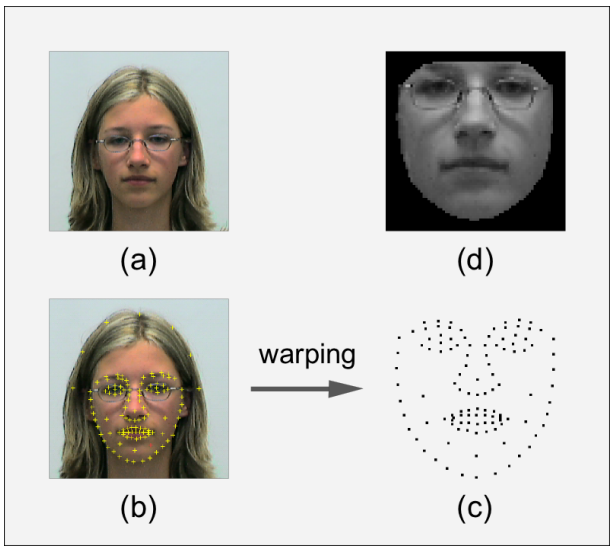


Abbildung 4.33: *Warping eines Grauwertgesichtes auf die mittlere Form. (a) Ursprüngliche Form des Gesichts. (b) Labelpunkte für das Gesicht. (c) Labelpunkte der bei der Erstellung des Formmodells berechneten mittleren Form. (d) Grauwertgesicht nach dem Warping.*

Abbildung 4.34: *Warping der Grauwertgesichter auf die mittlere Form. (a) Gesichter aus der Trainingsdatenbank. (b) Auf die mittlere Form gewarpte Grauwertgesichter.*

$$\alpha = \overline{\mathbf{g}}\mathbf{g}_i \tag{4.42}$$

$$\beta = \mathbf{g}_i/n \tag{4.43}$$

wobei n die Anzahl der Pixel des Grauwertgesichtes ist. Auf diese Weise entstehen grauwertnormierte Gesichter, die nun zur statistischen Analyse verwendet werden können. Dazu wird genau wie beim Formmodell eine Hauptkomponentenanalyse durchgeführt, wobei die Matrix $\mathbf{T}_G$ die m eigenwertgrößten Eigenvektoren $\mathbf{T}_G = (\mathbf{t}_1|\mathbf{t}_2|\ldots|\mathbf{t}_m)$ und λ_G die zugehörigen Eigenwerte enthält. Durch Projektion eines Grauwertbildes $\mathbf{g}$ auf die Eigenvektoren in $\mathbf{T}_G$ werden die Grauwertparameter $\mathbf{b}_G$ erzeugt.

$$\mathbf{b}_G \approx \mathbf{T}_G^T(\mathbf{g} - \overline{\mathbf{g}}) \tag{4.44}$$

Die Grauwertverläufe der Trainingsdaten können dann folgendermaßen approximiert werden:

$$\mathbf{g} \approx \overline{\mathbf{g}} + \mathbf{T}_G \mathbf{b}_G \tag{4.45}$$

Auch hier wird die Anzahl m der zu verwendenden Parameter so gewählt, dass ein bestimmter Prozentsatz der Variation in den Daten durch das Modell erklärt werden kann, siehe Gleichung 4.37.

4.6.2.4 Active-Appearance-Model

Nach der Erzeugung des Form- und des Grauwertmodells besteht der letzte Schritt zur Erzeugung des Active-Appearance-Models in der Kombination der beiden Modelle und der Erzeugung der Appearance-Parameter. Dazu wird in [Cootes et al., 1998] eine weitere statistische Analyse empfohlen, wodurch ein *kombiniertes Appearance-Model* (CAM) entsteht. In [Matthews and Baker, 2003] wird dagegen ein *unabhängiges Appearance-Model* (IAM) verwendet, bei dem auf die letzte Hauptkomponentenanalyse verzichtet wird. Beide Vorgehensweisen werden im Folgenden kurz vorgestellt.

Kombiniertes Appearance-Model Beim kombinierten Active-Appearance-Model wird eine weitere statistische Analyse der zusammengesetzten Form- und Grauwertmodelle durchgeführt. Hierzu muss jedoch eine Wichtung eines Datensatzes durchgeführt werden, damit im Ergebnis keine Dominanz von Form- oder Grauwertdaten entsteht, falls die Datensätze nicht im gleichen Wertebereich liegen. Dies wird erreicht, indem das Verhältnis der totalen Varianz der Formparameter $\mathbf{b}_S$ zur totalen Varianz der Grauwertparameter $\mathbf{b}_G$ ermittelt wird. Das Gewicht w für die Formparameter wird wie folgt berechnet:

$$w = \frac{\sum_{i=1}^{m_G} \lambda_{G,i}}{\sum_{i=1}^{m_S} \lambda_{S,i}} \tag{4.46}$$

Unter Verwendung von w ergibt sich der Appearance-Vektor $\mathbf{b}$ für jedes Trainingsbild wie folgt:

$$\mathbf{b} = \begin{pmatrix} w\mathbf{b}_S \\ \mathbf{b}_G \end{pmatrix} \tag{4.47}$$

Nach der Wichtung der Appearance-Vektoren wird deren Mittelwert $\overline{\mathbf{b}}$ bestimmt und schließlich die Hauptkomponentenanalyse durchgeführt. Mittels Projektion der Appearance-Vektoren $\mathbf{b}$ auf die dabei berechneten Eigenvektoren $\mathbf{T}$ lassen sich die Appearance-Parameter $\mathbf{c}$ berechnen:

$$\mathbf{c} = \mathbf{T}^{\mathbf{T}}(\mathbf{b} - \overline{\mathbf{b}}) \tag{4.48}$$

Da die Appearance-Parameter $\mathbf{c}$ durch eine Hauptkomponentenanalyse des zusammengesetzten Appearance-Vektors entstanden sind, kodieren sie Zusammenhänge zwischen Formveränderung

und damit einhergehender Grauwertveränderung. So sollte sich beispielsweise die Grauwertveränderung des Gesichtes beim Drehen des Kopfes aus oder in die Lichtrichtung erkennen lassen. Analog zum Form- und Grauwertmodell kann auch auf den Appearance-Parametern eine Datenreduktion durchgeführt werden.

Unabhängiges Appearance-Model Im Gegensatz zum CAM werden die Appearance-Vektoren keiner weiteren Hauptkomponentenanalyse unterzogen, es erfolgt also keine weitere Datenreduktion. Somit entsprechen die Appearance-Vektoren im IAM den Appearance-Parametern. Die resultierenden Appearance-Parameter $\mathbf{c}$ ergeben sich somit aus:

$$\mathbf{c} = \begin{pmatrix} \mathbf{b}_S \\ \mathbf{b}_G \end{pmatrix} \tag{4.49}$$

4.6.3 Modellanwendung

Die Anwendung eines Appearance-Models besteht aus einem iterativen Prozess, bei dem die Appearance-Parameter so adaptiert werden, dass die Differenz zwischen Eingangsgesicht und aus dem Modell synthetisierten Gesicht minimal wird. Zunächst soll erläutert werden, wie aus den Appearance-Parametern ein Gesicht synthetisiert werden kann.

4.6.3.1 Synthese eines Gesichts

Aus einem Satz von Appearance-Parametern $\mathbf{c}$ kann ein Grauwertgesicht synthetisiert werden. Die Appearance-Parameter $\mathbf{c}$ entsprechen beim IAM bereits dem Appearance-Vektor, beim CAM wird dieser mittels Rückprojektion auf die Eigenvektoren erzeugt. Der Appearance-Vektor $\mathbf{b}$ wird in Formparameter $\mathbf{b}_S$ und den Grauwertparameter $\mathbf{b}_G$ zerlegt. Nach Gleichung 4.45 kann das formnormierte Grauwertgesicht $\mathbf{g}$ synthetisiert werden. Entsprechend wird nach Gleichung 4.35 die Markierungspunktliste $\mathbf{x}$ erzeugt, die die Form des synthetisierten Gesichts beschreibt. Abschließend wird das formnormierte Grauwertgesicht auf die ermittelte Gesichtsform gewarpt. Abbildung 4.35 beschreibt den Prozess der Synthese eines Gesichtes.

4.6.3.2 Modellanpassung

Wenn ein Gesicht in einem Bild analysiert werden soll, müssen die Parameter des Apearance-Models so angepasst werden, dass das vom Modell synthetisierte Gesicht mit dem gegebenen Gesicht möglichst gut übereinstimmt. Abbildung 4.36 verdeutlicht den Vorgang der Modellanpassung, bei dem die Appearance-Parameter und die Transformationsparameter $\mathbf{t} = t_x, t_y, a$ und b für Translation, Skalierung und Rotation des synthetisierten Gesichtes ermittelt werden.

Zu Beginn sind sämtliche Appearance-Parameter Null. Nach Gleichung 4.35 und 4.45 entstehen also jeweils der Mittelwertvektor der Markierungspunktliste und des Grauwertvektors und damit das mittlere Gesicht. Somit existiert immer der gleiche, wohldefinierte Ausgangspunkt für

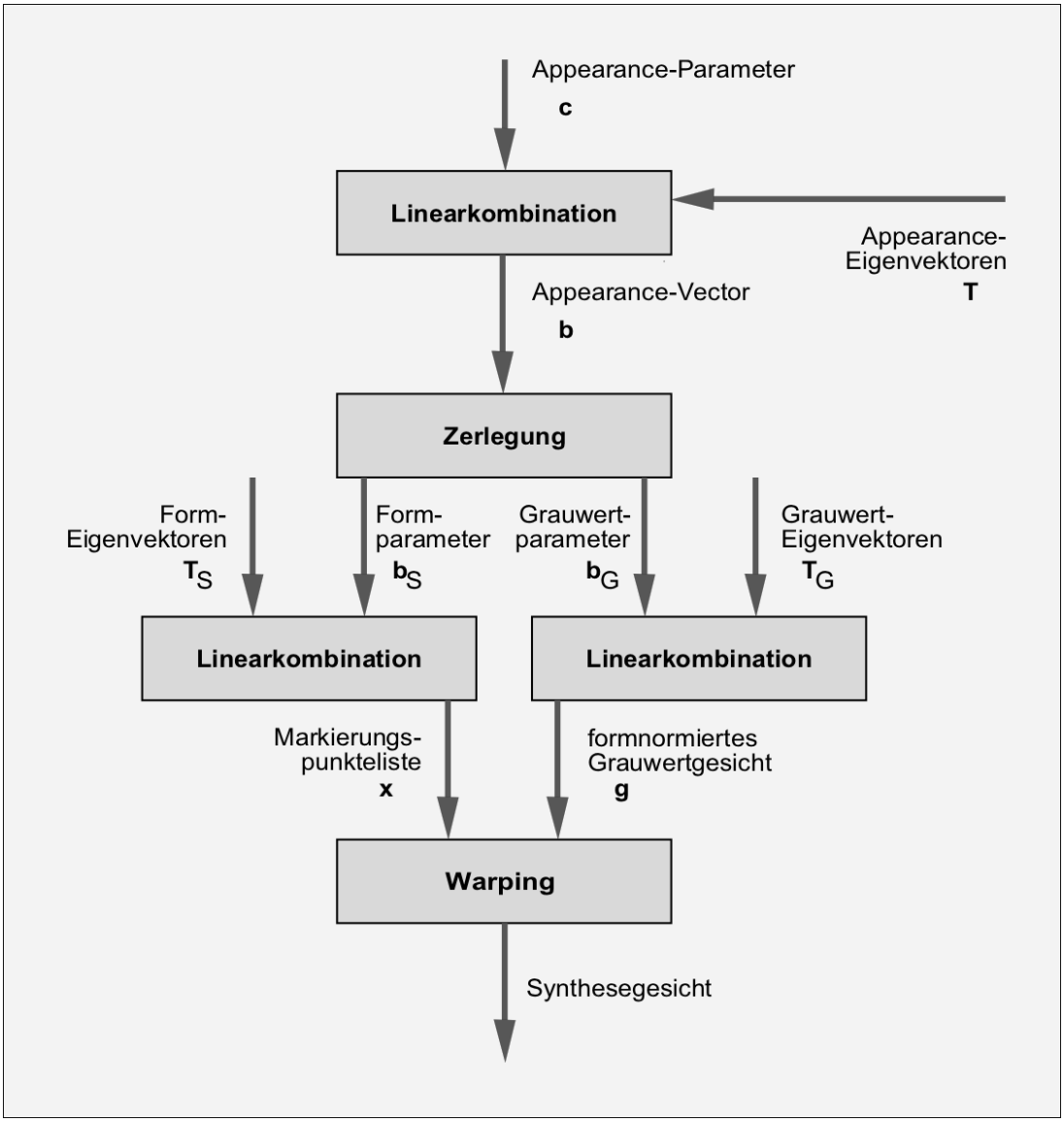


Abbildung 4.35: *Ablauf der Synthese eines Gesichtes aus Appearance-Parametern.*

den Suchprozess. Die Erzeugung des Differenzbildes findet im formnormierten Raum statt. Die Appearance-Parameter werden in die Markierungspunktliste $\mathbf{x}$ und das formnormierte Grauwertgesicht $\mathbf{g}^m$ überführt. Unter Verwendung von $\mathbf{x}$ sowie den geschätzten Transformationsparametern wird das Grauwertgesicht $\mathbf{g}^i$ aus dem Eingabebild von seiner geschätzten Position und Form auf die mittlere Form gewarpt und der beschriebenen Grauwertnormierung unterzogen. Die Differenz zwischen diesem formnormierten Eingangsgesicht $\mathbf{g}^i$ und dem, aus den Appearance-Parametern erzeugten, formnormierten Synthesegesicht $\mathbf{g}^m$ bildet das Residuum $\mathbf{r}$:

$$\mathbf{r}(\mathbf{p}) = \mathbf{g}^i + \mathbf{g}^m \tag{4.50}$$

mit $\mathbf{p} = (\mathbf{c}, \mathbf{t})^T$. Hierauf wird eine Taylor-Expansion erster Ordnung durchgeführt:

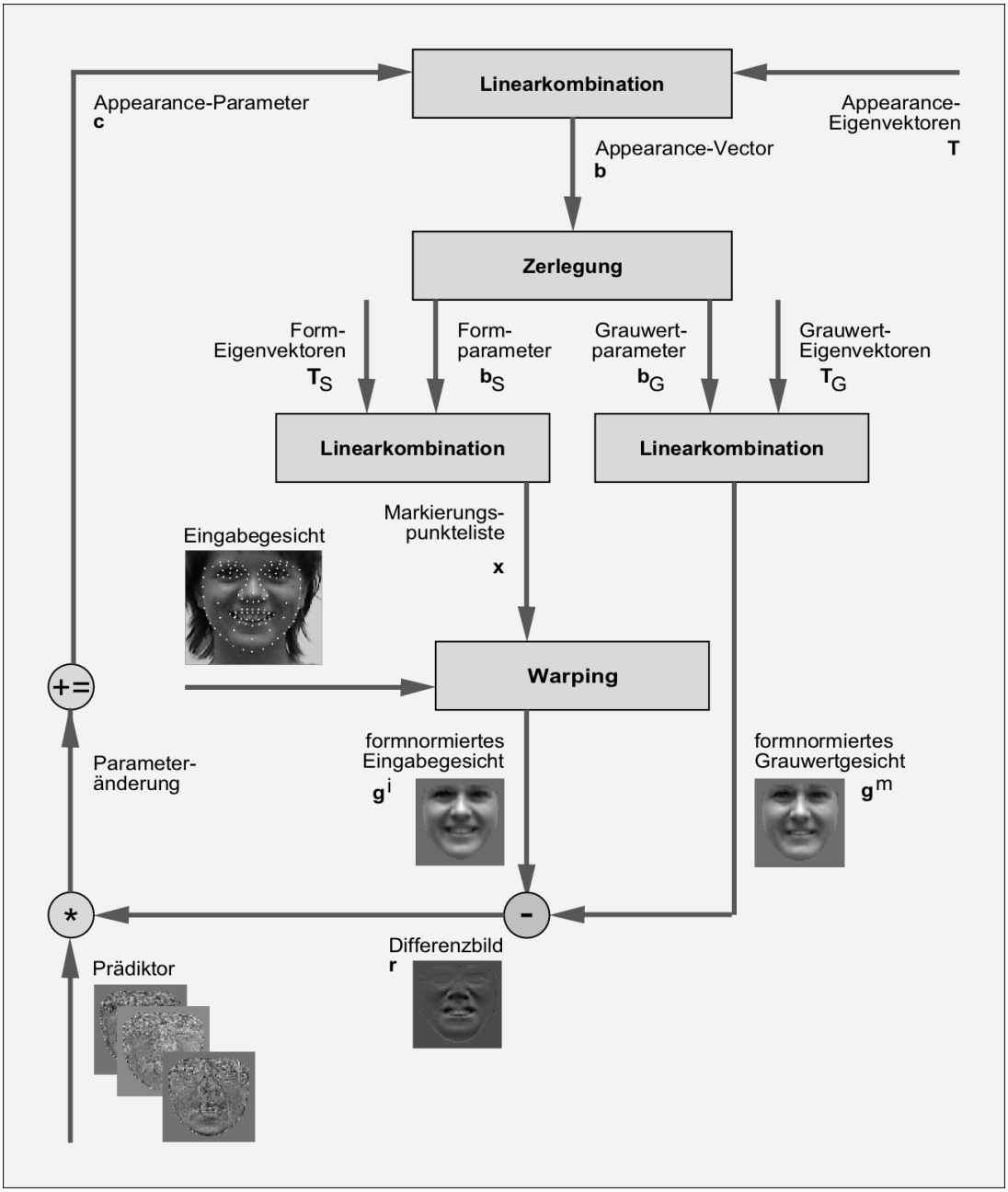


Abbildung 4.36: *Ablauf der Modellanpassung. Aus den aktuellen Appearance-Parametern werden die Form- und Grauwertparamerer ermittelt. Aus den Formparametern wird die Markierungspunkteliste bestimmt, anhand derer eine formnormierte Darstellung des Eingabegesichtes erzeugt wird. Aus den Grauwertparametern wird das formnormierte Synthesegesicht synthetisiert. Mit der Differenz zwischen formnormiertem Eingabegesicht und formnormiertem Synthesegesicht und dem Prädiktor wird eine Parameteränderung für jeden Appearance-Parameter ermittelt. Die Modellanpassung wird mit den neuen Appearance-Parametern wiederholt und solange fortgesetzt, bis die Parameteränderungen Null sind.*

$$\mathbf{r}(\mathbf{p}+\delta\mathbf{p}) = \mathbf{r}(\mathbf{p}) + \frac{\partial \mathbf{r}}{\partial \mathbf{p}}\delta\mathbf{p} \tag{4.51}$$

$\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ ist dabei die Matrix partieller Ableitungen, also ein Vektor von Grauwertänderungen in Abhängigkeit der Änderung der Appearance-Parameter. Ausgehend von einem aktuellen Parametervektor und dem zugehörigen Differenzbild $\mathbf{r}(\mathbf{p})$ soll $\delta\mathbf{p}$ so gewählt werden, dass $|\mathbf{r}(\mathbf{p}+\delta\mathbf{p})|^2$, d.h. die Energie im Differenzbild bei geänderten Parametern, minimiert wird. Durch Gleichsetzen von Gleichung 4.51 mit Null erhält man:

$$-\mathbf{r}(\mathbf{p}) = \frac{\partial \mathbf{r}}{\partial \mathbf{p}}\delta\mathbf{p} \tag{4.52}$$

und durch Umstellen:

$$\delta\mathbf{p} = -\mathbf{R}\mathbf{r}(\mathbf{p}) \quad \text{mit} \quad \mathbf{R} = \left(\frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T \frac{\partial \mathbf{r}}{\partial \mathbf{p}}\right)^{-1} \frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T \tag{4.53}$$

Normalerweise müsste $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ und damit auch die Pseudoinverse in jedem Iterationsschritt neu berechnet werden. Es kann aber davon ausgegangen werden, dass Parameteränderungen bei einem Gesicht aus dem Trainingsdatensatz und bei einem zu modellierenden Gesicht eine ähnliche Wirkung haben, so dass $\mathbf{R}$ bereits während des Trainings geschätzt werden kann. Dazu wird $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ aus den Trainingsdaten durch numerische Differentiation geschätzt.

Um die Prediktormatrix $\mathbf{R}$ zu berechnen, wird systematisch die Änderung genau eines Parameters p_j zu einem Zeitpunkt und deren Auswirkung auf das Differenzbild betrachtet. Innerhalb eines bestimmten Abweichungsintervalls (typischerweise -0.5 bis $+0.5$ Standardabweichungen) werden mehrere dieser Grauwertdifferenzen generiert. Durch anschliessendes Mitteln über die Anzahl der verwendeten Beispiele wird der Einfluss des Parameters auf das Differenzbild in Form eines Vektors erzeugt. Es handelt sich hierbei also um eine Approximation der partiellen Ableitungen durch einen Differenzenquotient:

$$\frac{d\mathbf{r}_i}{d\mathbf{p}_j} \approx \frac{1}{k}\sum_k \frac{1}{\delta p_{jk}} \left(\mathbf{r}_i(\mathbf{p} + \delta\mathbf{p}_{jk}) - \mathbf{r}_i(\mathbf{p})\right) \tag{4.54}$$

Hierbei bezeichnet δp_{jk} das $j$te Element des Vektors $\delta\mathbf{p}_k$, bei dem alle anderen Werte Null sind. $\mathbf{r}_i(\mathbf{p} + \delta\mathbf{p}_{jk}) - \mathbf{r}_i(\mathbf{p})$ ist die Grauwertdifferenz zwischen Bildern, die aus dem ursprünglichen Appearance-Vektor $\mathbf{r}_i(\mathbf{p})$ bzw. dem Appearance-Vektor mit Parameteränderung $\mathbf{r}_i(\mathbf{p} + \delta\mathbf{p}_{jk})$ erzeugt wurden. Bei der Betrachtung der Veränderung eines jeden Parameters in $\mathbf{p}$ entsteht somit ein Residuumvektor in $\mathbf{R}$, der die Grauwertdifferenz für jedes Pixel des Grauwertbildes enthält. $\frac{d\mathbf{r}_i}{d\mathbf{p}_j}$ ist ein Vektor von der Größe eines Bildes. Mit allen partiellen Ableitungen kann $\mathbf{R}$ nach Gleichung 4.53 berechnet werden. Die erzeugte Prediktormatrix enthält also Informationen darüber, wie sich das Differenzbild durch Veränderung der einzelnen Parameter beeinflussen lässt. Nachdem das Suchverfahren das Differenzbild zwischen formnormiertem Eingangsgesicht und formnormiertem Synthesegesicht erzeugt hat, werden durch Multiplikation jeder Zeile r_i der Prediktormatrix mit dem Differenzvektor die nötigen Parameteränderungen errechnet, die nötig sind, um das Synthesegesicht weiter zu verbessern.

4.6.4 Klassifikation

Für die Klassifikation der Appearance-Parameter hinsichtlich Alter, Geschlecht, Gesichtsausdruck und Identität wurden eine Reihe von Klassifikatoren vergleichend untersucht. Dabei handelt es sich wie bei der ICA um Nearest-Neighbor-Klassifikatoren, Multilayer-Perceptrons, Radial-Basis-Function-Netze und verschiedene Learning-Vector-Quantifier.

4.6.5 Voruntersuchungen

Bildgröße Durch den Aufbau von Active-Appearance-Models in verschiedenen Auflösungsstufen zeigte sich, dass die Auflösung einen relativ geringen Einfluss auf die Anzahl der zu verwendenden Grauwertparameter hat. Wie man in Tabelle 4.3 erkennen kann, bringt eine Steigerung der Auflösung von 25×25 auf 40×40 einen Gewinn an Information, da die Gesichtsbilder in der kleinsten Auflösung viele Details vermissen lassen. Dieser Effekt ist allerdings bei einer weiteren Vergrößerung des Modells nicht mehr so deutlich. Eine Erhöhung der Auflösung ist also nicht automatisch eine Garantie für bessere Syntheseergebnisse.

Auflösung	25×25	40×40	73×73	149×149
Grauwertparameter	91	164	187	192

Tabelle 4.3: *Einfluss der Auflösung auf die Anzahl der Grauwertparameter.*

Modellumfang Desweiteren hat sich gezeigt, dass das Hinzufügen zusätzlicher Grauwertinformationen sogar die Darstellungsfähigkeit des Modells einschränken kann. Durch Hinzufügen des Haaransatzes zu den Gesichtsbildern reduzierte sich die Anzahl der nötigen Grauwertparameter bei einer Bildgrösse von 86×69 auf lediglich 139, also auf weniger Parameter als das Grauwertmodell mit 40×40 Pixeln auf dem selben Datensatz umfasste. Daraus resultiert, dass sich auch die Syntheseergebnisse verschlechtern. Die Form des Gesichts wird zwar weiterhin richtig erkannt, obwohl auch die Zahl der Formparameter von 51 auf 47 sank, aber die Synthesebilder sind den Originalbildern weitaus unähnlicher als beispielsweise bei Verwendung des Modells der Größe 40×40. Grund für die starke Reduzierung der Grauwertparameter ist der hohe Varianzanteil der Kopfbehaarung. Der Trainingsdatensatz zeigt viele unterschiedliche Formen und Farben von Kopfbehaarung, die durch das Modell nachgebildet werden. Allerdings leidet darunter die Qualität der Nachbildung des Gesichts. Eine Möglichkeit, ein genaueres Modell zu erstellen, besteht darin, die Schwelle r, die den abgebildeten Varianzanteil festlegt, mit mehr als 98% festzulegen, siehe Gleichung 4.37. Hierbei steigt die Anzahl zu verwendender Parameter jedoch relativ schnell im Vergleich zum Zuwachs an Information, da die hinzukommenden Parameter nur eine sehr geringe Varianz haben.

Synthesegeschwindigkeit Das Diagramm in Abbildung 4.37(a) zeigt die durchschnittliche Schrittdauer in Sekunden und Abbildung 4.37(b) die Anzahl der nötigen Schritte bis zur Konvergenz. Diese Eigenschaften wurden für ein Modell der Größe 73×73 und ein Modell der Größe 149×149 ermittelt. Die Unterscheidung nach bekannt und unbekannt bezieht sich darauf, ob das Bild für die Erzeugung des Modells verwendet wurde oder nicht. Bei einem Modell mit doppelter Auflösung ist die durchschnittliche Schrittzahl bis zur Konvergenz der Anpassung zwar etwas geringer, dafür benötigen die einzelnen Schritte aber wesentlich mehr Zeit.

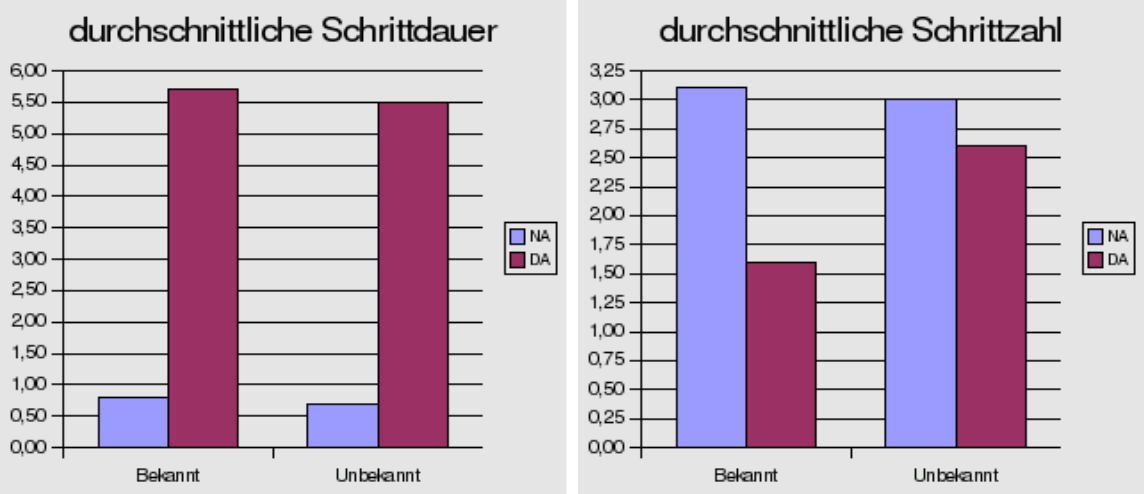
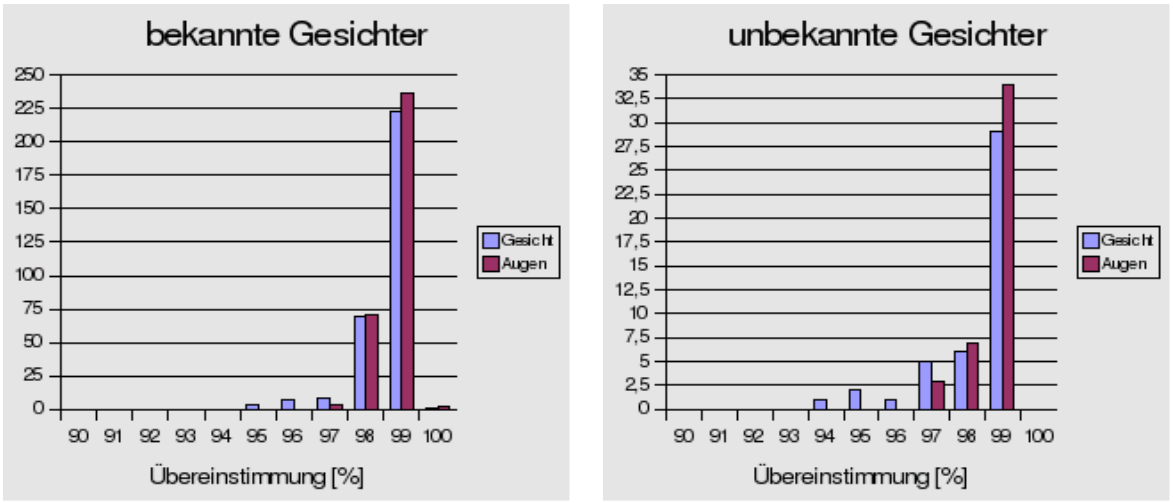


Abbildung 4.37: *(a) Durchschnittliche Schrittdauer, die ein Modell für einen Suchschritt auf einem bekannten bzw. unbekannten Bild benötigte. (b) Durchschnittliche Schrittzahl, die von den Modellen benötigt wurde, bis die Konvergenz auf einem bekannten oder unbekannten Bild erreicht wurde. NA ist die normale Auflösung von* 73×73 *und DA die doppelte Auflösung von* 149×149 *Pixeln.*

Synthesegenauigkeit Zur Überprüfung der Qualtität der erzeugten Synthesebilder wurde ein Trainingsdatensatz mit 313 Bildern bekannter sowie ein Datensatz mit 44 Bildern unbekannter Personen verwendet. Zur Bestimmung der Synthesequalität wurde das betragsnormierte Skalarprodukt zwischen dem Grauwertvektor des Gesichtes im Eingangsbild und dem Grauwertvektor des Synthesebildes gebildet. Das Skalarprodukt liefert genau dann Eins, wenn Eingangsbild und Synthesebild gleich sind. Für den Test wurden einmal die Augenpositionen durch die Labelpunkte vorgegeben und einmal nur die Grobschätzung des Gesichtsdetektors verwendet.

Wie Abbildung 4.38 zeigt, ist die Nachbildungsgenauigkeit des Modells generell sehr hoch. Bemerkenswert ist, dass die Genauigkeit auch bei der exakten Vorgabe der Augenpositionen nicht wesentlich ansteigt. Das lässt darauf schlussfolgern, dass das AAM auch dann in der Lage ist, sich an ein Eingabegesicht anzupassen, wenn die Initialposition nur grob vorgegeben ist, wie hier durch den Gesichtsdetektor.



(a) (b)

Abbildung 4.38: *Übereinstimmung von Originalbild und synthetisiertem Bild in %. (a) Für bekannte Gesichter ohne Vorgabe der Augenpositionen und mit Vorgabe der Augenposition zur Ermittlung der Startschätzung. (b) Für unbekannte Gesichter ohne Vorgabe der Augenpositionen und mit Vorgabe der Augenposition zur Ermittlung der Startschätzung.*

4.7 Vergleichende Untersuchungen

4.7.1 Trainingsablauf

Aus den beiden Datensätzen für Alter, Geschlecht und Identität bzw. für Mimik wurde jeweils ein Trainings-, ein Validierungs- und ein Testdatensatz erstellt. Der Trainingsdatensatz wird zum Aufbau der General-Face-Knowledge beim EGM bzw. zum Training der Klassifikatoren bei der ICA und den AAMs verwendet. Der Validierungsdatensatz dient dazu, eine optimale Trainingsdauer zu bestimmen und eine Überanpassung der Klassifikatoren zu verhindern. Im Anschluss an das Training werden mit Hilfe des Testdatensatzes die Generalisierungseigenschaften der Klassifikatoren bestimmt. Das Training der Merkmalsextraktion (ICA, AAM) wurde jeweils auf dem Gesamtdatensatz durchgeführt.

Für die Erstellung dieser Testdatensätze wurde eine *Leave-n-out*-Strategie angewendet, d.h. es wurde ein Teil der Daten aus dem Datensatz entnommen, der als Testmenge dient, ein Teil als Validierungsmenge und der Rest als Trainingsmenge. In diesen drei Teildatensätzen befanden sich immer Bilder unterschiedlicher Personen, wobei jede Person genau einmal in die Testmenge aufgenommen wurde. Dabei enthielten sowohl der Trainings-, der Test als auch der Validierungdatensatz immer alle Klassen der jeweiligen Klassifikationsaufgabe, das heißt, die fünf Altersklassen, die beiden Geschlechter, die sieben Basisemotionen bzw. die 70 Identitäten.

Tabelle 4.4 zeigt die jeweilige Aufteilung der Daten für die verschiedenen Erkennungsaufgaben. Für jede Zusammenstellung wurde ein Klassifikator trainiert und getestet und am Ende wurden die Erkennungsraten über die Anzahl der Durchläufe gemittelt. Durch diese Vorgehensweise wird die verhältnismäßig kleine Anzahl an Daten optimal ausgenutzt und es wird sichergestellt, dass die Ermittlung der Erkennungsraten bei der Generalisierung nicht mit den selben Personen erfolgt, die im Training verwendet wurden.

Ein Beispiel: bei der Geschlechtsschätzung auf dem Neutraldatensatz mit 70 Personen werden in jedem Durchlauf eine männliche und eine weibliche Person in den Testdatensatz aufgenommen. Da jeweils zwei Personen den Testdatensatz bilden, und jede Person genau einmal in den Testdatensatz aufgenommen wird, werden 35 Trainingsdurchläufe ausgeführt.

Aufgabe	**Personen**	**Bilder**	**Durchläufe**
Geschlecht	60/8/2	7/7/7	35
Alter	60/5/5	7/7/7	14
Mimik	25/4/1	7/7/7	30
Identität	70/70/70	3/2/2	1

Tabelle 4.4: *Aufteilung der Personen eines Datensatzes und der Bilder einer Person auf die Teildatensätze für* ***Training/Validierung/Test****. Der Datensatz mit neutralen Gesichtsausdrücken für die Klassifikation von Alter, Geschlecht und Identität enthielt 70 Personen mit jeweils sieben Bildern, der Datensatz mit Gesichtsausdrücken nur 30 Personen mit je sieben Bildern. Für die Klassifikation von Geschlecht und Alter werden die 70 Personen auf die drei Datensätze aufgeteilt, wobei alle sieben Bilder einer Person einem der drei Datensätze zugeordnet werden. Bei der Klassifikation von Gesichtsausdrücken werden die 30 Personen des Mimikdatensatzes auf die gleiche Weise auf die drei Datensätze aufgeteilt. Hier existieren für jede Person sieben Bilder, die je eine Basisemotion darstellen. Durch diese Aufteilung in Trainings-, Validierungs- und Testdatensatz wird sichergestellt, dass die Erkennungsraten, die ausschließlich auf den Testdaten ermittelt werden, nicht dadurch verfälscht werden, dass gleiche Personen im Trainings- und Testdatensatz enthalten sind. Bei der Personenidentifikation teilen sich die 7 Bilder einer Person auf die drei Datensätze auf. Drei Bilder sind im Trainings-, zwei im Validierungs- und zwei im Testdatensatz. Damit ist bei der Personenidentifikation nur ein Trainingsdurchlauf möglich.*

4.7.2 Ergebnisse

In diesem Abschnitt werden die Erkennungsraten für die Geschlechts-, Alters- und Mimikschätzung und für die Personenidentifikation für die verschiedenen Verfahren der Merkmalsextraktion und Klassifikation gegenübergestellt. Auf der Grundlage von Untersuchungen in [Eckardt, 2005] und den Voruntersuchungen aus Abschnitt 4.5.5 werden mit Ausnahme der Personenidentifikation für die ICA nur die Ergebnisse im Bildraum präsentiert, da die Verwendung des Pixelraums im allgemeinen schlechtere Ergebnisse liefert. Genauso wird hier auf die Darstellung der unabhängigen Active-Appearance-Models verzichtet, da sie keine besseren Ergebnisse liefern als die kombinierten Active-Appearance-Models [Trapp, 2005]. Bei der Geschlechts-, Alters- und Identitätsschätzung wurden neben den Tests auf dem Neutraldatensatz ebenfalls Tests auf der Vereinigung von Neutral- und Mimikdatensatz durchgeführt, um die Abhängigkeit der Erkennungsraten vom Gesichtsausdruck zu bestimmen.

Die Detektion der Augen erfolgte bei den Tests, falls nicht anders angegeben, immer automatisch durch den AdaBoost-Detektor aus Abschnitt 3.3.

Die Klassifikatoren wurden dabei wie folgt parametriert. Die Ausgabeschicht der Netze besitzt entsprechend der Erkennungsaufgabe k Neuronen (Geschlecht: $k = 2$, Alter: $k = 5$, Mimik: $k = 7$ und Identität $k = 70$). Für den Teach-Vektor wird eine 1-aus-N-Kodierung verwendet. Das MLP besitzt eine Hidden-Schicht mit Skalarproduktaktivierung und sigmoider Ausgabefunktion mit 40 Neuronen. Das RBF verwendet in der Hidden-Schicht 40 Neuronen mit radialsymmetrischer Aktivierungsfunktion, deren Zentren durch ein LVQ-Training platziert werden und später durch das Backpropagation-Training nicht mehr beeinflusst werden. Bei den LVQ-Netzwerken wurden typischerweise 20 Neuronen pro Klasse verwendet. Für die RBF- und LVQ-Klassifikatoren werden die Eingabedaten normiert. Beschreibungen zu den verwendeten Klassifikatoren und Lernverfahren können [Zell, 1994] entnommen werden.

4.7.2.1 Geschlechtsschätzung

Bei der Geschlechtsschätzung befinden sich immer zwei Personen im Testdatensatz, eine männliche und eine weibliche. Damit ergeben sich bei 70 Personen insgesamt $70/2 = 35$ Trainingsdurchläufe. Jeweils acht Personen bilden den Validierungsdatensatz und die restlichen 60 den Trainingsdatensatz. Die schlechteste Erkennungsrate (die durch Raten des Geschlechtes erreicht werden kann) liegt bei $0,5$. Abbildung 4.39 zeigt die Erkennungsraten von EGM, ICA und AAM bei der Geschlechtsschätzung für unterschiedliche Klassifikatoren auf dem Neutral- und auf dem Gesamtdatensatz.

Bei der Geschlechtsschätzung auf dem Neutraldatensatz wurden mit 93% die besten Erkennungsraten mit einem Active-Appearance-Model in Kombination mit einem MLP erreicht. Bei der ICA lag die beste Erkennungsrate bei Verwendung eines GLVQ bei 83% und beim EGM nur noch bei 74%. Auffallend ist, dass das MLP mit der ICA besonders schlecht abschneidet, während es mit den AAMs die besten Ergebnisse liefert. Eine Diskussion dieser Zusammenhänge erfolgt im Abschnitt 4.7.2.6.

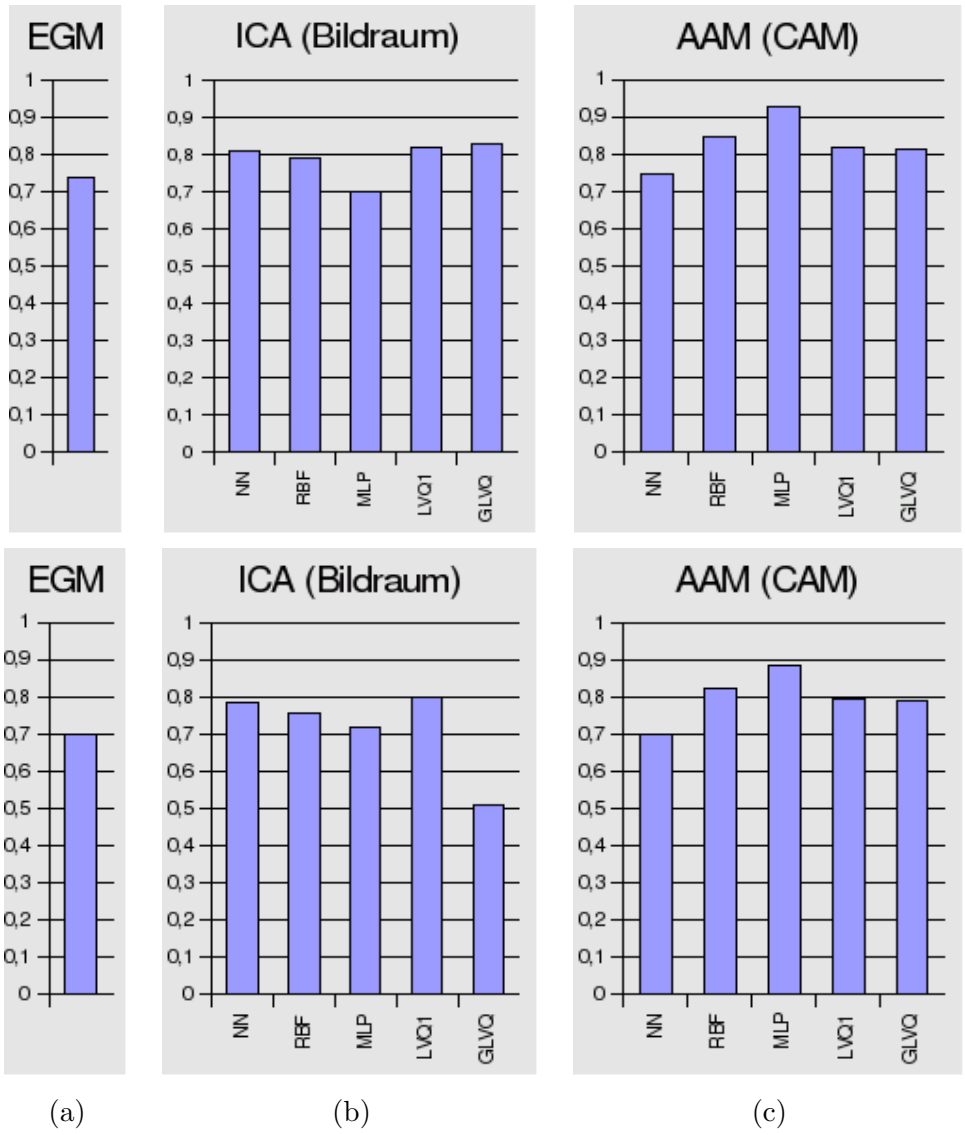
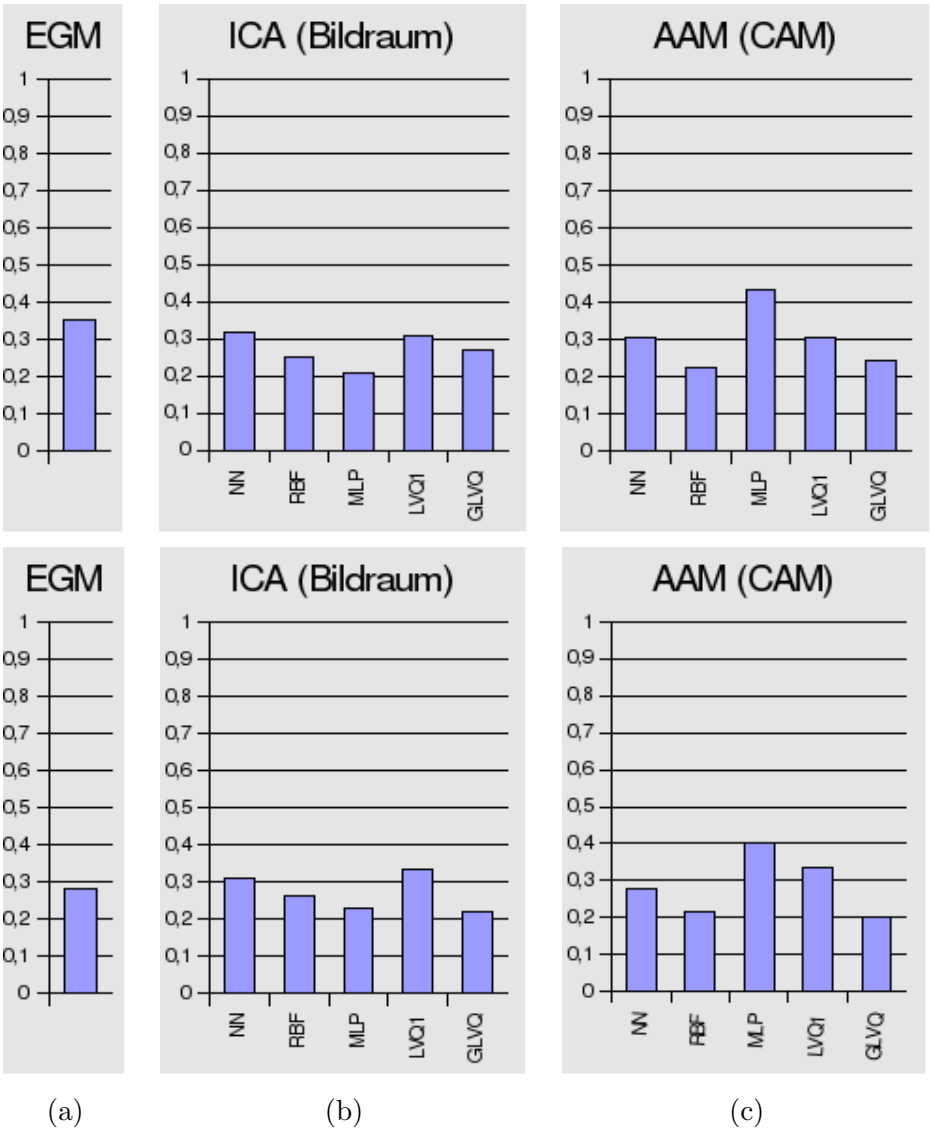


Abbildung 4.39: *Erkennungsraten bei der* ***Geschlechtsschätzung*** *(a) EGM (b) ICA im Bildraum (unreduziert) und (c) AAM (CAM). Obere Reihe: Erkennungsraten auf dem Datensatz mit neutralen Gesichtsausdrücken. Untere Reihe: Erkennungsraten auf dem Gesamtdatensatz mit Neutral- und Mimikdaten. Bei den AAMs wird die beste Erkennungsrate mit dem MLP erreicht, während bei der ICA die NN- und LVQ-Klassifikatoren am besten abschneiden. Das EGM liefert mit 74% eine relativ schlechte Erkennungsrate. Die Erkennungsraten auf dem Gesamtdatensatz mit Mimikdaten liegen meistens geringfügig unter den auf dem Neutraldatensatz erreichten Erkennungsraten.*

4.7.2.2 Altersschätzung

Bei der Altersschätzung befinden sich immer fünf Personen im Testdatensatz, eine aus jeder der fünf Altersgruppen. Damit ergeben sich bei 70 Personen $70/5 = 14$ Trainingsdurchläufe. Ebenfalls fünf Personen bilden den Validierungsdatensatz und die restlichen 60 den Trainingsdatensatz. Die schlechteste Erkennungsrate (die durch Raten des Alters bei fünf möglichen Klassen erreicht werden kann) liegt bei $0,2$. Abbildung 4.40 zeigt die Erkennungsraten von EGM, ICA und AAM bei der Alterschätzung für unterschiedliche Klassifikatoren auf dem Neutral- und auf dem Gesamtdatensatz.



(a) (b) (c)

Abbildung 4.40: *Erkennungsraten bei der* ***Altersschätzung****. (a) EGM (b) ICA im Bildraum und (c) AAM (CAM). Obere Reihe: Erkennungsraten auf dem Datensatz mit neutralen Gesichtsausdrücken. Untere Reihe: Erkennungsraten auf dem Gesamtdatensatz mit Neutral- und Mimikdaten. Auch hier scheiden die AAMs in Kombination mit einem MLP am besten ab.*

Wie bei der Geschlechtsschätzung schneiden auch bei der Altersschätzung die AAMs in Kombination mit einem MLP am besten ab. Bei Verwendung des Gesamtdatensatzes inklusive der

Mimikdaten fallen die Erkennungsraten wie zu erwarten bei den meisten Verfahren und Klassifikatoren geringfügig ab. Bei der Altersschätzung mit fünf Altersgruppen sind die Erkennungsraten bei allen drei Verfahren sehr schlecht. Auf dem Neutraldatensatz wurden mit dem EGM 35%, mit der ICA und einem Nearest-Neighbor-Klassifikator 32% und mit einem AAM und einem MLP 43% erreicht. Um diese Ergebnisse besser interpretieren zu können, werden in Tabelle 4.5 die Konfusionsmatrizen für die drei Verfahren angegeben. Es wird ersichtlich, dass prinzipiell junge von alten Personen unterschieden werden können, dass aber die Einteilung in fünf Altersgruppen zu je zehn Jahren für die Anwendung des Systems nicht geeignet ist. Deshalb wird eine gröbere Einteilung in die drei Klassen jung, mittel und alt vorgenommen. Zu diesem Zweck werden die Klassifikatoren nicht neu trainiert, sondern die Klassifikatorausgaben werden wie in Abbildung 4.41 gezeigt zusammengefasst.

EGM	**10**	**20**	**30**	**40**	**50**
10	62	17	14	2	3
20	43	28	17	7	3
30	23	29	22	14	10
40	22	10	23	18	25
50	12	8	14	22	42

(a)

ICA	**10**	**20**	**30**	**40**	**50**
10	49	18	16	4	11
20	9	42	22	5	20
30	20	33	21	11	13
40	8	18	27	24	21
50	10	32	20	14	22

(b)

AAM	**10**	**20**	**30**	**40**	**50**
10	65	12	13	3	5
20	7	56	13	14	8
30	20	19	14	29	16
40	4	9	15	36	34
50	10	11	7	26	44

(c)

Tabelle 4.5: *Konfusionsmatrizen und Erkennungsraten für die Altersschätzung mit 5 Klassen. (a) EGM (Erkennungsrate* $0,351$*) (b) ICA + NN (Erkennungsrate* $0,322$*) (c) AAM + MLP (Erkennungsrate* $0,439$*). Senkrecht: wahre Klasse, waagerecht: geschätzte Klasse. Es wird ersichtlich, dass die Verfahren prinzipiell in der Lage sind, alte von jungen Personen zu unterscheiden. Verwechslungen treten hauptsächlich zwischen benachbarten Altersgruppen auf.*

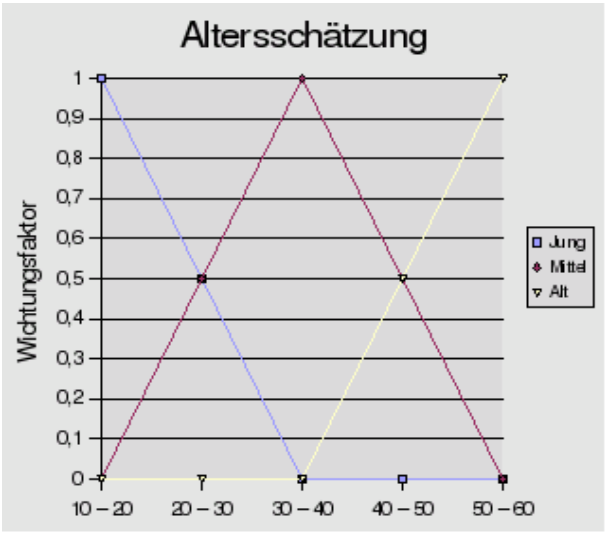


Abbildung 4.41: *Abbildung der Klassifikatorausgaben auf die drei Alterstufen jung, mittel und alt. Die Altersklassen* $10-20$*,* $30-40$*, und* $50-60$ *werden jeweils zu* 100% *den Klassen jung, mittel bzw. alt zugeordnet. Eine Klassifikation in der Altersklasse* $20-30$ *wird zu* 50% *als jung und zu* 50% *als mittel eingestuft. Entsprechend wird eine Klassifikation in der Altersklasse* $40-50$ *zu* 50% *als mittel und zu* 50% *als alt eingestuft.*

Wie bereits erwähnt, wurde kein erneutes Training der Klassifikatoren durchgeführt. Um eine Abschätzung der Erkennungsraten mit der in Abbildung 4.41 gezeigten Zusammenfassung der Altersgruppen zu geben, wurden die trainierten Systeme auf dem Neutraldatensatz angewendet und die Konfusionsmatrizen ermittelt. Hierfür wurden die wahren Klassen so gebildet, dass Personen unter 25 Jahren als jung, Personen über 45 Jahren als alt und dazwischen als mittel gelten. Tabelle 4.6 zeigt die so ermittelten Konfusionsmatrizen der drei Systeme.

EGM	jung	mittel	alt
jung	90	63	1
mittel	42	162	6
alt	19	76	31

(a)

ICA	jung	mittel	alt
jung	135	16	3
mittel	34	160	16
alt	10	26	90

(b)

AAM	jung	mittel	alt
jung	117	37	0
mittel	9	163	38
alt	1	11	114

(c)

Tabelle 4.6: *Konfusionsmatrizen und Erkennungsraten für die Altersschätzung mit 3 Klassen. (a) EGM (Erkennungsrate* $0,578$*) (b) ICA + NN (Erkennungsrate* $0,786$*) (c) AAM + MLP (Erkennungsrate* $0,804$*). Senkrecht: wahre Klasse, waagerecht: geschätzte Klasse. Bei Verwendung von drei Altersklassen liegt die Erkennungsrate bei Raten des Alters bei* $0,333$*.*

4.7.2.3 Mimikschätzung

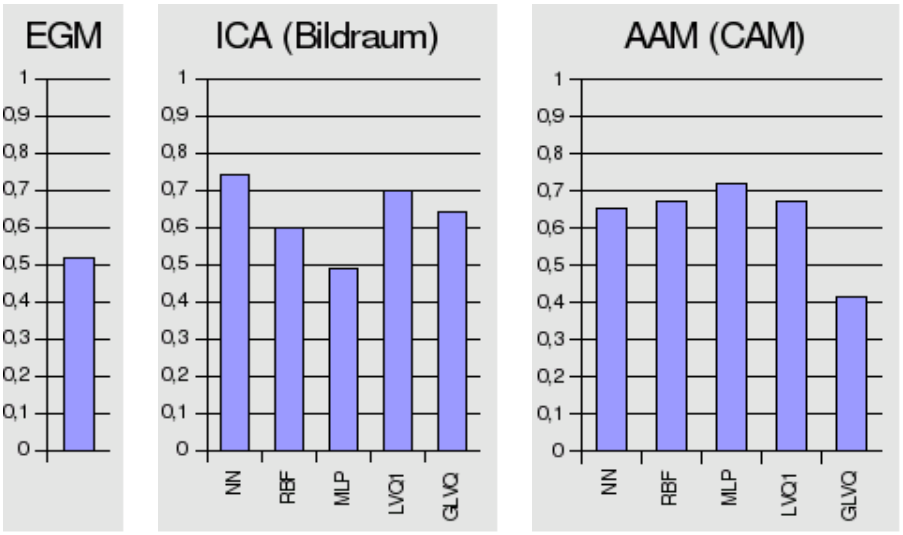
Bei der Mimikschätzung befindet sich immer eine Person im Testdatensatz, mit je einem Bild pro Gesichtsausdruck. Damit ergeben sich bei 30 Personen $30/1 = 30$ Trainingsdurchläufe. Vier Personen bilden den Validierungsdatensatz und die restlichen 25 den Trainingsdatensatz. Die schlechteste Erkennungsrate (die durch Raten des Gesichtsausdrucks bei sieben möglichen Klassen erreicht werden kann) liegt bei $0,143$. Abbildung 4.42 zeigt die Erkennungsraten von EGM, ICA und AAM für unterschiedliche Klassifikatoren.

Die besten Erkennungsraten bei der Mimikschätzung liefern mit 74% die ICA in Kombination mit einem Nearest-Neighbor-Klassifikator und mit 72% ein AAM wiederum in Kombination mit einem MLP. Das Elastic-Graph-Matching schneidet bei diesem Vergleich mit lediglich 52% verhältnismäßig schlecht ab.

Auch bei der Mimikschätzung ist es interessant, einen Blick auf die Konfusionsmatrizen zu werfen. Beispielhaft wird in Tabelle 4.7 die Konfusionsmatrix der Mimikschätzung bei Verwendung der ICA (Bildraum, reduziert) mit einem Nearest-Neighbor-Klassifikator dargestellt. Diese Kombination aus Merkmalsextraktion und Klassifikator erreicht eine Erkennungsrate von 0.724%. Aus der Konfusionsmatrix wird ersichtlich, welche Gesichtsausdrücke typischerweise verwechselt werden.

4.7.2.4 Identifikation

Bei der Personenidentifikation befinden sich immer alle 70 Personen mit je zwei Bildern pro Person im Testdatensatz. Zwei Bilder pro Person bilden den Validierungsdatensatz und die restlichen drei Bilder den Trainingsdatensatz. Im Gegensatz zur Geschlechts-, Alters- und Mimikschätzung werden also bei der Personenidentifikation nicht alle sieben Bilder einer Person



(a) (b) (c)

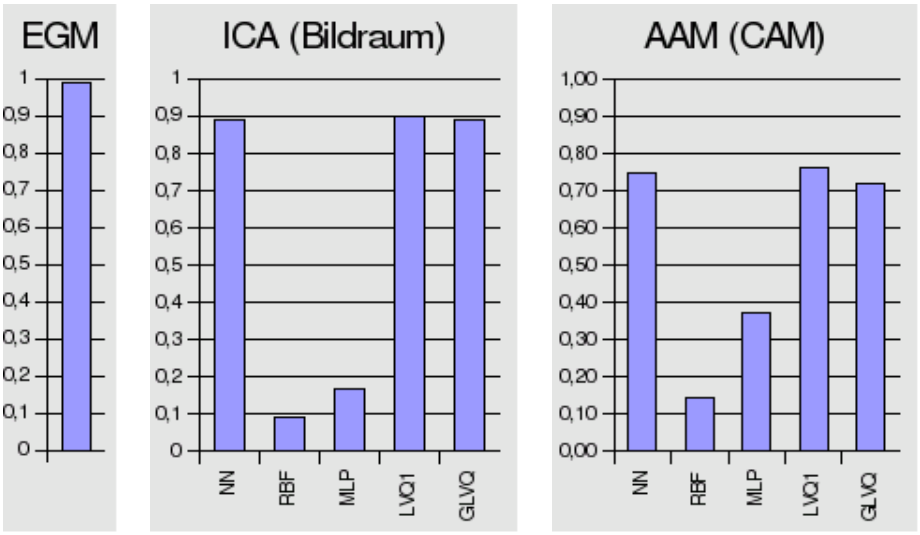
Abbildung 4.42: *Erkennungsraten bei der* ***Mimikschätzung****. (a) EGM (b) ICA (Bildraum, unreduziert) (c) AAM (CAM). Die besten Erkennungsraten liefert hier mit 74% die ICA mit einem Nearest-Neighbor-Klassifikator, gefolgt von einem AAM in Kombination mit einem MLP mit 72%.*

ICA	N	Ü	T	W	A	E	F
Neutral	29	0	0	0	1	0	0
Überraschung	1	27	2	0	0	0	0
Trauer	4	1	19	5	1	0	0
Wut	1	0	5	23	0	0	1
Angst	5	1	2	1	14	3	4
Ekel	0	0	2	2	2	15	9
Freude	0	0	0	2	1	2	25

Tabelle 4.7: *Konfusionsmatrix der ICA (Bildraum, reduziert) mit einem Nearest-Neighbor-Klassifikator bei der Mimikschätzung. Senkrecht: wahre Klasse, waagerecht: geschätzte Klasse. Besonders häufig falsch geschätzt wurden die folgenden Gesichtsausdrücke: Trauer → Wut, Trauer → Neutral, Wut → Trauer, Angst → Neutral, Angst → Freude, Ekel → Freude. Die Verwechslung eines Gesichtsausdrucks mit Neutral ist vermutlich auf eine schwache Ausprägung des Gesichtsausdrucks zurückzuführen, während andere Verwechslungen, z.B. Wut und Trauer bzw. Ekel und Freude auf starke Ähnlichkeiten in den jeweiligen Gesichtsausdrücken zurückzuführen sind. Die häufig falsch klassifizierten Gesichtsausdrücke Angst, Ekel und Trauer stimmen erstaunlich gut mit denjenigen überein, die auch bei der manuellen Vorklassifikation der NIFace2-Datenbank Probleme bereiteten, siehe Abbildung 4.4.*

dem Trainings-, Validierungs- oder Testdatensatz zugeordnet, sondern auf diese drei Datensätze aufgeteilt. Die Anzahl der Klassen beträgt somit 70 und es wird nur ein Trainingsdurchlauf durchgeführt. Abbildung 4.43 zeigt die Erkennungsraten von EGM, ICA und AAM für unterschiedliche Klassifikatoren.

Bei diesem Test schneidet das Elastic-Graph-Matching mit Abstand am besten ab. Es erreicht eine Erkennungsrate von 99%. Die ICA erreicht mit einem LVQ1-Klassifikator eine Erkennungs-



(a) (b) (c)

Abbildung 4.43: *Erkennungsraten bei der Personenidentifikation. (a) EGM (b) ICA (Bildraum, reduziert) (c) AAM (CAM). Im Gegensatz zu den Anderen Erkennungsaufgaben liefert hier das EGM mit Abstand die besten Erkennungsraten.*

rate von 90% und das AAM ebenfalls mit einem LVQ1-Klassifikator 76%.

Für den Einsatz im Baumarkt ist diese Vorgehensweise der Personenidentifikation allerdings nicht praktikabel, denn letztendlich soll ein $1 : n$-Vergleich durchgeführt werden, bei dem für eine Testperson entschieden wird, um welche Person aus einer Datenbank es sich handelt. Die Zusammensetzung dieser Datenbank ändert sich jedoch mit der An- und Abmeldung von Personen ständig und kann im Extremfall auch nur aus einer Person, nämlich dem aktuellen Nutzer, bestehen. In diesem Fall wird nur noch ein $1 : 1$-Vergleich durchgeführt, es wird also entschieden, ob es sich bei einer Person um diesen Nutzer handelt oder nicht. Für diesen Anwendungsfall wird die Ähnlichkeit zwischen einem Modell, das für die Person erstellt wurde und allen in der Datenbank enthaltenen Modellen berechnet. Beim EGM wird hierfür das betragsbasierte Ähnlichkeitsmaß für Graphen verwendet, siehe Gleichung 4.10. Bei der ICA und dem AAM wird das normierte Skalarprodukt zwischen den Fitwertvektoren bzw. den Appearance-Parametern der Modelle berechnet. Befindet sich nur eine Person in der Datenbank, wird anhand einer Ähnlichkeitsschwelle θ (siehe Abbildung 4.46) entschieden, ob es sich um die Testperson handelt oder nicht. Befinden sich mehrere Personen in der Datenbank, wird dies für die ähnlichste Person entschieden.

Bei diesem Vergleich kann es zu zwei Arten von Fehlern kommen. Zum einen kann eine Person, die in der Datenbank hinterlegt ist, fälschlicherweise zurückgewiesen werden (False-Rejection). Zum anderen kann eine Person, die nicht in der Datenbank hinterlegt ist, als bekannter Nutzer angenommen werden (False-Acceptance). Die Häufigkeit von Falsch-Rückweisungen und Falsch-Annahmen wird von der Ähnlichkeitsschranke θ bestimmt. Bei einer sehr kleinen Schranke ist die Anzahl der Falsch-Annahmen hoch und bei einer sehr großen Schranke ist die Anzahl der Falsch-Rückweisungen hoch. Die Leistungsfähigkeit der Verfahren in Bezug auf die Personenerkennung kann mit der so genannten FAR/FRR-Kurve veranschaulicht werden. Um eine solche

Kurve zu erstellen, wird für einen Datensatz zunächst für jedes Bild ein Modell erstellt. Dann wird zwischen je zwei Modellen die Ähnlichkeit berechnet. Nicht betrachtet wird die Ähnlichkeit eines Modells mit sich selbst. Die Ähnlichkeitswerte werden in einer Liste gespeichert, wobei jeder Wert mit der Information versehen wird, ob die beiden Modelle von einer Person stammen (intrapersonal), oder von unterschiedlichen Personen (extrapersonal). Diese Liste wird nun aufsteigend nach den Ähnlichkeiten sortiert. Aus der sortierten Liste wird dann eine False-Acceptance-Rate (FAR)-Kurve und eine False-Rejection-Rate (FRR)-Kurve erzeugt. Die False-Acceptance-Rate (FAR) ist die Anzahl derjenigen Paare, deren Ähnlichkeit größer als die Ähnlichkeitsschranke θ ist, die also als intrapersonal angesehenen werden, obwohl sie in Wirklichkeit extrapersonal sind. Diese Anzahl wird im Verhältnis zur Gesamtanzahl extrapersonaler Bildpaare betrachtet:

$$FAR(\theta) = \frac{\#\text{falsch als intrapersonal erkannte Paare}}{\#\text{extrapersonale Paare}} \tag{4.55}$$

Die False-Rejection-Rate ist die prozentuale Anzahl der als extrapersonal angesehenen Paare, die in Wirklichkeit intrapersonal sind. Sie ist also das Gegenstück zur FAR:

$$FRR(\theta) = \frac{\#\text{falsch als extrapersonal erkannte Paare}}{\#\text{intrapersonale Paare}} \tag{4.56}$$

Abbildung 4.44 zeigt die FAR/FRR-Kurven der drei Verfahren für die neutralen Gesichtsausdrücke und für den Gesamtdatensatz inklusive Mimikdaten jeweils bei Verwendung der gelabelten Augenpositionen bzw. bei Schätzung der Augenpositionen mit AdaBoost. Es wird ersichtlich, dass die Art der Bestimmung der Augenpositionen bei allen drei Verfahren nur einen relativ geringen Einfluss auf die False-Acceptance- und False-Rejection-Rates hat.

Eine wichtige Kenngröße für die Güte der Personenidentifikation ist die Equal-Error-Rate (EER). Sie wird im Schnittpunkt der FRR- und der FAR-Kurve bestimmt. Wie in Abbildung 4.44 zu sehen ist, steigen die Equal-Error-Rates für alle Verfahren, wenn neben neutralen Gesichtsausdrücken auch Mimikdaten verwendet werden, da dann die Innerklassenvarianzen durch die größeren Unterschiede zwischen Bildern einer Person zunehmen. Die Sensitivität ist beim EGM am größten, gefolgt vom AAM. Die Personenidentifikation mittels ICA ist am robustesten gegenüber variierenden Gesichtsausdrücken. Tabelle 4.8 und Abbildung 4.45 zeigen die Equal-Error-Rates für die drei Verfahren.

Wenn die Positionen der Augen nicht vorgegeben, sondern durch das in Abschnitt 3.3.4.3 beschriebene Verfahren zur Augendetektion ermittelt werden, steigen die Equal-Error-Rates mit Ausnahme bei den AAMs ebenfalls an. Die Robustheit der AAMs lässt sich durch deren sehr gute Fähigkeit zur Modellanpassung bei schlechter Schätzung der Augenpositionen erklären. Selbst wenn die Positionen der Augen nicht exakt gefunden wurden, werden die AAMs oft noch korrekt auf das Eingangsbild angepasst. Ein stärkerer Anstieg der EER ist für das EGM und die ICA zu verzeichnen.

Ausschlaggebend für die Auswahl eines Verfahrens zur Personenidentifikation bzw. -verifikation sollten die Equal-Error-Rates bei automatischer Detektion der Augen und unter Verwendung

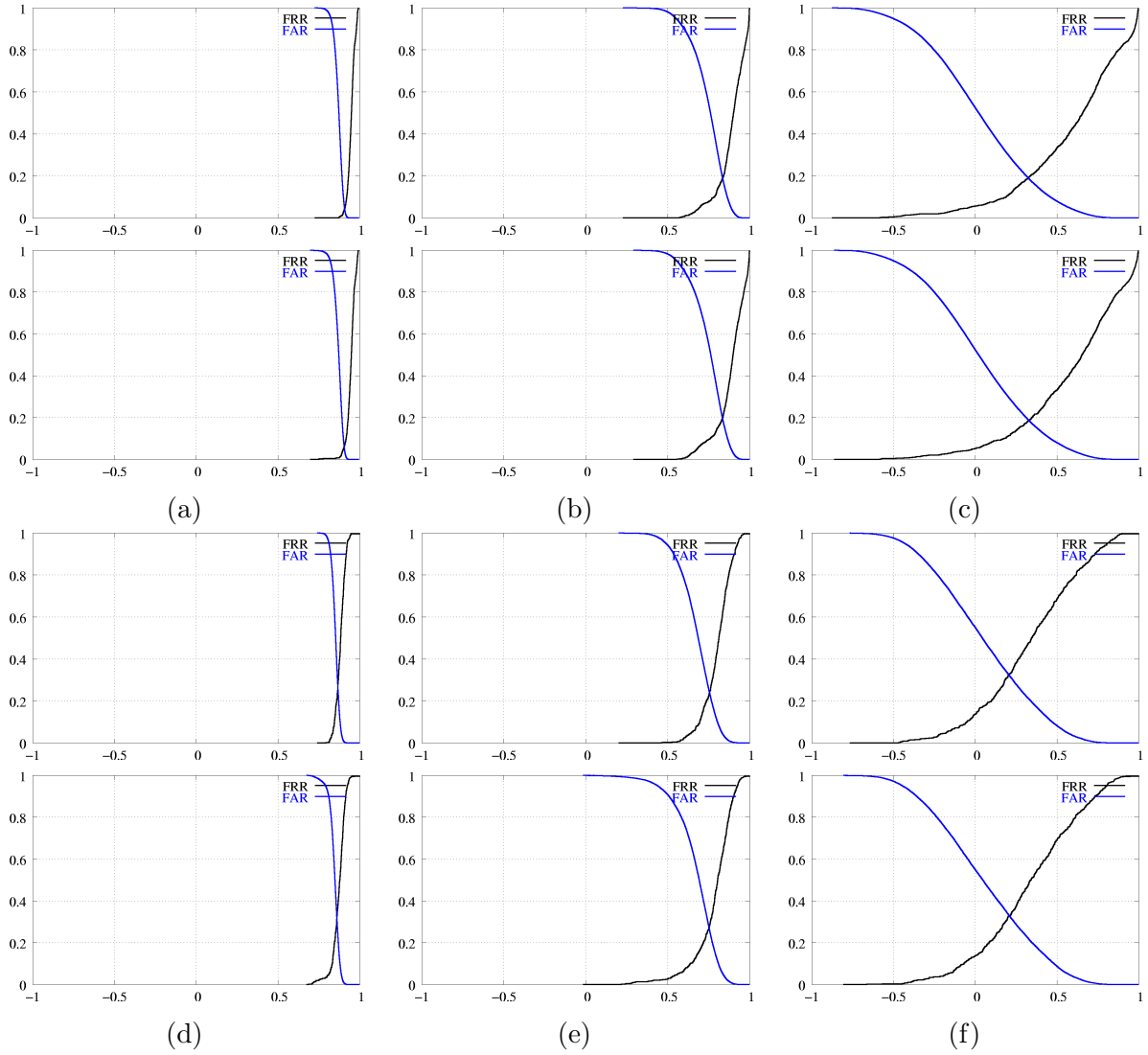
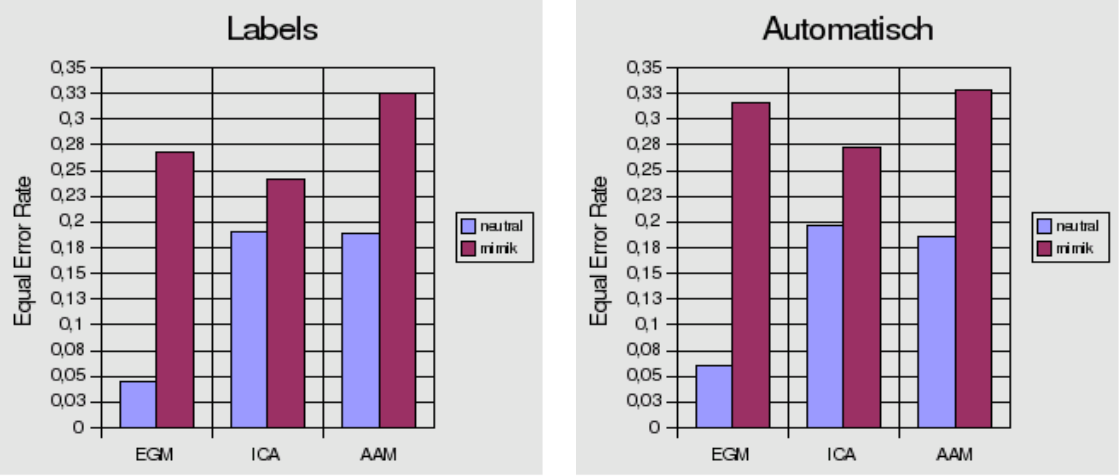


Abbildung 4.44: *(a)-(c) FRR/FAR-Kurven für die Personenidentifikation mit (a) EGM (b) ICA (c) AAM auf dem Datensatz mit neutralen Gesichtsausdrücken. Obere Reihe: bei Verwendung der gelabelten Augenpositionen, untere Reihe: Schätzung der Augenpositionen mit AdaBoost. (d)-(f) FRR/FAR-Kurven für die Personenidentifikation mit (d) EGM (e) ICA (f) AAM auf dem Mimik-Datensatz. Obere Reihe: bei Verwendung der gelabelten Augenpositionen, untere Reihe: Schätzung der Augenpositionen mit AdaBoost. Die EER wird im Schnittpunkt der FRR- und der FAR-Kurve bestimmt und sollte einen möglichst kleinen wert besitzen. Es wird ersichtlich, dass die Art der Schätzung der Augenpositionen einen relativ geringen Einfluss auf die Equal-Error-Rates haben, vgl. Tabelle 4.8, wohingegen die Verwendung von Mimikdaten bei allen drei Verfahren zu einem signifikanten Anstieg der Equal-Error-Rates im Vergleich zu Neutraldaten führt.*

der Mimikdaten sein, da dieser Testfall am ehesten der Realität entspricht. Interessanterweise schneidet hier die ICA am besten ab, obwohl das EGM auf den Neutraldaten mit Abstand die besten Erkennungsraten geliefert hat, vgl. Abbildung 4.43. Um eine Entscheidung treffen zu können, ob eine Person akzeptiert oder zurückgewiesen wird, muss eine Schwelle θ für die Ähnlichkeit festgelegt werden. Diese wird so eingestellt, dass eine bestimmte False-Acceptance-Rate

	manuell			automatisch		
	EGM	**ICA**	**AAM**	**EGM**	**ICA**	**AAM**
Neutral	0,04	0,19	0,19	0,06	0,2	0,19
Mimik	0,27	0,24	0,33	0,32	0,27	0,33

Tabelle 4.8: *Equal-Error-Rates der drei Verfahren auf dem Neutral- und auf dem Mimikdatensatz, siehe auch Abbildung 4.45. Die EERs wurden einmal bei manueller Vorgabe der Augenpositionen und einmal bei automatischer Bestimmung durch AdaBoost ermittelt. Einmal wurden nur neutrale Ansichten der Personen verwendet und einmal Ansichten mit verschiedenen Gesichtsausdrücken.*



(a) (b)

Abbildung 4.45: *Equal Error Rates. (a) Bei Vorgabe der Augenpositionen durch die Labelpunkte und (b) bei automatischer Detektion der Augen durch AdaBoost.*

erreicht (nicht überschritten) wird. Aus der FRR-FAR-Kurve kann dann für diese Schwelle die zugehörige False-Rejection-Rate abgelesen werden, siehe Abbildung 4.46. Diese sollte natürlich möglichst klein sein. Tabelle 4.9 und Abbildung 4.47 zeigen die Ähnlichkeitsschwellen und False-Rejection-Rates für verschiedene vorgegebene False-Acceptance-Rates.

4.7.2.5 Rechenzeit

Neben den Erkennungsraten ist natürlich auch der durchschnittliche Zeitbedarf für die Modellanpassung von Interesse für die Auswahl eines Verfahrens für das Gesamtsystem. Um diese Kenngrößen zu bestimmen, wurden die drei Verfahren auf einem AMD Athlon XP 3000+ auf dem gesamten Neutral- bzw. dem gesamten Mimikdatensatz angepasst. Dabei wurden die Zeiten aus Tabelle 4.10 ermittelt, wobei die Zeiten für die Klassifikation nicht berücksichtigt werden, da sie im Vergleich zur Modellanpassung vernachlässigt werden können.

Die Komplexität der Verfahren spiegelt sich direkt im Zeitbedarf für die Modellanwendung wieder. Bei der ICA erfolgt lediglich eine Unterraumprojektion, d.h. eine Reihe von Matrixmultiplikationen. Beim Local-Move des EGM wird die maximale Ähnlichkeit eines Jets aus

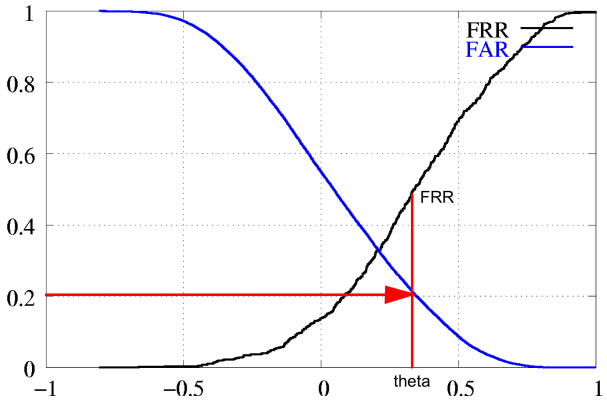


Abbildung 4.46: *Art der Bestimmung der Ähnlichkeitsschwelle θ und der False-Rejection-Rate bei Vorgabe einer False-Acceptance-Rate. In diesem Beispiel wird eine zu erreichende FAR von $0,2$ vorgegeben. Für die entsprechende Stelle in der Kurve kann dann die Ähnlichkeitsschwelle und die zugehörige FRR abgelesen werden.*

0,3	**EGM**	**ICA**	**AAM**
θ	0,862	0,745	0,237
FRR	0,332	0,257	0,357

(a)

0,2	**EGM**	**ICA**	**AAM**
θ	0,870	0,773	0,347
FRR	0,405	0,340	0,510

(b)

0,1	**EGM**	**ICA**	**AAM**
θ	0,880	0,810	0,482
FRR	0,511	0,524	0,668

(c)

0,05	**EGM**	**ICA**	**AAM**
θ	0,890	0,837	0,570
FRR	0,605	0,633	0,754

(d)

Tabelle 4.9: *Ähnlichkeitsschwellen θ und False-Rejection-Rates für verschiedene vorgegebene False-Acceptance-Rates. (a) FAR 0,3 (b) FAR 0,2 (c) FAR 0,1 (d) FAR 0,05. Die Werte wurden auf dem Mimikdatensatz unter Verwendung der automatischen Augendetektion mit AdaBoost bestimmt, also dem Testszenario, das der realen Anwendung am nächsten kommt. Die ermittelten Werte werden in Abbildung 4.47 nochmals dargestellt.*

EGM	**ICA**	**AAM**
898	21	2225

Tabelle 4.10: *Durchschnittlicher Zeitbedarf in Millisekunden für die Modellanpassung auf dem Neutral- und dem Mimikdatensatz.*

dem Eingabebild in einer lokalen Umgebung des entsprechenden Jets aus dem Average-Graphen gesucht. Beim AAM wird schließlich eine iterative Anpassung der Appearance-Parameter durchgeführt, bis die Energie der Differenz zwischen formnormiertem Eingabebild und synthetisiertem Bild minimal wird. Dieser Prozess ist rund 100 mal langsamer als die Unterraumprojektion bei der ICA. Dazu ist zu sagen, dass bei keinem der Verfahren eine Optimierung hinsichtlich der Rechenzeit durchgeführt wurde.

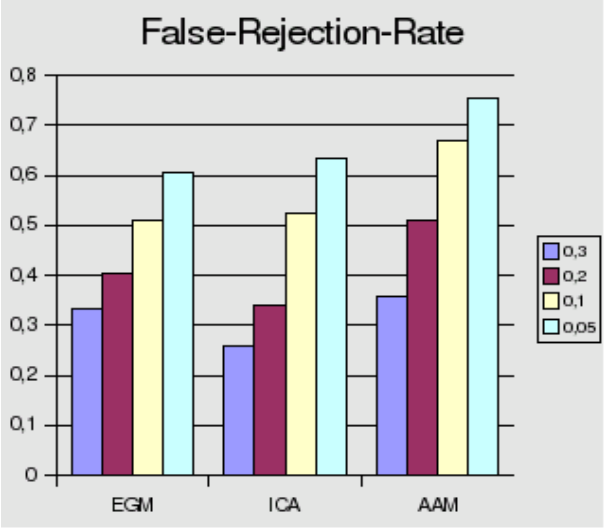


Abbildung 4.47: *False-Rejection-Rates der drei Verfahren aus Tabelle 4.9 in Abhängigkeit von der vorgegebenen False-Acceptance-Rate. Auch hier schneidet die ICA sehr gut ab. Nur bei False-Acceptance-Rates kleiner* $0,1$ *schneidet das EGM besser ab.*

4.7.2.6 Fazit

Die Ergebnisse bei der Geschlechts-, Alters- und Mimikschätzung sind sehr ähnlich. In allen Fällen werden sehr gute Erkennungsraten von Active-Appearance-Models in Kombination mit MLPs erreicht. Dies lässt darauf schließen, dass die Active-Appearance-Parameter hinsichtlich der jeweiligen Erkennungsaufgabe so beschaffen sind, dass sie durch eine globale Trennfunktion gut separiert werden können. Diese Beobachtung lässt sich dadurch erklären, dass oftmals einzelne Appearance-Parameter eine bestimmte Änderung im Gesicht kodieren. So lässt beim kombinierten AAM der vierte Appearance-Parameter einen deutlichen Einfluss auf das Geschlecht des synthetisierten Gesichtes erkennen, siehe Abbildung 4.48(b). Der Einfluss der einzelnen Parameter wird, wie bereits erwähnt, lediglich aus den statistischen Eigenschaften der Trainingsdaten bestimmt und nicht aufgrund einer bestimmten Erkennungsaufgabe. Wenn also typische Unterschiede in den Grauwertverläufen oder Formen zwischen bestimmten Klassen existieren, werden diese je nach ihrer Ausgeprägtheit auch durch einen oder mehrere Appearance-Parameter kodiert, siehe Abbildung 4.48.

Ein weiterer Vorteil der AAMs besteht in deren Toleranz gegenüber der Genauigkeit der Startschätzung für die Augenpositionen. Bei der Personenidentifikation ist es das einzige Verfahren, bei dem die Equal-Error-Rate bei automatischer Detektion im Vergleich zur manuellen Vorgabe der Augenpositionen nicht fällt.

Die Stärke der Independent-Component-Analysis zeigt sich dagegen eher bei Verwendung von Nearest-Neighbor- bzw. LVQ-Klassifikatoren. Dies lässt vermuten, dass die ICA-Fit-Werte hinsichtlich der einzelnen Klassen mehrere kleine stärker innenander verzahnte Clustern bilden, die durch LVQ-Prototypen besser beschrieben werden können als durch die global wirkende konvexe Trennfunktion eines MLP.

Obwohl die Active-Appearance-Models ihr Potential für die Analyse von Gesichtern bewiesen haben, spricht die für die Modellanpassung benötigte Rechenzeit noch gegen einen Einsatz im realen Betrieb. Zum gegenwärtigen Zeitpunkt ist die Personenidentifikation das wichtigste Einsatzfeld der Gesichtsanalyse auf dem mobilen Serviceroboter PERSES. Bei dieser Aufgabe schneidet die ICA sehr gut ab, da die Equal-Error-Rates bei Verwendung der Mimikdaten im Vergleich zum EGM und den AAMs nur geringfügig steigen. Diese Tatsache im Zusammenhang mit der mit rund 21ms sehr schnellen Modellberechnung lässt die ICA zum jetzigen Zeitpunkt als das für den realen Einsatz am besten geeignete Verfahren der Merkmalsextraktion erscheinen.

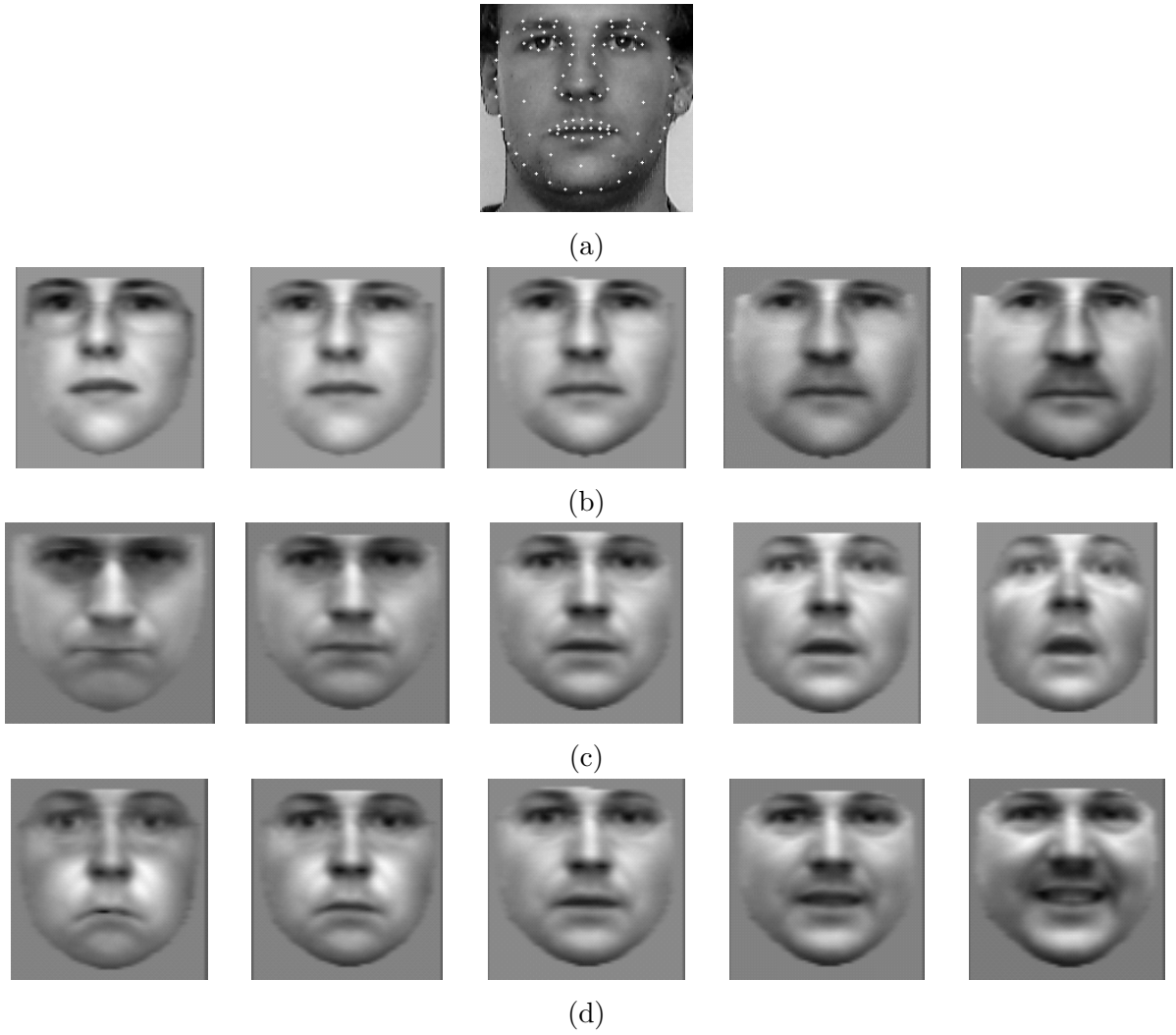


Abbildung 4.48: *Einfluss verschiedener Appearance-Parameter auf ein synthetisiertes Gesicht. (a) Eingangsbild (b) von einem auf den Neutraldaten erstellten Active-Appearance-Model synthetisiertes Bild bei -3,0/-1,5/0/+1,5/+3,0 Standardabweichungen des vierten Appearance-Parameters. Neben der Blickrichtung hat dieser Parameter auch einen deutlichen Einfluss auf das Geschlecht des synthetisierten Gesichtes, was sich insbesondere am Bartwuchs aber auch an der Form manifestiert. (c) Von einem auf den Mimikdaten erstellten Active-Appearance-Model synthetisiertes Bild bei -3,0/-1,5/0/+1,5/+3,0 Standardabweichungen des ersten Appearance-Parameters. Der Parameter variiert die Öffnung des Mundes und die Stellung der Augenbrauen, die offensichtlich in den Trainingsdaten korreliert sind. Der Gesichtsausdruck variiert entsprechend von ärgerlich zu überrascht. (d) Von einem auf den Mimikdaten erstellten Active-Appearance-Model synthetisiertes Bild bei -3,0/-1,5/0/+1,5/+3,0 Standardabweichungen des dritten Appearance-Parameters. Der Parameter variiert die Stellung der Mundwinkel von unten nach oben und die Öffnung des Mundes. Entsprechend variiert der Gesichtsausdruck von traurig nach freundlich.*

Kapitel 5

Integration der Teilsysteme und experimentelle Untersuchungen

5.1 Aufgabenstellung

An dieser Stelle geht es darum, eine Software-Architektur zu konzipieren, in die die entwickelten Module integriert werden können. Diese soll so beschaffen sein, das verschiedene Anwendungen des Robotersystems realisiert werden können, von der Demonstration von Teilleistungen wie Interaktion oder Navigation, bis zur vollständigen Anwendung für das Baumarktszenario. Die Leistungsfähigkeit der Teilmodule des peripheren und fovealen Vision-Systems und deren Zusammenspiel in der vorgestellten Software-Architektur werden dann anhand einiger Tests mit Versuchspersonen ermittelt.

5.2 Software-Architektur

Die Architektur ist das Ergebnis eines iterativen Entwicklungsprozesses am Fachgebiet für Neuroinformatik und Kognitive Robotik, in den mehrere Jahre Erfahrung mit Steuerarchitekturen eingeflossen sind. Beim Entwurf fanden die folgenden Punkte Berücksichtigung:

Hardware-Transparenz: Indem die Rohdaten von den Sensoren des Roboters auf eine zentrale Kommunikationsstruktur, das so genannte Blackboard, geschrieben werden, können die Berechnungsmodule, da sie nicht direkt auf die Hardware zugreifen, Hardware-unabhängig implementiert werden. So können z.B. die Bilder für das foveale Vision-System aus einer Digitalkamera, einer WebCam oder einer Datei ausgelesen werden.

Software-Transparenz: Um eine möglichst große Transparenz zu erreichen, kommunizieren die einzelnen Module nicht direkt miteinander, sondern ebenfalls über das Blackboard. Auf diese Weise wird eine direkte Anbindung von Modulen untereinander vermieden und das Gesamtsystem bleibt auch bei Hinzunahme weiterer Module übersichtlich und wartbar.

Jedes Modul liest die für seine Berechnungen notwendigen Daten vom Blackboard und schreibt auch die Ergebnisse dorthin. Module können so unabhängig voneinander entwickelt und relativ einfach in die Architektur integriert werden.

Anwendungstransparenz: Eine weitere Abstraktion besteht auf der Anwendungsebene. Eine Anwendung wird definiert durch Zustände und damit verbundene Aktionen und durch Bedingungen ausgelöste Zustandsübergänge. Eine solche Bedingung kann z.B. sein, dass das foveale Vision-System einen bestimmten Nutzer wiedererkannt hat. An dieser Stelle ist es nicht mehr von Interesse, wie diese Information ermittelt wird, d.h. welches Verfahren für die Wiedererkennung verwendet wurde, EGM, ICA oder AAM. Ein Austausch des entsprechenden Erkennungsmoduls bleibt somit ohne Auswirkung auf der Anwendungsebene. Damit kann eine Anwendung auf eine andere Roboterplattform mit anderen Erkennungsmodulen übertragen werden.

Abbildung 5.1 zeigt die Systemarchitektur von PERSES. Im Folgenden werden die einzelnen Komponenten dieser Architektur näher vorgestellt.

Hardware: Hier werden zum einen die Sensordaten ausgelesen und auf das Blackboard geschrieben und zum anderen Stelleingriffe auf die Hardware realisiert, z.B. durch das periphere Vision-System auf die PTU realisiert.

Blackboard: Das Blackboard ist die zentrale Kommunikationsstruktur, auf der die Module Rohdaten von der Hardware oder Berechnungsergebnisse anderer Module abfragen können und ihre Berechnungsergebnisse ablegen. Durch das Blackboard wird sowohl eine weitgehende Unabhängigkeit der Module von der Hardware als auch eine Unabhängigkeit der Module untereinander erreicht.

Module: Die Module realisieren bestimmte Teilleistungen des Systems, wie Navigation und Hindernisvermeidung, Selbstlokalisation und Nutzer-Tracking und Nutzeranalyse. Hier gliedern sich also die beiden in dieser Arbeit entwickelten Module an.

Variablen-Interface: Das Variablen-Interface stellt eine abstrakte Schnittstelle zwischen dem hardwarespezifischen Blackboard und den anwendungsspezifischen Komponenten dar. Diese ist so realisiert, dass Abfragen von Sensordaten oder Berechnungsergebnissen der Module durchgeführt werden können, ohne dass die konkrete Hardware oder Implementierung im Berechnungsmodul bekannt ist.

Zustandsautomat: Der Zustandsautomat bestimmt den Ablauf einer Anwendung. Es handelt sich um einen deterministischen endlichen Automaten, mit dem sowohl einfache automatisch und ohne Interaktion ablaufende Demonstrationen als auch komplexe Anwendungen, wie z.B. für das Baumarktszenario, siehe Abbildung 1.1, realisiert werden können. Der Ablauf kann durch eine XML-Datei konfiguriert werden, so dass nicht in die Programmstruktur eingegriffen werden muss. Zustandsübergänge werden durch Bedingungen, wie z.B. das Vorhandensein eines Nutzers oder das Betätigen einer Taste auf dem Display ausgelöst.

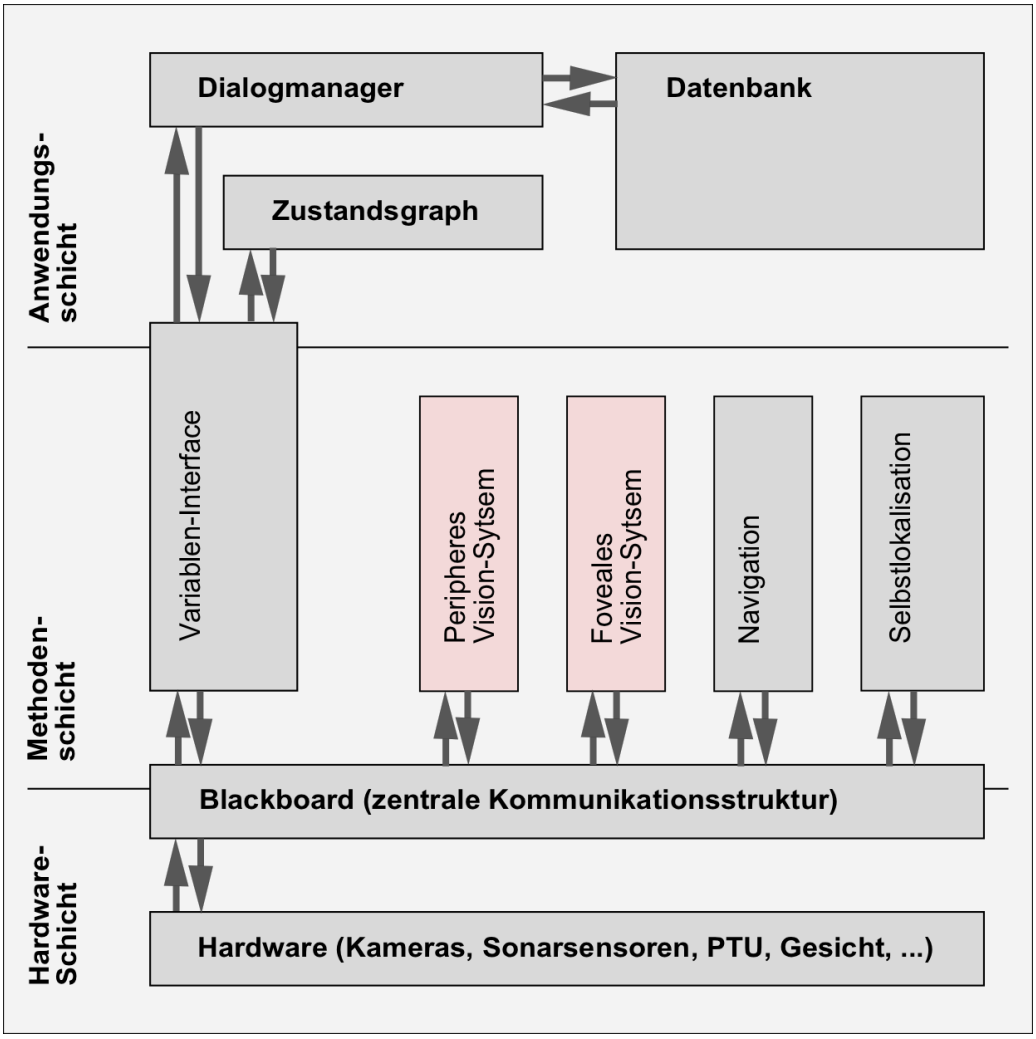


Abbildung 5.1: *Die Systemarchitektur besteht aus drei Schichten. Die unterste Schicht enthält die Hardware-Anbindung. Die mittlere Schicht besteht aus Modulen für die Umweltwahrnehmung und Navigation, die unabhängig von der eigentlichen Anwendung entwickelt und betrieben werden können. Hier gliedern sich die in dieser Arbeit entwickelten Module des peripheren und des fovealen Vision-Systems ein. Die oberste Schicht ist schließlich applikationsspezifisch. Hier wird der Ablauf einer Anwendung und die Art und Weise der Interaktion mit Nutzern spezifiziert. Die einzelnen Komponenten werden im Text näher erläutert.*

Dialogmanager: Der Dialogmanager verwaltet den Dialog mit dem Nutzer. Abhängig vom aktuellen Zustand des Zustandsautomaten werden Sprachausgaben ausgelöst und bestimmte Seiten auf dem Display angezeigt. Hier werden also multiple Ausgabemodalitäten geeignet kombiniert. Der Dialogmanager soll später in der Lage sein, geeignete Aktionen für jeden Systemzustand durch Lernprozesse zu ermitteln.

Datenbank: In der Datenbank sind schließlich die für die jeweilige Anwendung notwendigen Daten hinterlegt. Im Baumarktszenario befindet sich an dieser Stelle z.B. die Anbindung an das Warenwirtschaftssystem.

5.3 Untersuchungen

Das periphere und das foveale Vision-System wurden bis jetzt isoliert voneinander betrachtet. An dieser Stelle soll das Zusammenspiel der beiden entwickelten Teilsysteme untersucht werden. Dazu wird eine Reihe von Versuchspersonen vor den Roboter treten. Das System soll die Person kontinuierlich tracken und die Analyse des Gesichtes durchführen. Über den Zeitraum der Interaktion werden die extrahierten Informationen Geschlecht, Alter, Identität und Gesichtsausdruck protokolliert. Das System sollte die Person anschauen und dabei Identität, Alter und Geschlecht und den neutralen Gesichtsausdruck richtig und über den gesamten Zeitraum möglichst stabil schätzen.

5.3.1 Berechnung der Klassifikationsergebnisse

In der Anwendung des Gesamtsystems im realen Betrieb werden im Unterschied zu den Offline-Untersuchungen noch folgende Mechanismen bei der Berechnung der Klassifikationsergebnisse verwendet.

Modellgüte Vor der Klassifikation ist es sinnvoll, ein Gütemaß für das aus dem aktuellen Eingangsbild erstellte Modell zu schätzen. Durch Faktoren wie einen schlechten Bildkontrast, Beleuchtungsgradienten, zu starke Variationen der Pose oder eine schlechte Schätzung der Augenpositionen kann es vorkommen, dass ein erstelltes Modell nicht sinnvoll klassifiziert werden kann, da es entweder zu weit von den im Training verwendeten Daten abweicht oder gar kein Gesicht beschreibt. Für alle drei Modelltypen wurde daher ein Gütekriterium definiert, das vor der Klassifikation evaluiert wird. Bei der ICA wurde ein Fitwertvektor für das aus den Trainingsdaten berechnete Mittelwertgesicht bestimmt. Die Güte ergibt sich aus dem normierten Skalarprodukt zwischen diesem mittleren und dem aus dem aktuellen Eingangsbild ermittelten Fitwertvektor. Die Klassifikation erfolgt nur, wenn das Skalarprodukt einen empirisch bestimmten Schwellwert überschreitet. Beim AAM kann die Güte durch die Länge des Vektors der Appearance-Parameter bestimmt werden. Da der Null-Vektor für die mittlere Form und die mittlere Grauwertverteilung steht, wird ihm eine Güte von 1 zugewiesen. Weicht ein Appearance-Parameter von Null ab, bedeutet dies eine Abweichung in Form oder Grauwert vom Mittelwertgesicht. Es wurde wiederum empirisch ein Schwellwert festgelegt, bei dem typische Variationen in der Beleuchtung und Form toleriert werden. Beim EGM schließlich wird die Ähnlichkeit zwischen dem Bildgraph und dem Average-Graph als Gütemaß verwendet. Auch hier wird keine Klassifikation durchgeführt, wenn ein Schwellwert für die Ähnlichkeit unterschritten wird. Die Gütemaße werden jeweils auf den Bereich $[0, 1]$ normiert.

Kombination mehrerer Klassifikatoren Bei der Nutzeranalyse mit ICA und AAM werden neuronale Klassifikatoren verwendet. Da beim Training dieser Analyseverfahren eine Leave-n-out-Strategie zum Einsatz kam, stehen jeweils mehrere trainierte Klassifikatoren zur Verfügung. Daher ist es nahe liegend, in der Anwendung nicht nur mit einem Klassifikator

zu arbeiten, sondern das Klassifikationsergebnis durch eine Mittelung über die Ausgabe mehrerer Klassifikatoren zu berechnen.

5.3.2 Beispielinteraktionen

In diesem Abschnitt werden einige Beispielinteraktionen gezeigt. Dabei trat jeweils eine Person vor den Roboter und meldete sich bei diesem an. Im weiteren Verlauf wurde die Person durch das periphere Vision-System kontinuierlich getrackt und durch das foveale Vision-System wurden Geschlecht, Alter und Mimik geschätzt. Dabei erfolgte die Merkmalsextraktion durch eine Unterraumprojektion auf die im Bildraum berechneten unabhängigen Komponenten (ICA) und die Klassifikation durch Nearest-Neighbor-Klassifikatoren. Nach dem Anmelden wurden jeweils drei Bilder für die Erstellung eines Nutzermodells verwendet, das im weiteren Verlauf für eine Personenidentifikation eingesetzt wurde. Die Ergebnisse werden wie in Abbildung 5.2 präsentiert. Eine genauere Analyse des Gesichtes erfolgt dabei nur, wenn die entsprechend Abschnitt 5.3.1 ermittelte Modellgüte eine bestimmte Schwelle überschreitet. Auf den folgenden Seiten werden in den Abbildungen 5.3 bis 5.10 einige solcher Interaktionenssequenzen dargestellt.

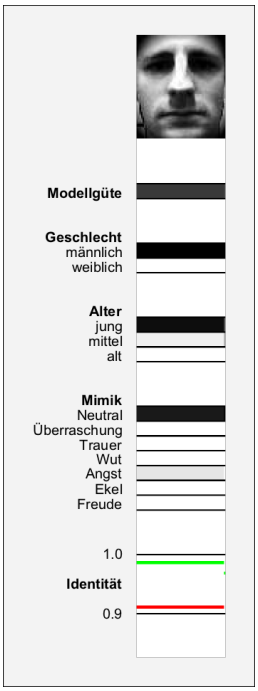


Abbildung 5.2: *Die Modellgüte beschreibt die Ähnlichkeit des für das aktuelle Eingangsbild ertellten Modells zu einem mittleren Gesichtsmodell. Die Klassifikation des Modells wird nur dann durchgeführt, wenn die Modellgüte eine verfahrensabhängige Schwelle überschreitet, vgl. Abschnitt 5.3.1. Die Grauwerte kodieren die Ausgabe der Klassifikatoren: weiß steht für* 0 *und schwarz für* 1*. Zum Beginn jeder Sequenz wird jeweils ein Modell mit drei Bildern für jede Person erstellt, das im weiteren Verlauf zur Personenidentifikation herangezogen wird. Dabei werden alle Modelle angezeigt, deren Ähnlichkeit mit dem Modell des Eingangsbildes die Akzeptanzschwelle von* $0,9$ *überschreiten. Das Modell, das dem Interaktionspartner entspricht, wird grün angezeigt, die anderen rot. In der Bildunterschrift wird jeweils beschrieben, um welche Person es sich handelt, welches Geschlecht und Alter diese hat, welchen Zeitraum die Sequenz umfasst und welche Personen in der Nutzerdatenbank für die Personenidentifikation hinterlegt sind. Am linken Rand der Darstellung ist die Legende dargestellt und am rechten Rand eine zeitliche Mittelung der Schätzung von Geschlecht und Alter.*

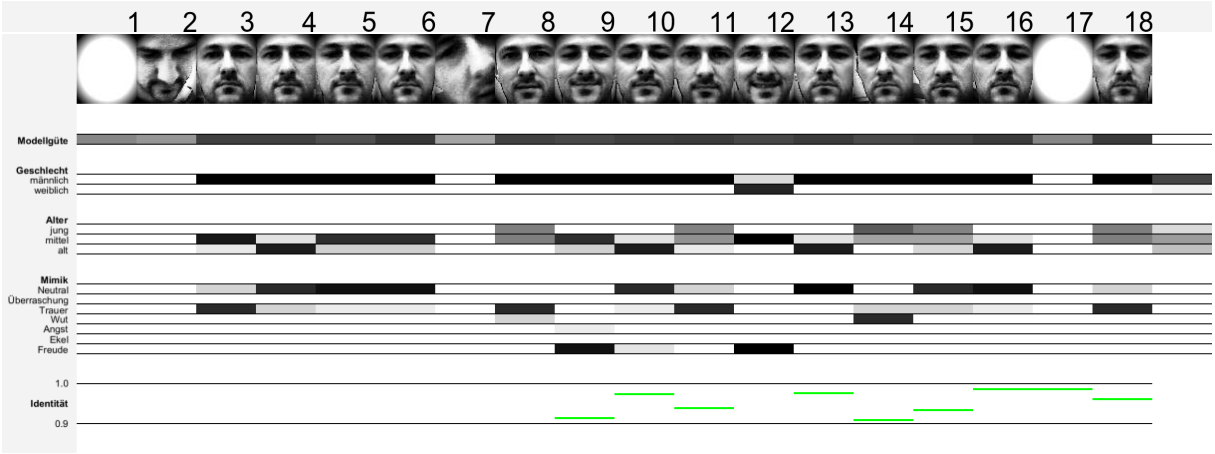


Abbildung 5.3: *Person P1 (männlich, 31), Sequenzlänge 164s. In der Datenbank enthalten war nur Person P1. Die Modellgüte ist in den Bildern 1,2, 7 und 17 zu niedrig für eine weitere Auswertung. Bei Bild 1 und 17 handelt es sich um eine Fehldetektion des Gesichtsdetektors, in Bild 2 liegt eine zu starke out-of-plane-Rotation des Gesichtes vor und in Bild 7 wurde ein Auge an der falschen Position detektiert. Das Geschlecht wurde bis auf Bild 12 richtig geschätzt. Die Altersschätzung ist relativ sprunghaft. Der Gesichtsausdruck wird mit Neutral, Trauer und in Bild 9 und 12 richtigerweise mit Freude geschätzt. Ab Bild 9 lag die Ähnlichkeit mit dem am Anfang der Sequenz erstellten Modell meist über der Schwelle von* $0,9$.

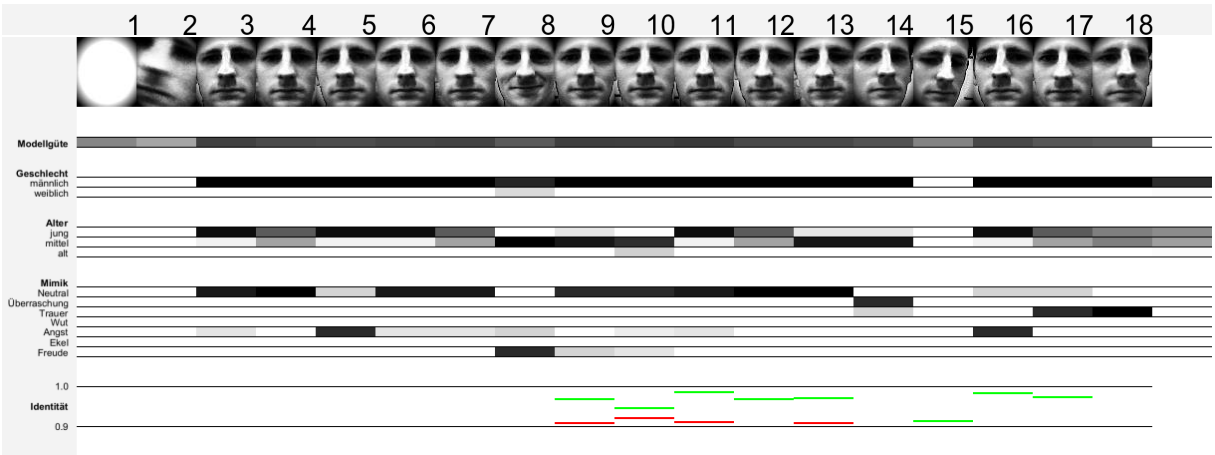


Abbildung 5.4: *Person P2 (männlich, 25), Sequenzlänge 168s. In der Datenbank enthalten waren Person P1 und P2. Das Geschlecht wurde auch in dieser Sequenz immer korrekt geschätzt. Die Altersschätzung schwankt zwischen jung und mittel, die Person liegt mit 25 Jahren auch genau an der Grenze zwischen der Klasse jung und mittel, vgl. Abbildung 4.41. Der Gesichtsausdruck wird meist korrekt auf Neutral geschätzt. Für die Bilder 5 und 16 wird fälschlicherweise Angst geschätzt, wobei auffällt, dass die Person in beiden Bildern leicht nach links schaut. In Bild 8 wird richtigerweise Freude geschätzt. Die Identität wird ab Bild 9 fast immer erkannt. In den Bildern 9, 10, 11 und 13 liegt auch Person P1 über der Ähnlichkeitsschranke von* $0,9$ *jedoch immer hinter Person P2.*

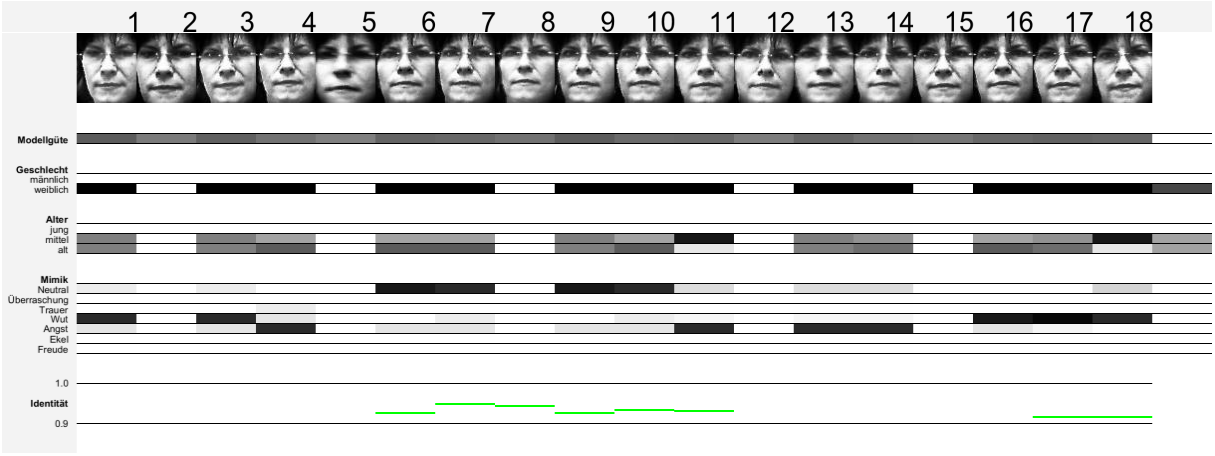


Abbildung 5.5: *Person P3 (weiblich, 42), Sequenzlänge 103s. In der Datenbank enthalten waren Person P1, P2 und P3. In dieser Sequenz wird das Geschlecht immer korrekt auf weiblich geschätzt. Das Alter schwankt zwischen mittel und alt, wobei die Grenze zwischen diesen beiden Klassen bei 45 Jahren liegt. Neben dem neutralen Gesichtsausdruck wird häufig Wut und Angst geschätzt. Die Identität wird ab Bild 6 mit einer Lücke von Bild 12 bis 16 immer richtig bestimmt.*

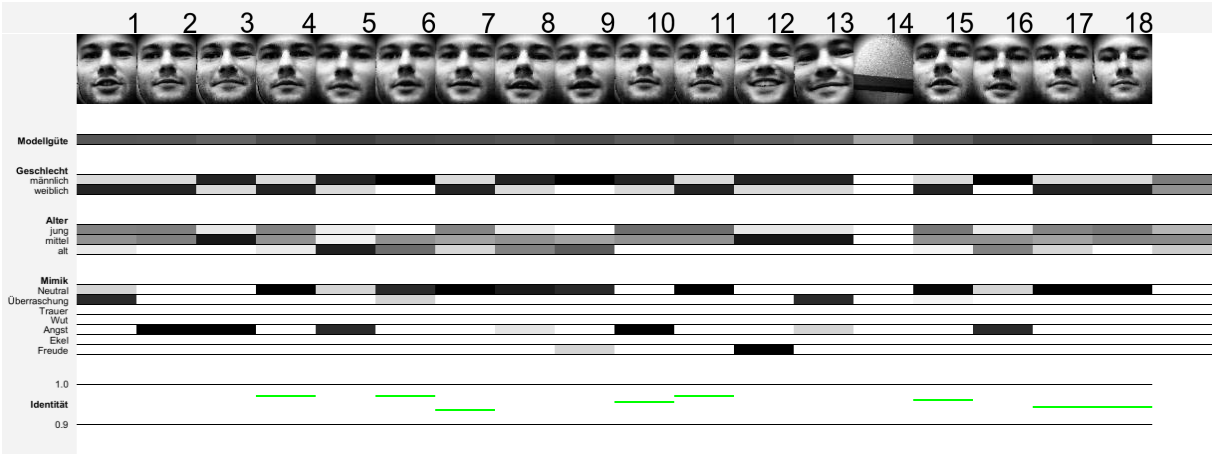


Abbildung 5.6: *Person P4 (männlich, 29), Sequenzlänge 119s. In der Datenbank enthalten waren Person P1, P2, P3 und P4. Bei dieser Person wird in 8 von 18 Bildern das Geschlecht fälschlicherweise mit weiblich bestimmt. In der zeitlichen Integration liegt die Schätzung bei männlich. Auch die Altersschätzung schwankt relativ stark, liegt aber im Mittel korrekterweise bei einem mittleren Alter. Der Gesichtsausdruck wurde meist korrekt mit neutral und in Bild 12 ebenfalls korrekt mit Freude geschätzt. In 5 Bildern wurde fälschlicherweise Angst geschätzt, was sich zumindest für Bild 2 und 3 durch ein leichtes Lächeln erklären lässt, dass durch die zurückgezogenen Mundwinkel oft mit Angst verwechselt wird. Die Identität wurde auch hier immer richtig bestimmt, andere Personen haben die Ähnlichkeitsschwelle nicht überschritten.*

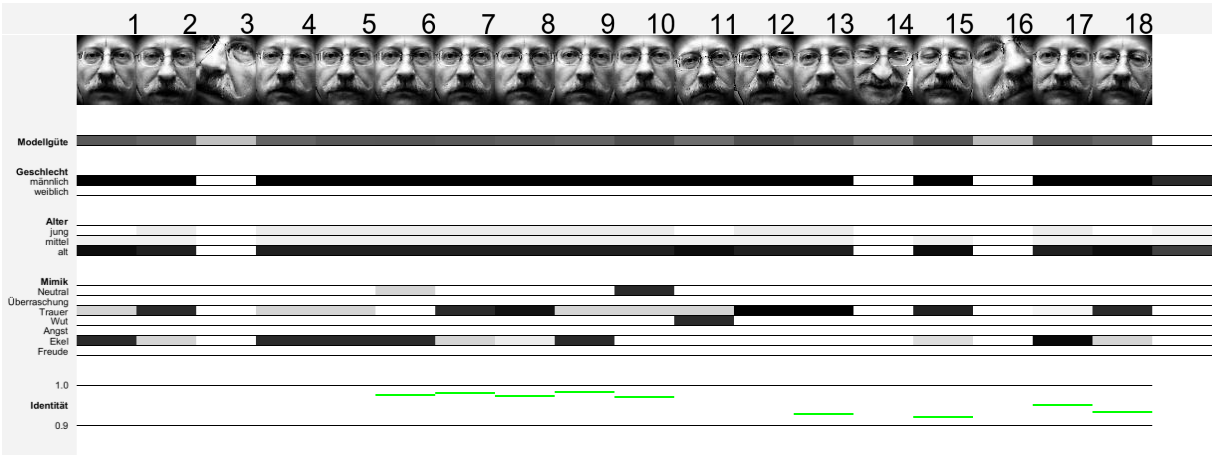


Abbildung 5.7: *Person P5 (männlich, 51), Sequenzlänge 141s. In der Datenbank enthalten waren Person P1, P2, P3, P4 und P5. In den Bildern 3 und 16 wurde jeweils ein Auge an der falschen Position detektiert. In dieser Sequenz wird das Geschlecht immer auf männlich und das Alter auf alt geschätzt. Der Gesichtsausdruck dagegen wird häufig auf Trauer bzw. Ekel geschätzt. Es ist zu vermuten, dass sich der starke Bartwuchs positiv auf die Geschlechts- und Altersschätzung auswirkt und eher negativ auf die Mimikschätzung. Insbesondere die Schätzung auf Trauer könnte durch die Ähnlichkeit des Bartes mit nach unten gezogenen Mundwinkeln entstehen. Die Identität wurde wiederum immer korrekt bestimmt, andere Personen haben die Ähnlichkeitsschwelle nicht überschritten.*

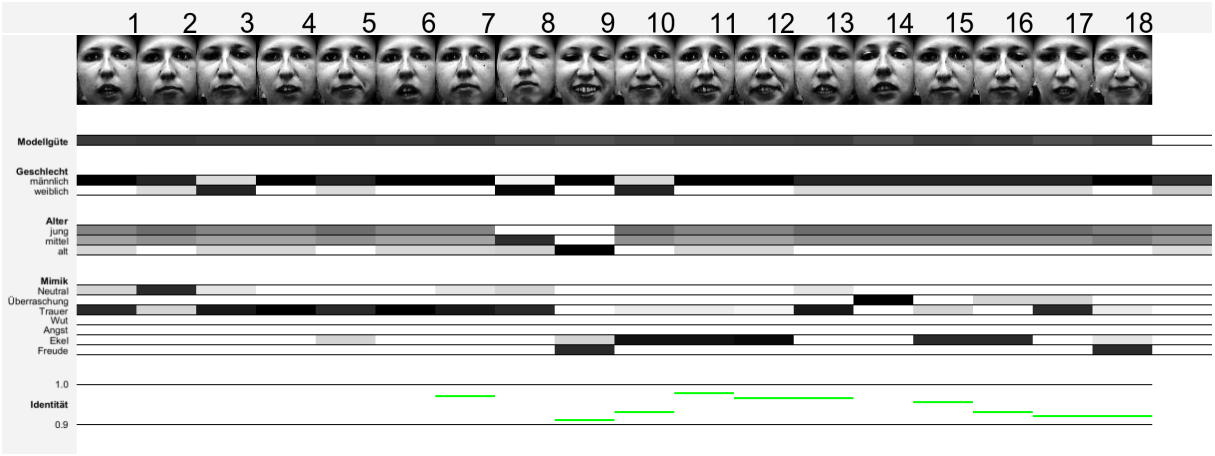


Abbildung 5.8: *Person P6 (weiblich, 27), Sequenzlänge 175s. In der Datenbank enthalten waren Person P1, P2, P3, P4, P5 und P6. In dieser Sequenz wird das Geschlecht fast immer falsch als männlich geschätzt. Das Alter wird korrekt auf jung bis mittel geschätzt. Der Gesichtsausdruck wird in Bild 9 korrekt als Freude und in Bild 14 als Überraschung eingestuft. Am Beginn der Sequenz wird der Gesichtsausdruck hauptsächlich als traurig eingestuft. Die Schätzung des Gesichtsausdrucks auf Ekel in den Bildern 10 bis 12 könnte aufgrund der etwas stärker gerümpften Nase zustande kommen. Die Identität wird ab Bild 7 immer richtig erkannt.*

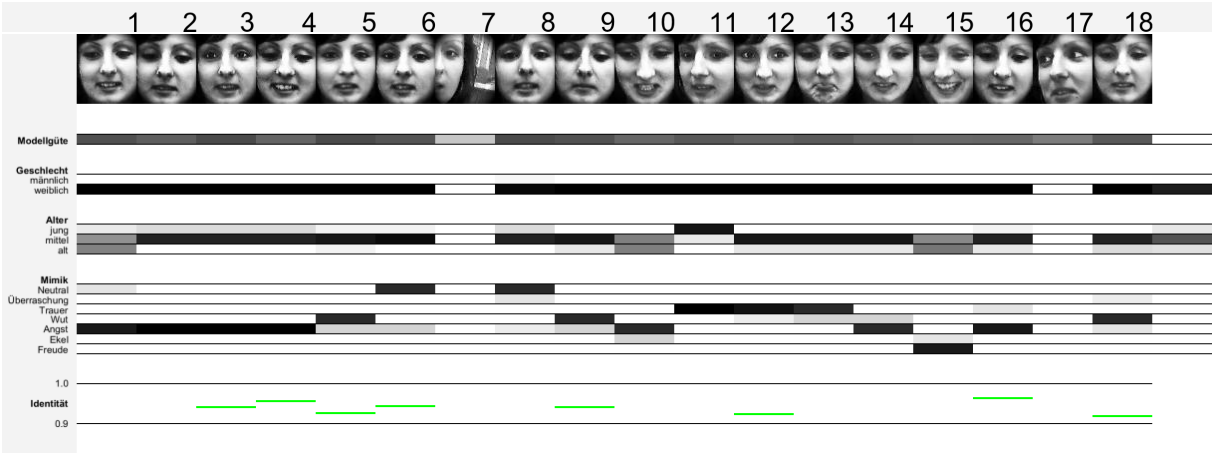


Abbildung 5.9: *Person P7 (weiblich, 23), Sequenzlänge 129s. In der Datenbank enthalten waren Person P1, P2, P3, P4, P5 und P7. Das Geschlecht wird immer richtig geschätzt. Obwohl die Person mit 23 Jahren unter der Grenze zwischen jung und mittel liegt, wird das Alter meistens auf mittel geschätzt. In den Bildern 1 bis 4 und 10 ist der Mund leicht geöffnet, wodurch die Klassifikation des Gesichtsausdrucks als Angst erklärt werden kann. Bild 15 wurde korrekt als Freude klassifiziert.*

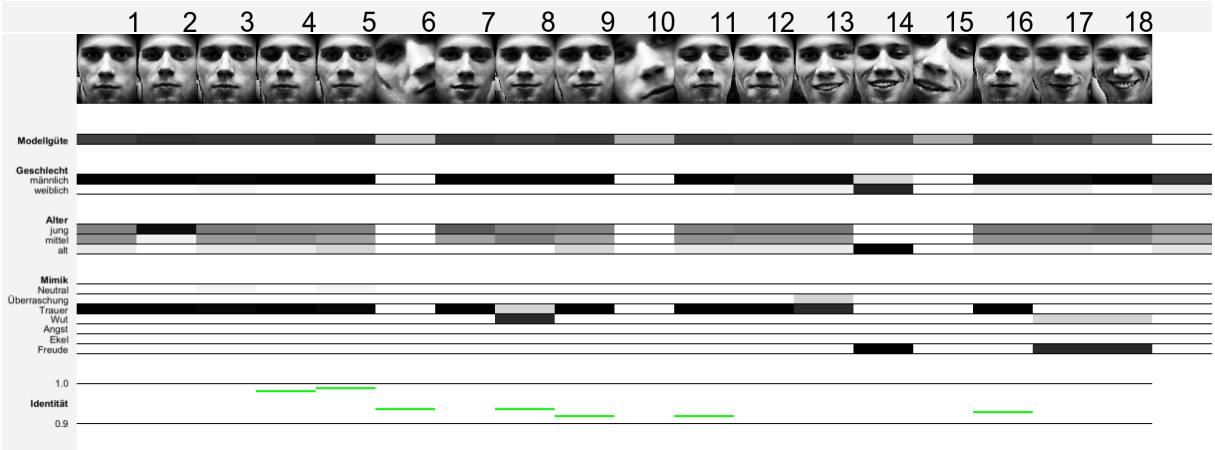


Abbildung 5.10: *Person P8 (männlich, 26), Sequenzlänge 109s. In der Datenbank enthalten waren Person P1, P2, P3, P4, P5 und P8. Die Schätzungen für Geschlecht und Alter sind wiederum meistens korrekt. Der Gesichtsausdruck wird in den Bildern 14, 17 und 18 korrekt als Freude klassifiziert, ansonsten aber oft falsch als Trauer.*

An dieser Stelle soll ein Experiment vorgestellt werden, bei dem sich eine Person beim Roboter anmeldet und sich dann von diesem entfernt, ohne sich abzumelden. Damit bleibt sie als aktueller Nutzer eingebucht und andere Personen haben nicht die Möglichkeit, die Dienste des Roboters in Anspruch zu nehmen.

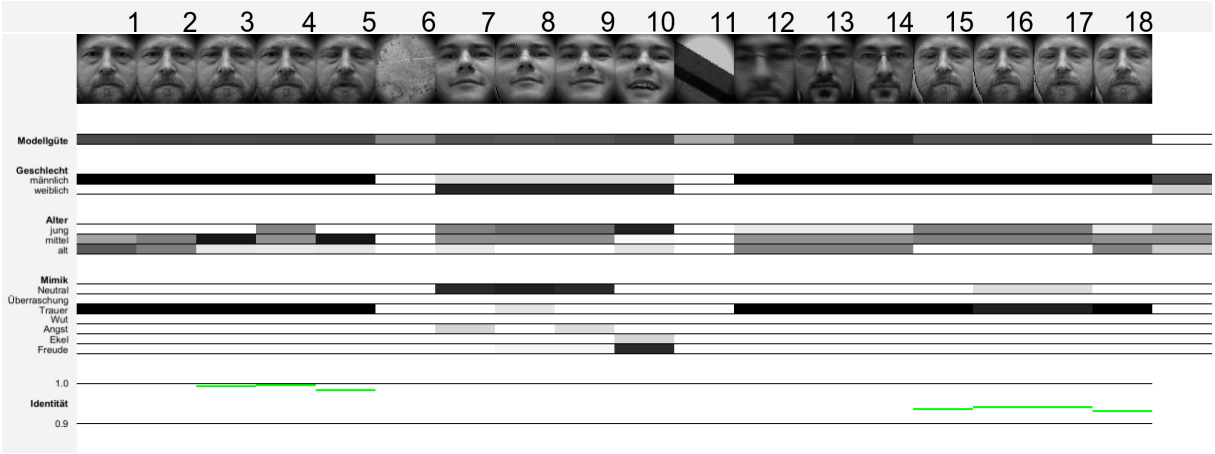


Abbildung 5.11: *Sequenzlänge 138s. Person A wird am Anfang der Sequenz als aktueller Nutzer in die Datenbank aufgenommen. In den Bildern 7 bis 10 wird Person B analysiert. Es wird erkannt, dass es sich nicht um den aktuellen Nutzer handelt. Ebenso überschreitet die Ähnlichkeit des für Person A hinterlegten Modells mit den Modellen für die Bilder 12 bis 14 (Person C) nicht die Ähnlichkeitsschwelle von* $0,9$*. In den Bildern 15 bis 18 interagiert wieder Person A mit dem Roboter und wird dabei auch als Person A erkannt. Der Gesichtausdruck von Person A und C wird immer als traurig eingestuft, was auf den Bartwuchs bei beiden Personen zurückzuführen sein könnte.*

5.3.3 Fazit

Die eben vorgestellten Experimente erlauben zwar keine quantitative Aussage über die Eignung des Gesamtsystems für eine Anwendung im Baumarkt. Trotzdem können einige Schlussfolgerungen getroffen werden. Zum einen hat sich die Personenidentifikation als sehr robust erwiesen, so dass diese sehr wichtige Funktion bereits auf dem Shopping-Assistenten angewendet werden konnte. Die Geschlechts und Altersschätzung liefern in der Regel stabile Ausgaben, es kann aber natürlich nicht garantiert werden, dass einige Personen falsch klassifiziert werden. Ein Beispiel hierfür ist die Schätzung des Geschlechts von Person P6.

Weiterhin auffällig war eine realtiv starke Abhängigkeit der Altersschätzung von der Mimik. Beispiele hierfür sind Bild 9 bei Person P6. Die Mimik wurde hier korrekt mit Freude klassifiziert, die Klassifikation des Alters sprang aber für dieses Bild zu 100% auf alt. Ein ähnliches Verhalten ist für Bild 10 und 15 bei Person P7 und für Bild 14 bei Person P8 zu beobachten. Diese starke Abhängigkeit der Altersschätzung vom Gesichtsausdruck war aus den Ergebnissen auf dem abgeschlossenen Datensatz aus Abschnitt 4.7.2.2 nicht zu erwarten. Eine Mögliche Lösung für dieses Problem besteht in einer hierarchischen Klassifikation, bei der zunächst die Mimik geschätzt wird und daraufhin entweder die Alterschätzung bei bestimmten Gesichtsausdrücken deaktiviert wird oder aber spezialisierte Klassifikatoren für das Alter eingesetzt werden.

Kapitel 6

Alternative Arbeiten

In diesem Kapitel werden verschiedene aktuelle Arbeiten, die in den Themenbereich dieser Dissertationsschrift fallen, vorgestellt und jeweils daraufhin untersucht, inwieweit sie den gestellten Anforderungen an eine natürliche Interaktion zwischen Mensch und Maschine gerecht werden. Dabei müssen ganz unterschiedliche Forschungsbereiche berücksichtigt werden. Zum einen sind dies aktuelle Projekte aus dem Bereich Servicerobotik und soziale Robotik, aber zum anderen auch Arbeiten aus dem Bereich Human-Computer-Interaction (HCI), bei denen nicht mit mobilen Roboterplattformen gearbeitet wird, die aber ebenfalls eine natürliche und intuitive Kommunikation zwischen Mensch und Maschine anstreben.

Für die Vorstellung der Projekte wird folgendes Schema verwendet:

PROJEKTNAME	Projektzeitraum cortex.informatik.tu-ilmenau.de

ENTWICKLER
Literaturverweise
SENSORIK: Sensorik
AUFMERKSAMKEIT: Aufmerksamkeitssystem
NUTZERANALYSE: Nutzeranalyse

Bei der Aufzählung wird nur auf die für die Interaktion wichtigen Teilaspekte eingegangen. Wenn es zu einem Projekt besonders interessante Aspekte zu benennen gibt, werden diese im Anschluss an das Schema aufgeführt. Naturgemäß unterscheiden sich die vorgestellten Projekte hinsichtlich der Mensch-Maschine-Interaktion sehr stark. Trotzdem lassen sich einige Trends erkennen, die am Ende dieses Kapitels aufgezeigt werden sollen.

6.1 Serviceroboter

In diesem Abschnitt werden einige Projekte aus dem Bereich Servicerobotik vorgestellt, die aus Sicht der Mensch-Maschine-Interaktion als interessant erscheinen. In dieser Rubrik finden sich Tourguides und Pflegeroboter, aber auch Unterhaltungsroboter. Da der Schwerpunkt bei der Entwicklung dieser Roboter auf ganz unterschiedlichen Teilaspekten lag, wie z.B. auf der Navigation bei Tourguides oder einem möglichst natürlichen Bewegungsablauf bei den Humanoiden, haben diese Roboter ein sehr unterschiedliches Äußeres und sehr unterschiedliche Funktionalitäten für die Mensch-Roboter-Interaktion.

Hermes	2002 www.unibw-muenchen.de/hermes/

Universität der Bundeswehr in München
[Bischoff and Graefe, 2002]
Sensorik: Active-Vision mit zwei Kameras, Sprachein- und -ausgabe
Aufmerksamkeit: Gesichtsdetektion
Nutzeranalyse: -

Care-O-BotII	2003 www.morpha.de/php_d/morpha_Projekt.php3

Morpha Konsortium, Deutschland
[Ehrenmann et al., 2001][Graf et al., 2004]
Sensorik: Spracheingabe, Touch-Display
Aufmerksamkeit: -
Nutzeranalyse: -

Rhino	2000 www.informatik.uni-bonn.de/~rhino/tourguide/

Universität Bonn, Carnegie Mellon University
[Burgard et al., 1999]
Sensorik: Stereokamera, Sonar- und Infrarot-Sensoren, Laser-Scanner, Touch-Display
Aufmerksamkeit: Personendetektion (Laser-basiert)
Nutzeranalyse: -

Minerva	2000 www-2.cs.cmu.edu/~minerva/

Carnegie Mellon University
[Thrun et al., 1999][Schulte et al., 1999]
Sensorik: Laser-Scanner, Kopf, Gesicht
Aufmerksamkeit: Personendetektion (Laser-basiert)
Nutzeranalyse: -

Colin	1999 www-ra.informatik.uni-tuebingen.de/forschung/service/welcome_e.html

Universität Tübingen
[Feyrer and Zell, 1999]
Sensorik: Active-Vision-Kameras, Sonarsensoren, Laser-Scanner
Aufmerksamkeit: Personendetektion und kontinuierliches Personen-Tracking mittels Hautfarbe, Bewegung, Kontur und Stereoinformationen; Eigenbewegungskompensation
Nutzeranalyse: -

Wakamaru	2004 www.sdia.or.jp/mhikobe-e/products/etc/robot.html

Mitsubishi
-
Sensorik: omnidirektionale Kamera, Frontalkamera, Mikrofone, taktile Sensoren, IR- und Ultraschallsensoren, Sensoren zur Ermittlung des Neigungsgrades, Spracherkennung (10000 Wörter)
Aufmerksamkeit: Personendetektion (Bewegung, Gesichter, Wärmequellen, Geräusche)
Nutzeranalyse: Personenerkennung (10 Personen)

Ifbot	2003 www.business-design.co.jp/product/001/index.html

Business Design Laboratory, University of Nagoya (Japan)
-
Sensorik: Spracherkennung (10000 Wörter)
Aufmerksamkeit: -
Nutzeranalyse: Personenerkennung (10 Personen), Emotionserkennung aus Stimme und Vokabular

PaPeRo	2003 http://www.incx.nec.co.jp/robot/

NEC

-

Sensorik: zwei CCD-Kameras, ein Ultraschall-Sensor, zwei Mikrophone, Spracherkennung

Aufmerksamkeit: Personendetektion: Stereovision, Template-Matching, Soundlokalisation

Nutzeranalyse: Personenerkennung

SIG2	? winnie.kuis.kyoto-u.ac.jp/SIG/oldsig/

Kitano Symbiotic Systems (Japan)

[Okuno et al., 2002]

Sensorik: zwei frontal ausgerichtete CCD-Kameras, mehrere Mikrofone

Aufmerksamkeit: Gesichtsdetektion, Sound-Quellen-Lokalisation, Stereo-Vision, Kombination von hautfarb- und korrelationsbasierter Gesichtsdetektion

Nutzeranalyse: Gesichtserkennung und Sprecheridentifikaktion

RoboVie	2004 www.irc.atr.co.jp/~m-shiomi/Robovie/index.html

Kitano Symbiotic Systems (Japan)

[Shiomi et al., 2004] [Littlewort et al., 2003]

Sensorik: omnidirektionale Kamera, Pan-Tilt-Kameras mit Zoom, Mikrophon, Ultraschallsensoren, taktile Sensoren

Aufmerksamkeit: Hautfarb- und Bewegungsdetektion auf Bild der omnidirektionalen Kamera

Nutzeranalyse: Mimikanalyse in dem mit der Frontalkamera aufgenommenen Gesicht

ASIMO	2004 world.honda.com/ASIMO/

Honda

-

Sensorik: Kamera, Spracherkennung (100 Wörter)

Aufmerksamkeit: Bewegungsdetektion, Erkennung der Bewegungsrichtung, Verfolgen von Bewegungen, Geräuschquellenlokalisation

Nutzeranalyse: Posen- und Gestenerkennung, Gesichtserkennung (bis 10 Personen)

QRIO	2003 www.sony.net/SonyInfo/QRIO/

Sony

-

Sensorik: zwei Kameras, sieben Mikrophone, Spracherkennung (20000 Wörter)
Aufmerksamkeit: Stereobildverarbeitung, Gesichtsdetektion, Geräuschquellenlokalisation
Nutzeranalyse: Gesichtserkennung, Unterscheidung von Personen anhand der Stimme

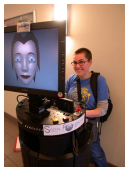
Es existiert noch eine Vielzahl anderer Serviceroboter, zu denen aber leider keine detaillierten Informationen im Hinblick auf die Mensch-Maschine-Kommunikation verfügbar waren. Hierzu zählen: Der ToyotaRobot (Toyota), die Roboter Wendy und Wamoeba (Intelligent Machine Lab, Waseda University in Japan), Isamu (Kawada Industries), ISAC (Vanderbilt University), AMI (Korea Institute of Science and Technology) und der Interbot (Navy Center for Applied Research in Artificial Intelligence, USA).

Bei einigen Projekten, vor allem aus dem kommerziellem Sektor, wurden die Angaben hinsichtlich Aufmerksamkeitssystem und Nutzeranalyse aus Beschreibungen auf Internetseiten entnommen. Diese Angaben sind, sofern keine Referenz auf eine wissenschftliche Publikationen angegeben wurde, nicht belegbar und es ist auch oft nicht nachvollziehbar, mit welchen Methoden diese Leistungen erbracht werden.

6.2 Soziale Roboter

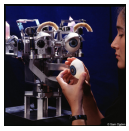
Im Gegensatz zu Servicerobotern dienen die hier vorgestellten Roboter nicht dazu, dem Menschen in irgendeiner Weise behilflich zu sein, sondern ausschließlich dazu, mit ihm in eine Interaktion zu treten. Da diese Robotertypen speziell für die Interaktion entwickelt werden, sind ihre Kommunikationsmöglichkeiten in der Regel deutlich besser ausgeprägt.

Vikia	2004 www.cs.cmu.edu/afs/cs/project/robocomp/social/www/vikia.html



Robotics Institute, Human Computer Interaction Institute (Carnegie Mellon University)
[Bruce et al., 2002]
Sensorik: Spracherkennung
Aufmerksamkeit: visuelle Informationsverarbeitung
Nutzeranalyse: -

KISMET	2000 www.ai.mit.edu/projects/humanoid/~robotics/~group/kismet/kismet.html



MIT
[Breazeal and Scassellati, 2000] [Breazeal, 1999]
SENSORIK: Stereo Active-Vision-Kopf (zwei Weitwinkelkameras, zwei hochauflösende Kameras), Mikrofon (am Interaktionspartner)
AUFMERKSAMKEIT: Auffälligkeitssystem auf Weitwinkelbildern, Bewegungsdetektion, Entfernungsschätzung, Personen-Tracking
NUTZERANALYSE: Gesichtsdetektion

COG	- www.ai.mit.edu/projects/humanoid/~robotics/~group/cog/cog.html

MIT
[Varshavskaya, 2002][Aryananda, 2002]
SENSORIK: Stereo Active-Vision-Kopf (zwei Weitwinkelkameras, zwei hochauflösende Kameras), Spracherkennung
AUFMERKSAMKEIT: kontinuierliches Tracken
NUTZERANALYSE: Gesichtsdetektion, Augendetektion

LEONARDO	2004 robotic.media.mit.edu/projects/Leonardo/Leo-intro.html

MIT
[Breazeal et al., 2004]
SENSORIK: Stereo Active-Vision-Kopf (zwei Weitwinkelkameras, zwei hochauflösende Kameras), Spracherkennung
AUFMERKSAMKEIT: kontinuierliches Verfolgen von Objekten (Farbe, Form und Bewegung), Personendetektion (Hautfarbe)
NUTZERANALYSE: Augendetektion zum Herstellen von Augenkontakt, Tracken von Gesichtsmerkmalen, Gestenerkennung

ROBITA	- www.pcl.cs.waseda.ac.jp/robita/

PERCEPTUAL COMPUTING GROUP, WASEDA UNIVERSITY, TOKYO
[Tojo et al., 2000]
SENSORIK: CCD-Kamera, Mikrophone, Spracherkennung
AUFMERKSAMKEIT: Geräuschlokalisation
NUTZERANALYSE: Gesichtserkennung, Kopfposenerkennung

6.3 HCI-Projekte

Interessante Forschungsarbeiten findet man nicht nur im Bereich Mensch-Roboter-Interaktion, sondern auch im Bereich Mensch-Computer-Interaktion. In den folgenden Abschnitten sollen deshalb auch solche Projekte aufgeführt werden, die zwar nicht aus der Robotik kommen, aber durchaus relevant für diese Arbeit sind. Dabei geht es in der Regel um eine Verbesserung der Interaktion zwischen Mensch und Computer durch die Entwicklung natürlicher und intuitiver Schnittstellen für die Dialogführung. Die Darstellung erfolgt hier aufgrund des stark variierenden Charakters der Systeme nicht tabellarisch, sondern jeweils durch eine kurze verbale Beschreibung

Emotions in HCI	- www.iha.bepr.ethz.ch/research/groups/manmachine/design/emotionsinhci

ETH Zürich

[Zimmermann et al., 2003]

Ziel ist es, das Verhalten einer Software adaptiv in Hinblick auf die Gemütsverfassung des Nutzers zu gestalten. Das Verfahren beruht auf den Auswirkungen des Gemütszustandes einer Person auf deren motorisches Verhalten, wie z.B. die Anzahl der Anschläge auf der Tastatur oder die Anzahl der Mausklicks pro Minute und die Bewegungen der Maus. Es werden also keine Sensoren am Körper des Nutzers angebracht, sondern die beim Umgang mit dem Computer üblichen Eingabegeräte verwendet.

SmartKom	2003 www.smartkom.org/

SmartKom-Konsortium: DFKI GmbH, DaimlerChrysler AG, Philips GmbH, Siemens AG, Universität Stuttgart

[Wahlster, 2002][Reithinger et al., 2003]

Dieses Projekt hatte das Ziel der Erforschung und Entwicklung selbsterklärender, benutzeradaptiver Schnittstellen für die Mensch-Technik-Interaktion. Es ging um eine intelligente Verknüpfung verteilter Informationsquellen und verschiedenartiger Kommunikationsdienste und eine robuste Verarbeitung von möglicherweise ungenauen, mehrdeutigen oder teilweise inkorrekten Eingaben. Als Sensorik wurden eine prosodische Analyse, eine Spracherkennung, graphische Bedienoberflächen, Gestik und Gesichtsausdrücke eingesetzt. Im Ergebnis des Projektes wurden die Prototypen SmartKOM-Public (Multimodale Kommunikationszelle), SmartKOM-Mobil (mobiler Kommunikationsassistent) und SmartKOM-Home/Office (intuitives Arbeiten mit dem Computer) vorgestellt. Teilweise wurde mit eingeschränkten Umgebungsbedingungen gearbeitet, z.B. erfolgte die Gestenerkennung mit Hilfe einer Infrarotkamera und einem über der Projektionsfläche eines LCD-Videoprojektors angebrachten Infrarotsender.

COMIC	2005 www.hcrc.ed.ac.uk/comic/

COMIC-Konsortium: Max Planck Institut für Psycholinguistik, Max Planck Institut für Biologische Kybernetik, University of Nijmegen, University of Sheffield, University of Edinburgh, DFKI, ViSoft GmbH

[ten Bosch et al., 2004]

Ziel dieses Projektes war die Entwicklung generischer kognitiver Modelle für die multimodale Interaktion, die Fusion verschiedener Eingabemodalitäten, dem Dialog- und Aktionsmanagement und der Koordinierung verschiedener Ausgabemodalitäten. Als Eingabemodalitäten kamen dabei Sprache und Stift zum Einsatz. Es wurden Untersuchungen zur zwischenmenschlichen Interaktion und eine quantitative Analyse des zeitlichen Verhaltens beim Wechsel von Sprecher und Zuhörer durchgeführt. Normalerweise gibt es keine Verzögerungen bei solchen Wechseln. Wenn aber eine Person die andere nicht sieht, kommt es zu längeren Pausen und zu einer signifikant langsameren Kommunikation. Der fließende Wechsel von Sprecher und Zuhörer ist ein Ergebnis des Bestrebens der Gesprächspartner, die Pausen zwischen den Turns zu minimieren und dabei Überschneidungen zu vermeiden. Aus diesen Beobachtungen wurde auf die Wichtigkeit von visuellen Informationen, insbesondere der Wahrnehmung des Gesichtes des Gesprächspartners, für die Konversation geschlossen. Lippenbewegungen tragen zum Verständnis von undeutlichen Äußerungen bei und dienen auch dem Erkennen des Endes einer sprachlichen Äußerung. Mit dem Gesicht erfolgt nicht nur der Ausdruck von Emotionen wie Vergnügen, Überraschung oder Angst, sondern auch nicht-verbale Rückmeldungen über den Dialogfluss wie Verwirrung, Verständnis, Zustimmung oder die Betonung und die Unterstreichung von Aussagen.

Im Rahmen des Projektes erfolgte eine Aufnahme einer Vielzahl von Gesichtsausdrücken von mehreren Personen und eine experimentelle Bestimmung, wie glaubwürdig diese Ausdrücke waren. Mit diesen Kenntnissen wurde ein Avatar mit für Sprachausgaben synchronisierten Lippenbewegungen entwickelt, der solche menschlichen Gesichtsausdrücke nachahmen kann. Die Gesichtsausdrücke dieses Avatars konnten leicht erkannt werden, selbst wenn der Kontext der Konversation nicht verfügbar war.

Auch wenn im COMIC-Projekt keine Wahrnehmung des Nutzers erfolgte, macht es die Wichtigkeit des visuellen Kanals bei der Mensch-Mensch- und der Mensch-Maschine-Interaktion deutlich und unterstützt die These, dass solche Komponenten für eine natürliche Interaktion unabdingbar sind.

6.4 Fazit

In diesem Kapitel wurden die Interaktionsmöglichkeiten einiger aktueller Serviceroboter, sozialer Roboter und einiger HCI-Systeme vorgestellt.

Die meisten Serviceroboter verwenden relativ simple Mechanismen für die Mensch-Maschine-Interaktion. Personen werden in der Regel über entfernungsmessende Sensoren oder visuell über Hautfarbe oder Bewegung wahrgenommen. Nur wenige Systeme gehen bei der Bildverarbeitung über eine Gesichtsdetektion hinaus. Neuere Systeme, die ihre Einsatzfelder im Haushalts- oder Pflegebereich haben, setzen verstärkt auf natürlichere Interaktionsmöglichkeiten. So kommen neben Verfahren zur Gesichtserkennung auch Spracherkennung und Sprachsynthese zum Einsatz.

Naturgemäß finden sich die fortschrittlichsten Mensch-Maschine-Schnittstellen bei sozialen Robotern, wobei hier der Schwerpunkt in der Regel auf der Aktuatorik liegt. Roboter wie Kismet und Leonardo können sehr überzeugende Gesichtsausdrücke erzeugen, die visuelle Wahrnehmung

beschränkt sich aber in der Regel auf die Detektion und das kontinuierliche Verfolgen von Gesichtern oder Gesichtsmerkmalen wie Augen. Der Grund hierfür ist, dass diese Roboter für ihren Interaktionspartner möglichst natürlich wirken sollen. Im Vordergrund stehen daher in der Regel eine realistische Darstellung von Gesichtsausdrücken und vielfältige und angemessene Sprachausgaben. Die Wahrnehmung des Nutzers beschränkt sich zumeist auf eine Gesichtsdetektion, Entfernungsschätzung, eine Detektion der Augen, um Augenkontakt herzustellen zu können und eine Gesichtserkennung.

Bei den HCI-Projekten steht dagegen die Wahrnehmung des Nutzers im Vordergrund. Hier spielen intuitive Eingabemodalitäten wie Gesten eine Rolle oder die Erkennung von Stresssituationen beim Nutzer. Im Gegensatz zu den Robotikanwendungen wird hier allerdings oftmals eine Anpassung der Umgebungsbedingungen vorgenommen oder auf spezielle Eingabegeräte wie Tastatur und Maus zurückgegriffen.

Kapitel 7

Zusammenfassung und Ausblick

7.1 Zusammenfassung

In Kapitel eins dieser Arbeit wurden die Komponenten zwischenmenschlicher Kommunikation identifiziert und den typischerweise bei der Mensch-Roboter-Interaktion realisierten Komponenten gegenübergestellt. Dabei wurde herausgestellt, dass die Beziehungsebene der Kommunikation, die in der zwischenmenschlichen Kommunikation im Vergleich zur Sachebene den weitaus größeren Anteil einnimmt, entweder ganz außer Acht gelassen, oder nur teilweise berücksichtigt wird. Die Servicerobotik beschränkt sich häufig auf die reine Übertragung von Kommandos und Statusinformationen. Die Nutzerwahrnehmung ist in der Regel auf eine Personendetektion, ein -tracking und evtl. auf eine Identifikation beschränkt. Die Entwicklung sozialer Roboter widmet sich zwar der Beziehungsebene der Kommunikation, vernachlässigt dabei aber die Wahrnehmung von Informationen über den Nutzer. Soziale Roboter sollen auf den Nutzer einen intelligenten Eindruck machen und möglichst menschenähnlich wirken. Die Wahrnehmungsmechanismen dieser Roboter orientieren sich an biologischen Vorbildern, gehen aber auch nicht über Detektion, Tracking und Identifikation von Personen hinaus. Damit wurde eine verbesserte Wahrnehmung des Zustandes des Nutzers in der Robotik als Zielstellung dieser Arbeit definiert.

Beim konkreten Anwendungsfall handelt es sich um einen Shopping-Assistenten, der in einem Baumarkt den Kunden bei der Suche nach Produkten behilflich ist. Damit sollte die Interaktion zum einen für uneingewiesene Benutzer verständlich sein, zum anderen sollte der Roboter aber auch einen gewissen Umfang an sozialer Kompetenz zeigen, indem er z.B. Personen in seiner Umgebung anspricht und während der Interaktion kontinuierlich Blickkontakt hält. Um Nutzermodelle erstellen, kurzzeitig verlorene Nutzer wiedererkennen und den Gemütszustand des Nutzers abschätzen zu können, sollen Geschlecht, Alter, Identität und Gesichtsausdruck des Nutzers aus einem Bild ermittelt werden.

Für die Realisierung dieser Aufgabe wurde eine biologisch motivierte Aufteilung in ein peripheres und ein foveales Vision-System gewählt. Das periphere System arbeitet auf den Bildern einer omnidirektionalen Kamera und verfügt damit über einen sehr großen Sichtbereich, aber eine kleine Auflösung. In diesem System werden Hypothesen über die Position von Personen im Umfeld

des Roboters gebildet. Dafür werden Hautfarbe, Bewegung und Entfernung in einer Auffälligkeitskarte integriert und auffällige Bildbereiche mittels eines Multi-Target-Trackers verfolgt. Für die omnidirektionale Kamera wurde ein automatischer Weißabgleich entwickelt, der die Hautfarbdetektion unempfindlich gegen Änderungen der Beleuchtungsfarbe macht, so dass selbst mit einem einfachen nicht-adaptiven Farbmodell ein robustes Personen-Tracking ermöglicht wird.

Wurde eine Nutzerhypothese ausgewählt, wird der Kopf des Roboters kontinuierlich in die entsprechende Richtung ausgerichtet. Damit erhält der Nutzer zum einen eine Rückmeldung über die Aufmerksamkeit des Roboters während der Interaktion. Zum anderen kann der Roboter hochaufgelöste Bilder der Person aufnehmen, so dass eine Analyse durch das foveale Vision-System ermöglicht wird. Diese ist wiederum in zwei Teilschritte unterteilt, der Erzeugung einer normalisierten Darstellung des Gesichtes und der Analyse selbst. Der erste Schritt besteht aus einer Detektion des Gesichtes und einer anschließenden Detektion der Augen. Für die Realisierung wurden drei Gesichtsdetektionsverfahren implementiert und vergleichend untersucht. Dabei konnte gezeigt werden, dass die Verfahren von Rowley und Viola vergleichbare Erkennungs- und Falsch-Positiv-Raten liefern, das letztgenannte jedoch aufgrund seiner kaskadierten Klassifikatoren wesentlich schneller in der Abarbeitung ist. Folglicherweise kommt es auch für die anschließende Augendetektion zum Einsatz. Anhand der gefundenen Punkte wird das Gesicht schließlich in eine normalisierte Darstellung gebracht.

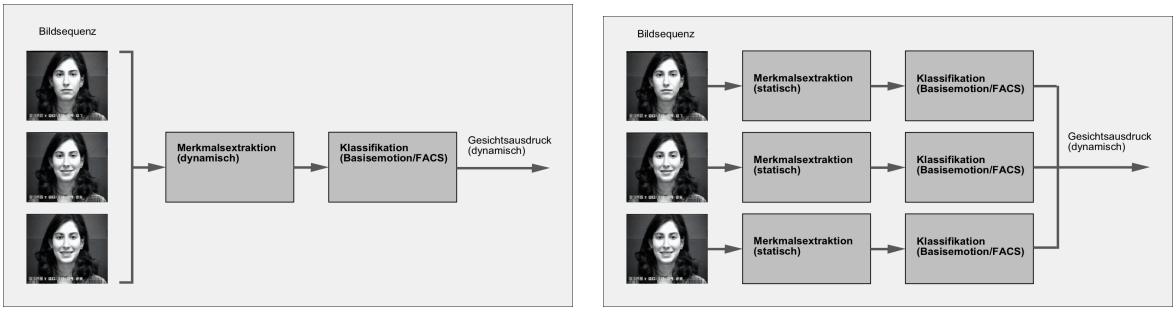
Für den Analyseschritt wurden das Elastic-Graph-Matching (EGM), die Independent-Component-Analysis (ICA) und die Active-Appearance-Modelle (AAM) implementiert und vergleichend untersucht. Unter Berücksichtigung der Anforderungen einer Geschlechts-, Alters-, Mimik- und Identitätsschätzung wurde hierfür eine umfassende Gesichtsdatenbank zum Training und zum Test der Verfahren angelegt. Die Gesichtsausdrücke werden in dieser Datenbank durch emotionale Klassen beschrieben.

Aufgrund der normalisierten Darstellung des Gesichtes konnte beim EGM auf die Schritte zur Grobpositionierung und Skalierung des Graphen verzichtet werden. Stattdessen wurde der Local-Move so adaptiert, dass die statistischen Variationen der Knotenpositionen in den Trainingsdaten für die Bestimmung der Größe des Suchbereichs verwendet werden. Die Klassifikation erfolgt hier anhand eines Vergleichs des Bildgraphen mit den Graphen der General-Face-Knowledge. Bei der ICA wurde die Darstellung im Bildraum und die Darstellung im Pixelraum bei jeweils unreduzierter und reduzierter Anzahl von Basisbildern gegenübergestellt. Dabei schnitt fast ausschließlich die Bildraumdarstellung mit unreduzierter Anzahl an Basisbildern am besten ab. Bei den Active-Appearance-Modells wurden die kombinierten und die unabhängigen Appearance-Models gegenübergestellt, wobei kein eindeutiger Vorteil eines Modelltyps festgestellt werden konnte. Bei der abschließenden Gegenüberstellung der drei Verfahren schnitten die Active-Appearance-Models bei der Geschlechts-, Alters- und Mimikschätzung in Kombination mit einem MLP als Klassifikator am besten ab. Die Independent-Component-Analysis war sehr gut in Kombination mit Nearest-Neighbor und LVQ-Klassifikatoren. Trotz der Überlegenheit der Active-Appearance-Models fiel die Wahl des eingesetzten Verfahrens bei der Integration zu einem Gesamtsystem auf die ICA, da diese in der Anwendung um den Faktor 100 schneller arbeitet als die AAMs.

Die Funktionsfähigkeit des integrierten Systems wurde schließlich anhand von Experimenten demonstriert. Dabei bewegten sich Personen frei vor dem Roboter, wurden durch das periphere Vision-System erfasst und getrackt und das Gesicht wurde durch das foveale Vision-System analysiert. Dabei erwieß sich die Personenidentifikation als sehr robust und für den Einsatz auf einem Shopping-Assistenten geeigent. Die Geschlechts- und Altersschätzung funktioniert in der Regel robust, es gibt aber naturgemäß Personen, für die eine diesbezügliche Klassifikation falsche Ergebnisse liefert. Dies liegt nicht zuletzt darin begründet, dass den untersuchten Verfahren nur Informationen über das Gesicht und nicht z.B. über Frisur oder Kleidung einer Person zur Verfügung stehen. Es ist kritisch zu hinterfargen, wie mit solchen Problemen beim Einsatz des Systems auf dem Shopping-Assistenten umzugehen ist. Bei der Mimikanalyse hat sich gezeigt, dass stark ausgeprägte Gesichtsausdrücke in der Regel richtig erkannt werden können, wie z.B. Lachen. Probleme gibt es z.B. mit Bärten, die offensichtlich dazu führen, dass der neutrale Gesichtsausdruck eher als Trauer eingestuft wird.

7.2 Ausblick

Peripheres Vision System Das Problem der Detektion und des Trackings von Personen wurde durch die Kombination der Merkmale Hautfarbe, Bewegung und Entfernung gelöst. Trotz der weitgehenden Toleranz dieses Systems gegen Änderungen in der Chrominanz der Beleuchtung durch die Verwendung eines automatischen Weißabgleichs treten bei manchen Realweltbedingungen Probleme auf. Dies ist in der Regel dann der Fall, wenn die Ausprägung der, der Detektion und dem Tracking zugrunde liegenden, Merkmale kleiner wird, z.B. wenn sich die Person zu weit von der Kamera entfernt. Da der CONDENSATION-Algorithmus nur Verteilungen mit einer bestimmten Mindestgröße tracken kann, werden diese in der Regel verloren, obwohl sie im Panoramabild noch gut sichtbar sind. Eine anderes Problem stellt die Beleuchtungsrichtung dar. Bei starkem Gegenlicht nimmt die Farbsättigung im Bild so stark ab, dass Hautfarbe selbst für den menschlichen Betrachter kaum noch als solche wahrnehmbar ist. Eine mögliche Lösung dieses Problems besteht in einer lokalen Helligkeitsanpassung im Panoramabild. Mögliche weiterführende Arbeiten könnten sich außerdem damit beschäftigen, das periphere Vision-System spezifischer für Personen zu machen. Es wäre z.B. denkbar, mit einer höher auflösenden Kamera und einem entsprechend guten Objektiv auch im Panoramabild Gesichter zu detektieren. Auf diese Weise könnten viele Falsch-Positive aussortiert werden, ohne dass sie zunächst mit der Frontalkamera angeschaut werden müssten. Für die Berechnung der Sample-Wichtungen könnte in diesem Fall die Anzahl der Kaskaden eines kaskadierten Gesichtsdetektors verwendet werden. D.h. je mehr Kaskaden ein Bildausschnitt durchläuft, bevor er aussortiert wird, desto höher könnte die Wichtung des zugehörigen Samples sein [Shakhnarovich et al., 2002].



(a) (b)

Abbildung 7.1: *Erkennung dynamischer Gesichtsausdrücke (a) mit bewegungsbasierter Merkmalsextraktion (b) mit statischer Merkmalsextraktion. Auch mit den in dieser Arbeit verwendeten statischen Methoden der Merkmalsextraktion ist es möglich, dynamische Aspekte von Gesichtsausdrücken zu erfassen. Dazu muss allerdings in den extrahierten statischen Merkmalen die Intensität des Gesichtsausdrucks hinterlegt werden, wie dies z.B. beim FACS der Fall ist.*

Foveales Vision System Mit dem Gesichtsdetektor aus [Viola and Jones, 2004] wird eine robuste Detektion bei variablen Beleuchtungsbedingungen erreicht. Wie in der Einleitung zu Abschnitt 3.2 erläutert, existieren aber weiterhin offene Fragestellungen. Verschiedene Blickwinkel, Gesichtsausdrücke, Verdeckungen, Beleuchtungsbedingungen und Bärte oder Brillen stellen immer noch ein Problem dar.

Wie im Abschnitt 4.3.2 bereits angemerkt wurde, stellt die Kodierung von Gesichtsausdrücken durch emotionale Klassen nur einen Kompromiss dar, da im Rahmen dieser Arbeit keine FACS-Kodierung der Daten möglich war. Durch eine solche Kodierung durch visuelle Klassen wäre eine wesentlich feinere Unterscheidung bei der Klassifikation möglich, wobei auch die Stärke der Ausprägung von Deformationen berücksichtigt werden könnte. Auf diese Weise und bei hinreichend schneller Abarbeitung könnten auch dynamische Aspekte von Gesichtsausdrücken erfasst werden, wobei weiterhin mit einer statischen Merkmalsextraktion gearbeitet werden könnte, siehe Abbildung 7.1.

Eine wesentliche Schwäche der verwendeten Analyseverfahren besteht in der Notwendigkeit frontaler Ansichten. Die Sensitivität ist dabei bei der ICA am größten, da hier kein Modell verwendet wird, dass sich an Änderungen in der Ansicht des Gesichtes anpassen könnte. Beim EGM kann der Graph bis zu einem gewissen Grad auch an leicht out-of-plane rotierte Gesichter angepasst werden. Das selbe gilt in stärkerem Maße auch für AAMs. Für stärkere Deformationen sind jedoch andere Ansätze notwendig, wie z.B. die dreidimensionalen Modelle aus [DeCarlo and Metaxas, 1997].

Die Gesichtsanalyse war im Rahmen dieser Arbeit auf eine Schätzung von Geschlecht, Alter, Gesichtsausdruck und Identität beschränkt. Sicherlich gibt es eine Vielzahl weiterer Informationen, die für eine intuitive Mensch-Maschine-Schnittstelle von Interesse sind. An erster Stelle zu nennen wäre hierbei die Blickrichtungsschätzung, die dem Roboter Auskunft darüber geben kann, worauf der Nutzer seine Aufmerksamkeit richtet. Mit den Active-Appearance-Models existiert bereits eine Grundlage für die Realisierung die-

ser Klassifikationsaufgabe, denn die Blickrichtung kann als Formvariation das Gesichtes aufgefasst und durch das Formmodell modelliert werden. Hierfür wäre lediglich ein nochmaliges Training eines Active-Appearance-Models mit einem entsprechenden Datensatz notwendig. Wenn Kopfposen bis hin zu Profilansichten erfasst werden sollen, bietet sich eine Kombination mehrerer Appearance-Models an [Cootes et al., 2000].

Aufgrund der Testergebnisse aus Abschnitt 4.7 wird den Active-Appearance-Models das größte Potential bei der Analyse von Gesichtern zugesprochen. Ein nicht zu vernachlässigender Nachteil der AAMs ist jedoch der relativ hohe Zeitbedarf bei der Modellanpassung. Bei weiterführenden Arbeiten sollte verstärkt Augenmerk auf die Optimierung im Hinblick auf die Rechenzeit gelegt werden. Dass AAMs sogar in Echtzeit angewendet werden können, zeigen [Matthews and Baker, 2003] und [Dornaika and Ahlberg, 2004].

Integration Mit dieser Arbeit wurden die Grundlagen für die Erstellung von Nutzermodellen und für die Auswertung von Gesichtsausdrücken geschaffen. Im Gesamtsystem tatsächlich eingesetzt, wird bis jetzt aber nur die Personenidentifikation. In weiterführenden Arbeiten soll ein Dialogmanager entwickelt werden, siehe Abbildung 5.1, der die ermittelten Informationen nutzt und den Iteraktionsprozess entsprechend anpasst.

Anhang A

Anhang

A.1 Bestimmung der Reglerparameter für den automatischen Weißabgleich

Abbildung A.1 zeigt die Struktur des digitalen Regelkreises für den automatischen Weißabgleich.

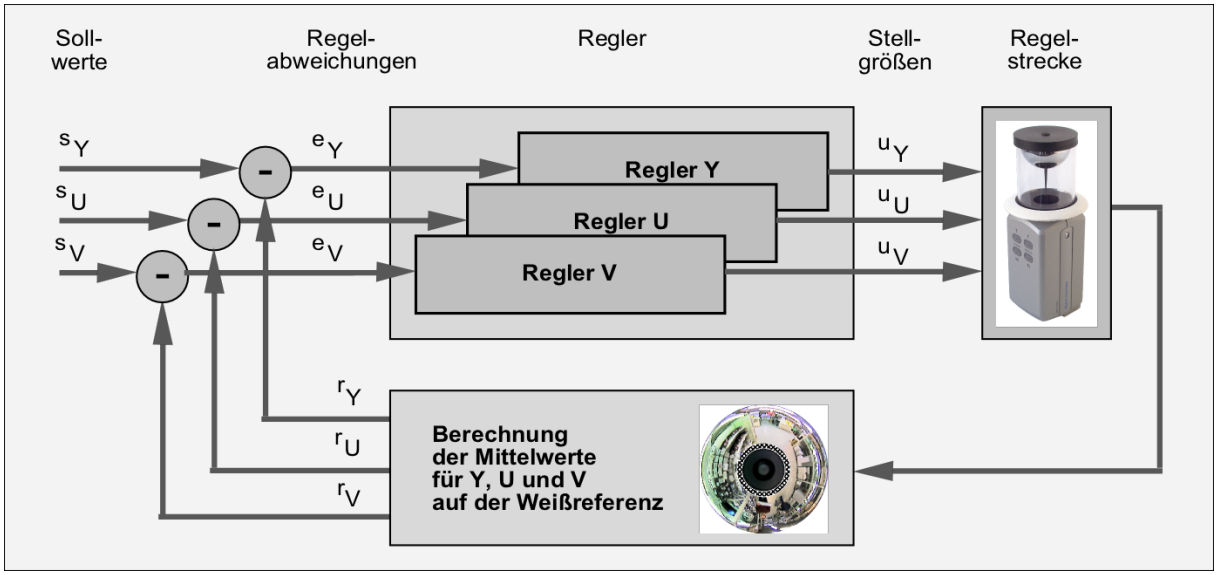


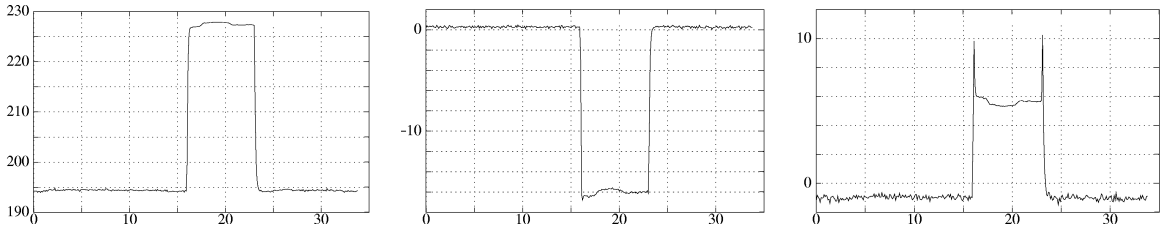
Abbildung A.1: *Struktur des digitalen Regelkreises für den automatischen Weißabgleich. Aus den Mittelwerten für Y, U und V aller Pixel, die auf der Weißreferenz liegen und den Sollgrößen $s_U = 0$, $s_V = 0$ und $s_Y = 0$ werden die Regelabweichungen e_U, e_V und e_Y berechnet. Drei separate PID-Regler berechnen die Stellgrößen für den Weißabgleich u_U, u_V und die Iris u_Y der Kamera.*

A.1.1 Regelstrecke

Um die Regler dimensionieren zu können, muss das Verhalten der Regelstrecke bekannt sein. Die zu regelnden Größen sind die Mittelwerte für Y, U und V aller Pixel, die auf der Weißreferenz

liegen. Die Sollwerte sind $U = 0$ und $V = 0$, da diese im YUV-Farbraum für weiß stehen. Da es nicht möglich ist, die Regelstrecke analytisch zu modellieren, muss experimentell ein Modell bestimmt werden. Zu diesem Zweck wurde die Sprungantwort $h(t)$ für die Mittelwerte der Komponenten Y, U und V aller Pixel auf der Weißreferenz aufgezeichnet, siehe Abbildung A.2. Dazu wurde das System durch An- und Ausschalten des Lichtes mit einem Sprung beaufschlagt. Die Abtastfrequenz betrug $9Hz$. Für die Erstellung eines Systemmodells wurde angenommen, dass es sich bei den Regelstrecken $G_S(s)$ um PT2-Glieder handelt mit der Übertragungsfunktion:

$$G_S(s) = \frac{K}{(1+T_1 s)(1+T_2 s)} \tag{A.1}$$



(Y) (U) (V)

Abbildung A.2: *Sprungantworten für die Mittelwerte von Y, U und V auf der Weißreferenz bei An- und Ausschalten des Lichtes. Die Messwerte sind über die Zeit in Sekunden aufgetragen.*

Hierzu wurden die Messwertkurven auf eine Höhe von 1 normiert, da die Verstärkung des Systems nur bestimmt werden kann, wenn die Höhe des Eingangssprunges bekannt ist, was bei einem Anschalten des Lichtes nicht der Fall ist. Es ergeben sich die Zeitkonstanten aus Tabelle A.1(a).

A.1.2 Regler

Nachdem ein Modell für die Regelstrecke existiert, werden im nächsten Schritt die Regler dimensioniert. Für PT2-Strecken erscheinen normale PID-Regler als geeignet, siehe Gleichung A.2. Die PID-Regler werden so eingestellt, dass ihre Nullstellen die Polstellen der Strecke kompensieren. Die Übertragungsfunktion eines PID-Reglers $G_R(s)$ ist:

$$G_R(s) = \frac{K}{1 + \frac{1}{T_N s} + T_V s} \tag{A.2}$$

Damit ergeben sich die Reglerparameter aus Tabelle A.1. Die zeitdiskrete Realisierung des PID-Reglers ist:

$$u(t) = Ke(t) + K\frac{T}{T_N}\sum_{i=0} te(i) + K\frac{T_V}{T}\left(e(t) - e(t-1)\right) + u_0 \tag{A.3}$$

A.1.3 Reglerverstärkungen

Für die Bestimmung der Reglerverstärkungen wurde die Verstärkung eines einzelnen Reglers variiert, während die der anderen Regler fest auf einen kleinen Wert eingestellt wurden. Dabei wurde das System mit Sprüngen beaufschlagt und eine Verstärkung gesucht, bei der das System möglichst schnell reagierte, aber stabil blieb. Im Gesamtsystem mit allen drei Reglern können die Verstärkungswerte so nicht verwendet werden, da sonst Instabilitäten auftreten. Deshalb wurden die Reglerverstärkungen sukzessive verkleinert, bis das Gesamtsystem stabil blieb. Die letztendlich gewählten Verstärkungsfaktoren sind in Tabelle A.1(b) abgebildet.

Strecke	K	T_1	T_2
Y	0.99408	0.09218	0.09218
U	1.00443	0.08243	0.08243
V	0.99134	0.04025	0.04025

(a)

Regler	K	T_N	T_V
Y	6.0	0.184	0.046
U	0.4	0.1648	0.0412
V	0.6	0.0805	0.0201

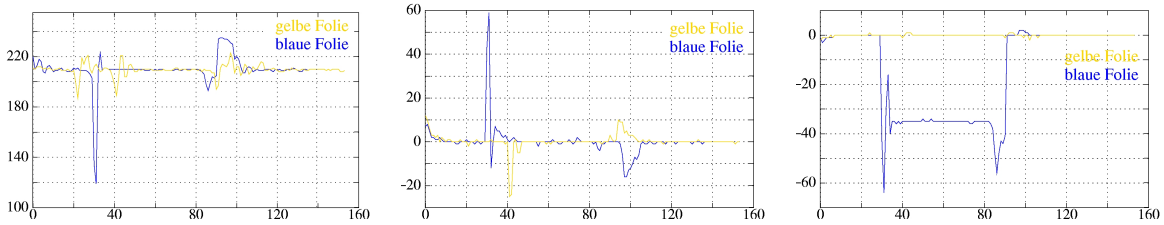
(b)

Tabelle A.1: *Zeitkonstanten und Verstärkungsfaktoren des Systems. (a) der Regelstrecken und (b) des Reglers.*

A.1.4 Ergebnisse

Abbildung A.3 zeigt die Auswirkungen des automatischen Weißabgleichs auf die Hautfarbdetektion. Die Stellgrößen für den Weißabgleich sind direkt nach dem Einschalten des Systems mit Vorgabewerten belegt. Nach nur zehn Bildern (rund 1 Sekunde) ist die Regelabweichung nahezu Null. Für einen Test der Stabilität des Gesamtsystems wurde die Kamera jeweils mit einer gelben und einer blauen Folie zur Simulation von Farbstichen abgedeckt. Trotz der extremen Farbänderung blieb das System in beiden Fällen stabil, siehe Abbildung A.4.

Abbildung A.3: *Verbesserte Detektionen des Hautfarbdetektors durch den automatischen Weißabgleich. Es handelt sich um das erste, das fünfte und das zehnte Bild einer Sequenz. Neben der eigentlichen Hautfarbregion entstehen auch mehr falsch-positive Detektionen auf holzfarbenen Objekten.*



(Y) (U) (V)

Abbildung A.4: *Systemverhalten bei Abdeckung der Kamera mit Farbfolien. Bei der blauen Farbfolie konnte die Regelabweichung für V nicht ausgeregelt werden, da die Stellgröße u_V in die Sättigung lief.*

A.2 Panoramatransformation

Wie in Abschnitt 2.6 erwähnt, wird ein omnidirektionales Objektiv mit hyperbolischem Spiegel verwendet. Abbildung A.5 zeigt eine schematische Darstellung eines hyperbolischen Spiegels.

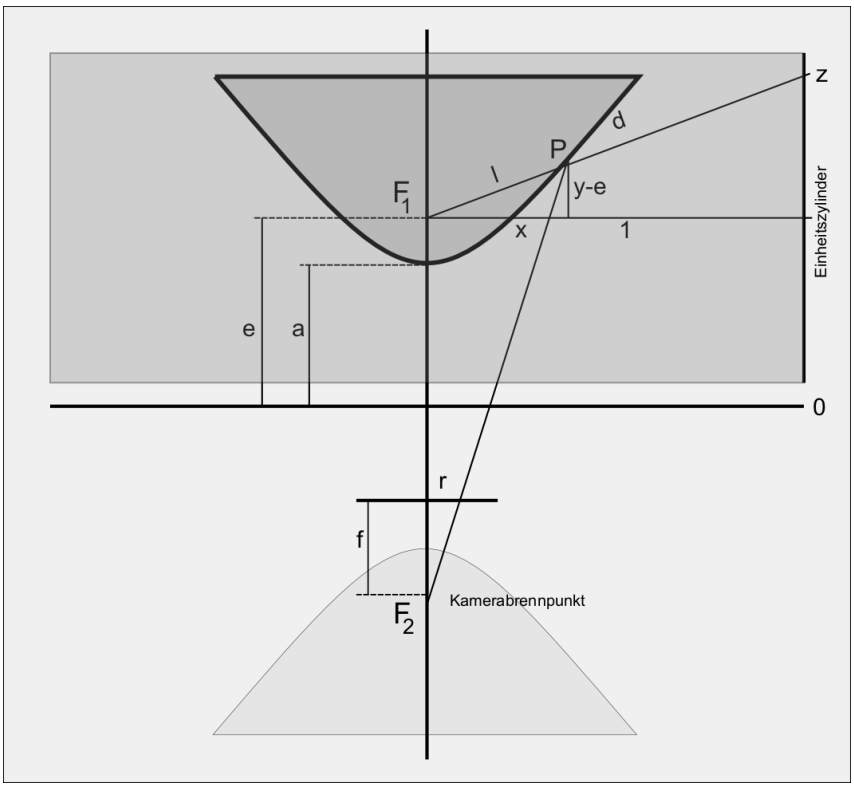


Abbildung A.5: *Schematische Darstellung eines hyperbolischen Spiegels.*

Im folgenden wird beschrieben, wie das omnidirektionale Bild in ein Panoramabild transformiert werden kann. Dies geschieht durch Projektion des Bildes auf einen virtuellen Einheitszylinder, der um den Brennpunkt des hyperbolischen Spiegels gelegt wird. Um die Transformation von einem omnidirektionalen Bild in ein Panoramabild durchführen zu können, wird ein Modell der Projektion am hyperbolischen Spiegel benötigt. Diese Projektion $\mathcal{H}$ soll zu jedem Punkt $p = (\phi, z)$ auf dem virtuellen Einheitszylinder seine Koordinaten $q = (\phi, r)$ im omnidirektionalen Bild liefern.

$$q = \mathcal{H}(p) \tag{A.4}$$

p wird dabei in Zylinder- und q in Polarkoordinaten angegeben. Für die Herleitung wird die Brennpunkteigenschaft der Hyperbel verwendet: Ein Strahl, der durch den Brennpunkt F_1 im hyperbolischen Spiegel verlaufen würde, wird so reflektiert, dass er durch den anderen Brennpunkt F_2 verläuft, in dem sich die Kamera befindet, siehe Abbildung A.5. Zunächst wird die Geradengleichung für den einfallenden Strahl über den Strahlensatz abgeleitet:

$$y = zx + e \tag{A.5}$$

Der Schnittpunkt $P(x, y)$ des einfallenden Strahles mit der Spiegeloberfläche ergibt sich wie folgt:

$$x = \frac{l}{d} \tag{A.6}$$

$$y = z\frac{l}{d} + e \tag{A.7}$$

Aus der Leitlinieneigenschaft für Hyperbeln und der Exzentrität $\varepsilon = \frac{e}{a}$ lässt sich folgende Beziehung für l ableiten:

$$l = \varepsilon y - a \tag{A.8}$$

damit kann Gleichung A.7 wie folgt formuliert werden:

$$\begin{aligned} y &= z\frac{\varepsilon y - a}{d} + e \\ &= \frac{ed - za}{d - \varepsilon z} \end{aligned} \tag{A.9}$$

Für x gilt entsprechend:

$$\begin{aligned} x &= \frac{y - e}{z} \\ &= \frac{\varepsilon e - az}{d - \varepsilon z} \end{aligned} \tag{A.10}$$

Der Schnittpunkt $P(x, y)$ kann nun durch

$$\frac{r}{f} = \frac{x}{y + e} \tag{A.11}$$

auf den CCD-Chip der Kamera projiziert werden. Durch Einsetzen von Gleichung A.9 und A.10 erhält man folgende Gleichung:

$$r = f\frac{k_1}{k_3\sqrt{z^2 + 1} - k_2 z} \tag{A.12}$$

mit den Konstanten: $k_1 = \varepsilon^2 - 1$, $k_2 = \varepsilon^2 + 1$ und $k_3 = 2\varepsilon$.

Damit kann die gesuchte Projektion wie folgt angegeben werden:

$$q = \mathcal{H}(p) = (\phi, f\frac{k_1}{k_3\sqrt{z^2+1} - k_2 z}) \tag{A.13}$$

Für jeden Pixel $p = (\phi, z)$ im Panoramabild (auf dem Einheitszylinder) kann nun das zugehörige Pixel $q = (\phi, r)$ im omnidirektionalen Bild berechnet und dessen Farbwert ins Panoramabild eingetragen werden. Um die Berechnung nicht für jedes Bild durchführen zu müssen, wird eine Look-Up-Tabelle verwendet. Gleichung A.13 enthält zwei Parameter, die durch eine geeignete Kalibrierung ermittelt werden müssen. Dabei handelt es sich um die Brennweite f der Kamera und die Exzentrität des hyperbolischen Spiegels ε. Für diese Kalibrierung wurde ein Testmuster verwendet, das in horizontaler und in vertikaler Richtung gleichabständige Kreuze zeigt. Durch die Variation von f kann die Skalierung des Panoramabildes festgelegt werden. ε muss so eingestellt werden, dass die Kreuze im Panoramabild in x- und y-Richtung gleichabständig abgebildet werden.

A.3 Bildvorverarbeitung

A.3.1 Histogrammausgleich

Beim Histogrammausgleich wird das Grauwerthistogramm des betrachteten Bildausschnittes nichtlinear gestreckt, so dass der gesamte zur Verfügung stehende Grauwertbereich genutzt und gleichzeitig annähernd eine Gleichverteilung der Grauwerte im Histogramm erreicht wird. Dieses Vorgehen führt zu einem verbesserten Kontrast und kompensiert Unterschiede, die bei der Aufnahme der Bilder aufgrund von unterschiedlichen Kameraempfindlichkeiten auftreten.

Beim Histogrammausgleich geht man von der Idee aus, daß der zur Verfügung stehende Grauwertbereich am besten ausgenutzt wird, wenn alle Grauwerte gleich oft vorkommen. Es wird also das Intensitätshistogramm H einer Gleichverteilung und somit das kumulative Histogramm H_c einer Gerade angenähert. Für die Erstellung des kumulativen Histogramms werden die Werte im Histogramm von links nach rechts aufsummiert. Es liefert also eine Aussage darüber, wie hoch der Anteil der Grauwerte im Bild unterhalb eines bestimmten Wertes ist:

$$\begin{aligned} H_c(k) &= \sum_{i=0}^{k} H(i) \\ k &= [0, K] \text{ Intensitätslevel (K=255 für 8bit Grauwertbilder)} \\ H(i) &- \text{ Intensitätshistogramm} \end{aligned} \tag{A.14}$$

Dieses kumulative Histogramm soll im Ergebnisbild einer Gerade entsprechen:

$$\begin{aligned} H_c(k') &= \frac{N}{I_{max}} k' \\ k' &= f(k) \\ N &- \text{ Anzahl der Pixel} \\ I_{max} &- \text{ Maximale Intensität (255)} \end{aligned} \tag{A.15}$$

Für k' ergibt sich somit folgender analytischer Ausdruck:

$$\begin{aligned} H_c(k') &= H_c(k) \\ \frac{N}{I_{max}} k' &= \sum_{i=0}^{k} H(i) \\ k' &= \frac{I_{max}}{N} \sum_{i=0}^{k} H(i) \\ &= \frac{I_{max}}{N} H_c(k) \end{aligned} \tag{A.16}$$

Das somit normierte kumulative Histogramm $H_c(k)$ kann nun als Look-Up-Table zur Bestimmung der Intensitäten im Ausgabebild benutzt werden:

$$\mathbf{I}_{output}(x, y) = k'(\mathbf{I}_{input}(x, y)). \tag{A.17}$$

Abbildung A.6 zeigt ein Beispiel für einen Histogrammausgleich.

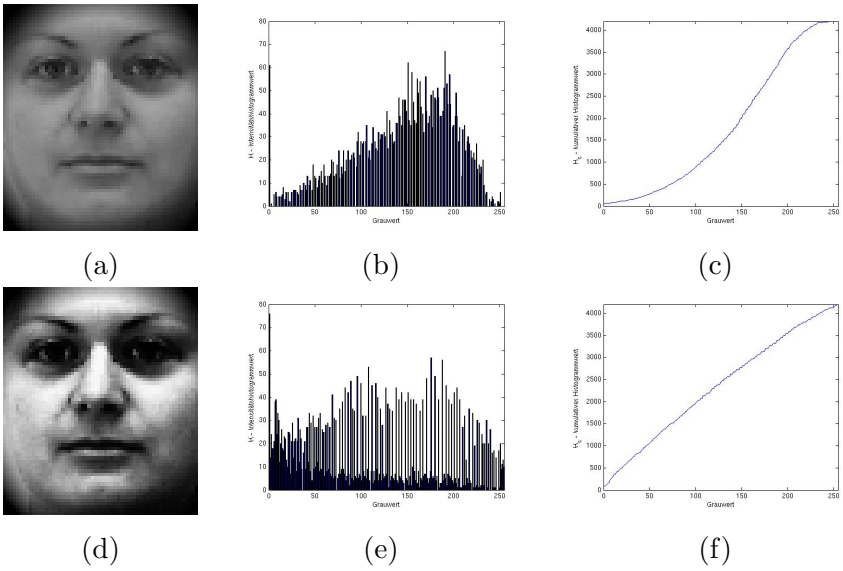


Abbildung A.6: *Beispielbild vor und nach dem Histogrammausgleich: (a) Ausgangsbild vor dem Histogrammausgleich, (b) dessen Histogramm und (c) dessen kumulatives Histogramm. (d) Ergebnisbild nach dem Histogrammausgleich, (e) dessen Histogramm und (c) dessen kumulatives Histogramm. Wie zu erkennen ist, liegen im Ergebnis die Grauwerte nahezu gleichverteilt vor, wodurch sich annähernd eine Gerade im kumulativen Histogramm ergibt.*

A.3.2 Intensitätsausgleich

Der Intensitätsausgleich dient dazu, eine ungleichförmige Ausleuchtung des Bildes zu beseitigen. Hierzu wird eine lineare Funktion berechnet, welche die durchschnittlichen Helligkeiten in den vier Quatranden des Bildausschnittes repräsentiert. Diese Funktion wird anschließend von den Intensitätswerten der Pixel im Fenster subtrahiert, so dass eine annähernd gleichmäßige Helligkeitsverteilung entsteht. Abbildung A.7 verdeutlich das Prinzip des Intensitätsausgleichs.

Aus den mittleren Grauwerten A, B, C und D in den vier Quadranten des Bildes wird eine Ebene berechnet:

$$m = 128 - \frac{A+B+C+D}{4} \tag{A.18}$$

$$r = \frac{2(B-A)}{s_x} \tag{A.19}$$

$$s = \frac{2(C-A)}{s_y} \tag{A.20}$$

$$t = \frac{4(D+A-B-C)}{(s_x s_y)} \tag{A.21}$$

dabei ist s_x und s_y die Bildgröße in x- bzw. in y-Richtung. Die Korrektur der Grauwerte des Eingangsbildes erfolgt dann wie folgt:

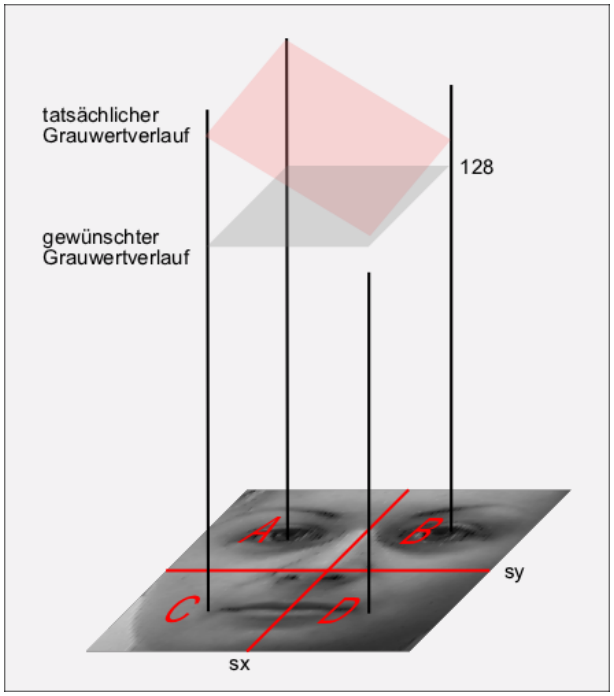
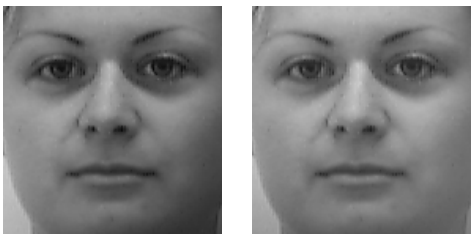


Abbildung A.7: *Prinzip des Intensitätsausgleichs.*

$$\mathbf{I}_{output}(x,y) = \mathbf{I}_{input}(x,y) - r\left(x - \frac{s_x}{2}\right) - s\left(y - \frac{s_y}{2}\right) - t\left(x - \frac{s_x}{2}\right)\left(y - \frac{s_y}{2}\right) + m \quad \text{(A.22)}$$

Abbildung A.8 zeigt ein Beispiel für einen Intensitätsausgleich.



(a) (b)

Abbildung A.8: *Beispielbild vor und nach dem Intensitätsausgleich.*

A.4 Integralbild

Das Integralbild $\mathcal{I}$ wird aus dem Eingabebild wie folgt berechnet:

$$\mathcal{I}(x,y) = \sum_{x' \leq x, y' \leq y} \mathbf{I}(x', y') \tag{A.23}$$

In Abbildung A.9 ist ein Eingabebild (a) und das zugehörige Integralbild (b) dargestellt.

Abbildung A.9: *(a) Eingabebild und (b) zugehöriges Integralbild. (c) Berechnung der Pixelsumme in einer Fläche im Integralbild.*

Mit dem Integralbild können die Pixelsummen in beliebigen Rechtecken des Originalbildes sehr einfach berechnet werden. Aus der Definition des Integralbildes ist leicht ersichtlich, dass für die Flächen der Rechtecke in Abbildung A.9 gilt:

$$\begin{aligned} P_1 &= F_1 + F_2 + F_3 + F_4 \\ P_2 &= F_2 + F_3 \\ P_3 &= F_3 \\ P_4 &= F_3 + F_4 \end{aligned} \tag{A.24}$$

Damit kann die Fläche des Rechtecks F_1 durch Addition von lediglich vier Werten berechnet werden:

$$F_1 = P_1 - P_2 + P_3 - P_4 \tag{A.25}$$

A.5 Parametrierung der Gabor-Filter

Die Fourier-transformierte eines zweidimensionalen Gabor-Wavelets ergibt einen gaußförmigen Bandpass im Realteil, wobei ein umgekehrt proportionaler Zusammenhang zwischen der Größe der Einhüllenden im Ortsraum und der Bandbreite im Frequenzraum besteht, siehe Gleichung 4.4. Durch die Wahl eines geeigneten Testmusters ist es möglich, die Orientierungs- und Frequenzabdeckung im Ortsraum zu überprüfen. Ein solches Muster wird in [Jähne, 1997] vorgestellt. Es besteht aus sinusförmigen konzentrischen Ringen, deren Wellenzahlen mit zunehmendem Abstand vom Zentrum zunehmen:

$$g(\mathbf{x}) = g_0 \sin\left(\frac{k_m|\mathbf{x}|^2}{2r_m}\right)\left(\frac{1}{2}\tanh\left[\frac{r_m - |\mathbf{x}|}{w}\right] + \frac{1}{2}\right) \tag{A.26}$$

Die Parameter sind r_m für den maximalen Radius des Musters, k_m für die maximale Wellenzahl am Rand. Der zweite Term realisiert einen weichen Übergang zwischen dem Ringmuster und der äußeren Kante, um Moiré-Effekte zu vermeiden, gesteuert durch den Parameter w. Das Muster enthält alle lokalen Orientierungen und ein großes Spektrum an Wellenzahlen. Abbildung A.10 zeigt das Testmuster und die zugehörige Antwort eines Gabor-Wavelets. Es ist gut zu erkennen, dass sich der Betrag der Antworten der Gabor-Wavelets nur langsam ändert, die Phase hingegen ändert sich schneller. Sie trägt zudem einen Sägezahn-Charakter, dessen Frequenz der des Gabor-Filters entspricht, mit dem das Bild gefaltet wurde.

Um eine gleichmäßige Abdeckung der Orientierungen und Frequenzen zu überprüfen, wurden die Beträge der Filterantworten aller verwendeten Filter in eine Abbildung eingezeichnet, wobei gleiche Frequenzen mit gleicher Farbe dargestellt wurden, die erste rot, die zweite grün, die dritte blau und die vierte wieder rot, siehe Abbildung A.10(f). Für die Darstellung wurden 8 Orientierungen und 4 Frequenzen verwendet. Die kleinste verwendete Wellenlänge beträgt 4 Pixel, da bei kleineren Wellenlängen die Filterantworten verschwindend klein werden. Für die Berechnung der Filtermasken wurden folgende Parameter eingestellt:

$$\theta(O) = O\frac{\pi}{8} + \frac{\pi}{8} \quad , 0 \leq O \leq 7 \tag{A.27}$$

$$\lambda = \frac{1}{\lambda_0 2^{F\Delta f}} \quad , 0 \leq F \leq 3 \tag{A.28}$$

$$\lambda_0 = 4\text{Pixel} \quad \text{(kleinste Wellenlänge)} \tag{A.29}$$

$$\Delta f = 0.5 \quad \text{(Halboktavabstand)} \tag{A.30}$$

$$b = 0.7 \quad \text{(Bandbreite)} \tag{A.31}$$

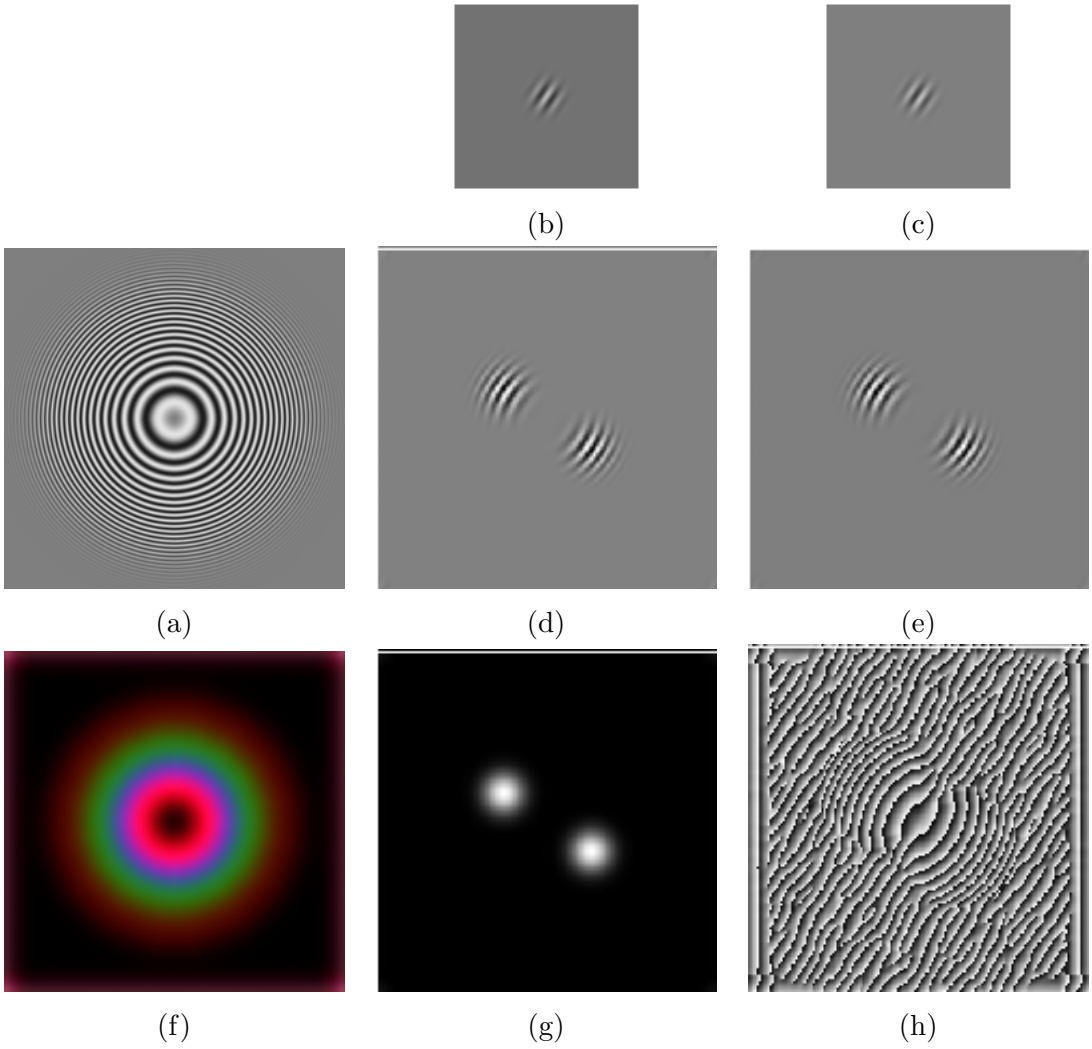


Abbildung A.10: *Antworten von Gabor-Wavelets auf einem synthetischem Testmuster. (a) Testmuster der Größe* 256×256 *mit einer Wellenzahl am Rand von* 0.25*. (b) und (c) Real- und Imaginärteil des Filters. (d) und (e) Real- und Imaginärteil der Filterantwort. (g) und (h) Betrag und Phase der Filterantwort. (f) Überlagerung der Beträge der Filterantworten aller* 32 *verwendeten Gabor-Wavelets. Die Antworten von Wavelets gleicher Frequenz werden mit gleicher Farbe dargestellt. Durch die Einstellung der Parameter entsprechend der Gleichungen A.27 bis A.31 wird eine gleichmäßige Abdeckung der Frequenzen und Orientierungen erreicht.*

A.6 Berechnung des Displacements

Als Ausgangspunkt der Berechnung des Displacements dient das phasenbasierte Ähnlichkeitsmaß. Dieses wird in seiner Taylor Approximation maximiert, indem nach $\mathbf{d}$ differenziert wird, siehe Gleichung A.32. Das Displacement gibt damit an, um welchen Betrag der Jet $\mathbf{j}'$ zu $\mathbf{j}$ verschoben werden muss, damit das Ergebnis der Approximation maximal wird. Wie auch beim phasenbasierten Ähnlichkeitsmaß kann über den *Fokus* festgelegt werden, welche Frequenzen verwendet werden sollen.

$$\mathcal{S}_\phi\left(\mathbf{j},\mathbf{j}'\right) \approx \frac{\sum_{n=1}^{N} a_n a'_n \left(1 - 0.5\left(\phi_n - \phi'_n - \mathbf{d}^T\mathbf{k}_n\right)^2\right)}{\sqrt{\sum_{n=1}^{N} a_n^2 \sum_{n=1}^{N} a'^2_n}} \tag{A.32}$$

Durch Setzen von $\frac{\partial \mathcal{S}_\phi}{\partial d_x} = 0$ und $\frac{\partial \mathcal{S}_\phi}{\partial d_y} = 0$ ergibt sich das Displacement mit der Bedingung $\gamma_{xx}\gamma_{yy} - \gamma_{xy}\gamma_{yx} \neq 0$ wie folgt:

$$\mathbf{d}\left(\mathbf{j},\mathbf{j}'\right) = \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \frac{1}{\gamma_{xx}\gamma_{yy} - \gamma_{xy}\gamma_{yx}} \begin{pmatrix} \gamma_{yy} & -\gamma_{yx} \\ -\gamma_{xy} & \gamma_{xx} \end{pmatrix} \begin{pmatrix} \varphi_x \\ \varphi_y \end{pmatrix} \tag{A.33}$$

mit den partiellen Ableitungen:

$$\varphi_x = \sum_{n=1}^{N} a_n a'_n k_{n\,x}\left(\phi_n - \phi'_n\right) \tag{A.34}$$

$$\varphi_y = \sum_{n=1}^{N} a_n a'_n k_{n\,y}\left(\phi_n - \phi'_n\right) \tag{A.35}$$

$$\gamma_{xx} = \sum_{n=1}^{N} a_n a'_n k_{n\,x} k_{n\,x} \tag{A.36}$$

$$\gamma_{xy} = \sum_{n=1}^{N} a_n a'_n k_{n\,x} k_{n\,y} \tag{A.37}$$

$$\gamma_{yx} = \sum_{n=1}^{N} a_n a'_n k_{n\,y} k_{n\,x} \tag{A.38}$$

$$\gamma_{yy} = \sum_{n=1}^{N} a_n a'_n k_{n\,y} k_{n\,y} \tag{A.39}$$

Die Bedingung $\gamma_{xx}\gamma_{yy} - \gamma_{xy}\gamma_{yx} = 0$ trifft auf alle horizontal bzw. vertikal ausgerichteten Gabor-Filter zu. Um das Displacement für alle Filter korrekt schätzen zu können, wurden diese um $\frac{\pi}{8}$ gedreht. Zusätzlich muss die Phasendifferenz auf den Bereich $[-\pi, \pi]$ normiert werden.

A.7 Eigenwertberechnung bei großen Kovarianzmatrizen

Bei der PCA im Pixelraum wird aufgrund der hohen Dimension der Spalten der Matrix $\mathbf{X}$ die Kovarianzmatrix sehr groß, was die Berechnung der Eigenwerte aufwendig macht. Dieses Problem kann durch einen Trick umgangen werden, bei dem nicht die Kovarianzmatrix im Pixelraum, sondern im Bildraum berechnet wird. Gegeben sei eine Matrix $\mathbf{X}$ der Größe $n \times m$ mit $n > m$.

Die Kovarianzmatrix von $\mathbf{X}$ ist

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T \tag{A.40}$$

mit der Größe $n \times n$. $\mathbf{T}$ sei die Kovarianzmatrix

$$\mathbf{T} = \frac{1}{m}\mathbf{X}^T\mathbf{X} \tag{A.41}$$

mit der Größe $m \times m$ und den Eigenvektoren $\mathbf{v}^{(i)}(i = 1, ..., m)$ und den Eigenwerten λ_i:

$$\mathbf{T}\mathbf{v}^{(i)} = \lambda^{(i)}\mathbf{v}^{(i)} \tag{A.42}$$

Dann gilt:

$$\frac{1}{m}\mathbf{X}^T\mathbf{X}\mathbf{v}^{(i)} = \lambda^{(i)}\mathbf{v}^{(i)} \tag{A.43}$$

$$\frac{1}{m}\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{v}^{(i)} = \lambda^{(i)}\mathbf{X}\mathbf{v}^{(i)} \tag{A.44}$$

$$S(\mathbf{X}\mathbf{v}^{(i)}) = \lambda^{(i)}(\mathbf{X}\mathbf{v}^{(i)}) \tag{A.45}$$

Wenn also $\mathbf{v}^{(i)}$ ein Eigenvektor von $\mathbf{T}$ ist, dann ist $(X\mathbf{v}^{(i)})$ ein Eigenvektor von $\mathbf{S}$ und hat den selben Eigenwert. Die m Eigenvektoren von $\mathbf{S}$ sind dann $p_i(i = 1, ..., m)$:

$$\mathbf{p}_i = \frac{1}{\sqrt{\lambda^{(i)}m}}\mathbf{X}\mathbf{v}^{(i)} \tag{A.46}$$

A.8 Der FastICA-Algorithmus

Der FastICA-Algorithmus basiert auf dem Prinzip der Maximierung der Nichtgaußhaftigkeit der geschätzten unabhängigen Quellen. Entsprechend des Zentralen Grenzwertsatzes der theoretischen Statistik ist die Summe mehrerer stochastisch unabhängiger Zufallsvariablen einer Normalverteilung ähnlicher als die einzelnen Verteilungen. Dieser Fakt kann genutzt werden, um ein Summensignal in seine unabhängigen Komponenten zu zerlegen. Hierbei wird angenommen, dass die Verteilungen der einzelnen Komponenten einer Gaußverteilung sehr unähnlich sind. Für die Schätzung einer unabhängigen Quelle wird folgende Gleichung betrachtet:

$$\hat{s}_i = \hat{\mathbf{w}}_i^T \mathbf{x}, \tag{A.47}$$

wobei $\hat{s}_i$ eine unabhängige Komponente und $\hat{\mathbf{w}}_i$ die zugehörige Zeile der Entmischmatrix $\hat{\mathbf{W}}$ ist. Aufgrund der Tatsache, dass $\hat{\mathbf{w}}_i$ nicht direkt bestimmt werden kann, wird ein Schätzer benötigt, um die Nichtgaußhaftigkeit von $\hat{\mathbf{w}}_i^T \mathbf{x}$ zu schätzen. Durch Maximierung dieser, erhält man eine unabhängige Quelle. Der im FastICA-Algorithmus verwendete Ansatz zur Bestimmung der Nichtgaußhaftigkeit, vergleicht den Informationsgehalt (Entropie) einer Zufallsvariable mit dem einer normalverteilten Zufallsvariable gleichen Mittelwerts und gleicher Varianz. Dieses Maß, genannt Negentropie, kann mathematisch folgendermaßen beschrieben werden:

$$J(\hat{s}_i) = H(y_{Gauss}) - H(\hat{s}_i) \tag{A.48}$$

$$H(y) = -\int p(y) \log p(y) dy, \tag{A.49}$$

wobei H die Entropie, p die Wahrscheinlichkeitsdichtefunktion und y_{Gauss} eine gaußverteilte Zufallsvariable darstellt. Da die Entropie einer gaußverteilten Zufallszahl unter allen Zufallszahlen anderer Verteilung, jedoch gleicher Varianz und gleichen Mittelwerts am größten ist, muss die Negentropie maximiert werden, um eine entsprechende gaußunähnliche Zufallszahl zu erhalten. Jedoch stellt die Berechnung der Entropie für hochdimensionale Verteilungen einen hohen rechentechnischen Aufwand dar. Aus diesem Grund ist es notwendig, eine Approximation der Negentropie vorzunehmen:

$$J(\hat{s}_i) \approx \left[E\{G(\hat{s}_i)\} - E\{G(y_{Gauss})\}\right]^2 \tag{A.50}$$

wobei G eine nichtquadratische Funktion darstellt. Laut [Hyvärinen and Karhunen, 2001] können hierfür folgende Nichlinearitäten G und deren erste Ableitung g genutzt werden:

$$G_1(\hat{s}) = \frac{1}{a_1} \log(\cosh(a_1 \hat{s})) \qquad g_1(\hat{s}) = \tanh(a_1 \hat{s}) \tag{A.51}$$

$$G_2(\hat{s}) = -\frac{1}{a_2} \exp(-\frac{1}{2} a_2 \hat{s}^2) \quad g_2(\hat{s}) = \hat{s} \exp(-\tfrac{1}{2} a_2 \hat{s}^2) \tag{A.52}$$

$$\tag{A.53}$$

Setzt man nun in Gleichung A.50 das inverse ICA-Modell ein, so ergibt sich:

$$J(\hat{\mathbf{w}}_i^T\mathbf{x}) \approx \left[E\{G(\hat{\mathbf{w}}_i^T\mathbf{x})\} - E\{G(y_{Gauss})\}\right]^2, \tag{A.54}$$

wobei $\hat{\mathbf{w}}_i$ eine Zeile der zu schätzenden Entmischmatrix $\hat{\mathbf{W}}$ darstellt, welche eine unabhängige Komponente schätzt. Somit ergibt sich folgendes Optimierungsproblem:

$$\sum_{i=0}^{n-1} J(\hat{\mathbf{w}}_i) \rightarrow max. \tag{A.55}$$

$$C\{(\hat{\mathbf{w}}_k^T\mathbf{x})(\hat{\mathbf{w}}_i^T\mathbf{x})^T\} = 0,\ i \neq k, \tag{A.56}$$

Im Ergebnis dieser Optimierung erhält man, unter Verwendung einer von angegebenen Lernregel der Entmischmatrix, den FastICA-Algorithmus [Hyvärinen and Karhunen, 2001].

Abkürzungsverzeichnis

AAM	Active-Appearance-Model
AU	Action Unit
EER	Equal-Error-Rate
EGM	Elastic-Graph-Matching
FAR	False-Acceptance-Rate
FRR	False-Rejection-Rate
ICA	Independent-Component-Analysis
LVQ	Learning-Vector-Quantifier
MLP	Multilayer-Perceptron
NN	Nearest-Neighbor
PCA	Principal-Component-Analysis
PTU	Pan-Tilt-Unit
RBF	Radial-Basis-Function
SVM	Support Vector Machine

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Aryananda, 2002] Aryananda, L. (2002). Recognizing and remembering individuals: Online and unsupervised face recognition for humanoid robot. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'02), Lausanne*.

[Backhaus, 2003] Backhaus, A. (2003). Implementierung und Untersuchung moderner neuronaler und probabilistischer Verfahren zur videobasierten Mimikanalyse. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

[Bartlett, 2001] Bartlett, M. (2001). *Face image analysis by unsupervised learning*. Kluwer Academic Publishers.

[Bartlett et al., 2003] Bartlett, M., Littlewort, G., Fasel, I., and Movellan, J. (2003). Real time face detection and expression recognition: Development and application to human-computer interaction. In *Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction (CVPR'03)*.

[Bartlett et al., 2002] Bartlett, M., Movelan, J., and T.J., S. (2002). Face recognition by independent component analysis. *IEEE Trans. on Neural Networks*, 13(6):1450–1464.

[Bell and Sejnowski, 1995] Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computing*, 7:1129–1159.

[Bendlin, 2004] Bendlin, A. (2004). Vision-basierte Personenidentifikation und Geschlechtsschätzung für die Mensch-Maschine-Schnittstelle eines mobilen Serviceroboters. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

[Bischoff and Graefe, 2002] Bischoff, R. and Graefe, V. (2002). Demonstrating the humanoid robot hermes at an exhibition: A long-term dependability test. In *Proc. Int. Conf. on Intelligent Robots and Systems (IROS'02), Lausanne*.

[Böhme et al., 1998] Böhme, H.-J., Braumann, U.-D., Brakensiek, A., Corradini, A., Krabbes, M., and Gross, H.-M. (1998). User localisation for visually-based human-machine interaction. In *Proc. 1998 IEEE Int. Conf. on Face and Gesture Recognition (FGR'98), Nara, Japan*, Seiten 486–491.

[Bradski, 1998a] Bradski, G. (1998a). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 2.

[Bradski, 1998b] Bradski, G. (1998b). Real time face and object tracking as a component of a perceptual user interface. In *Proc. 4th IEEE Workshop on Applications of Computer Vision*, Seiten 19–21.

[Braumann, 2001] Braumann, U. (2001). *Multi-Cue-Ansatz für ein dynamisches Auffälligkeitssystem zur visuellen Personenlokalisation.* Dissertation, TU Ilmenau.

[Breazeal, 1999] Breazeal, C. (1999). Robot in society: Friend or appliance? In *Agents99 workshop on emotion-based agent architectures, Seattle, WA.*, Seiten 18–26.

[Breazeal et al., 2004] Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Mulanda, D. (2004). Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*, 1(2).

[Breazeal and Scassellati, 1999] Breazeal, C. and Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proc. of the IEEE/RSJ Conf. on Intelligent Robots and Systems (IROS'99), Kyongju, Korea*, Seiten 858–863.

[Breazeal and Scassellati, 2000] Breazeal, C. and Scassellati, B. (2000). Infant-like social interactions between a robot and a human caretaker. *Special issue of Adaptive Behavior on Simulation Models of Social Agents*, 8.

[Bruce et al., 2002] Bruce, A., Nourbakhsh, I., and Simmons, R. (2002). The role of expressiveness and attention in human-robot interaction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA-02).*

[Burgard et al., 1999] Burgard, W., Cremers, A., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., and Thrun, S. (1999). Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2):3–55.

[Burgard et al., 1998] Burgard, W., Cremers, A., Fox, D., Lakemeyer, G., Hähnel, D., Schulz, D., Steiner, W., and Thrun, S. (1998). The interactive museum tour-guide robot. In *Proc. 15th Nat. Conf. on Artificial Intelligence (AAAI'98)*, Seiten 11–18.

[Burt and Perrett, 1995] Burt, D. and Perrett, D. (1995). Perception of age in adult caucasian male faces: computer graphic manipulation of shape and colour information. In *Proc. Royal Society B*, volume 259, Seiten 137–143.

[Cootes et al., 1992a] Cootes, T., Cooper, D., Taylor, C., and Graham, J. (1992a). A trainable method of parametric shape description. *Image and Vision Computing*, 10(5):289–294.

[Cootes et al., 1998] Cootes, T., Edwards, G., and Taylor, C. (1998). Active appearance models. *Lecture Notes in Computer Science*, 1407:484–498.

[Cootes and Taylor, 1992] Cootes, T. and Taylor, C. (1992). Active shape models - smart snakes. In *British Machine Vision Conference*, Seiten 266–275, Berlin. Springer Verlag.

[Cootes and Taylor, 1999] Cootes, T. and Taylor, C. (1999). Statistical models of appearance for computer vision. Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, Manchester M13 9PT.

[Cootes et al., 1992b] Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1992b). Training models of shape from sets of examples. In *Proc. British Machine Vision Conference*, Seiten 266–275, Berlin. Springer.

[Cootes et al., 2000] Cootes, T., Walker, K., and Taylor, C. (2000). View-based active appearance models. In *Int. Conf. on Face and Gesture Recognition*, Seiten 227–232.

[Corty and Marchand, 2003] Corty, N. and Marchand, E. (2003). Visual perception based on salient features. In *Proc. Int. Conf. on Intelligent Robots and Systems (IROS-03), Las Vegas*, Seiten 1024–1029.

[Costen et al., 1999] Costen, N., Cootes, T., Edwards, G., and C., T. (1999). Automatic extraction of the face identity-subspace. In *Proc. of the British Machine Vision Conference*, volume 2, Seiten 513–522.

[Dailey and Cottrell, 1999] Dailey, M. and Cottrell, G. (1999). PCA = Gabor for expression recognition. Technical Report CS-629, Universitiy of California San Diego (UCSD).

[Dailey et al., 2002] Dailey, M., Cottrell, G., Padgett, C., and Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, 14(8):1158–1173.

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. J. Murray, London.

[Daugman, 1985] Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169.

[DeCarlo and Metaxas, 1997] DeCarlo, D. and Metaxas, D. (1997). The integration of optical flow and deformable models with applications to human face shape and motion estimation. *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR'97)*, Seiten 231–238.

[Donato and Bartlett, 1999] Donato, G. and Bartlett, M. (1999). Classifying facial actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(10):974–989.

[Dornaika and Ahlberg, 2004] Dornaika, F. and Ahlberg, J. (2004). Fast and reliable active appearance model search for 3d face tracking. *IEEE Trans. on Systems, Man and Cybernetics, Part B*, 34:1838–1853.

[Eckardt, 2005] Eckardt, R. (2005). Vergleichende Untersuchung von Verfahren zur Erkennung von Mimik, Alter Geschlecht und Identität von Personen in Videodaten. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

[Edwards and Cootes, 1998] Edwards, G. and Cootes, T. (1998). Face recognition using active appearance models. *Proc. 5th Eur. Conf. on Computer Vision (ECCV'98)*, 2:581–695.

[Ehrenmann et al., 2001] Ehrenmann, M., Knoop, S., Zöllner, R., and Dillmann, R. (2001). Sensor fusion approaches for observation of user actions in programming by demonstration. In *Proc. Int. Conf. on Multi Sensor Fusion and Integration for Intelligent Systems (MFI)*, Seiten 227–232.

[Ekman, 1989] Ekman, P. (1989). *Handbook of Social Psychophysiology.* John Wiley, Chichester.

[Ekman and Friesen, 1975a] Ekman, P. and Friesen, W. (1975a). *Pictures of Facial Affect.* Consulting Psychologists Press, Palo Alto, CA.

[Ekman and Friesen, 1975b] Ekman, P. and Friesen, W. (1975b). *Unmasking the face. A guide to recognizing emotions from facial clues.* Prentice-Hall, Englewood Cliffs, New Jersey.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto.

[Erich, 2003] Erich, J. (2003). Kontinuierliches stereobasiertes Tracken von Personen mit einem mobilen Serviceroboter. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

[Fasel and Luettin, 2003] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275.

[Fellenz and Taylor, 1999] Fellenz, W. and Taylor, J. (1999). Comparing template-based, feature-based and supervised classification of facial expressions from static images. In *Proc. of Circuits, Systems, Communications and Computers (CSCC'99)*, Seiten 5331–5336.

[Feyrer and Zell, 1999] Feyrer, S. and Zell, A. (1999). Detection, tracking, and pursuit of humans with an autonomous mobile robot. In *Proc. Int. Conf. on Intelligent Robots and Systems (IROS'99)*, Seiten 864–869.

[Fieguth and Terzopoulos, 1997] Fieguth, P. and Terzopoulos, D. (1997). Color-based tracking of heads and other mobile objects at video frame rates. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'97)*, Seiten 21–27.

[Fong et al., 2002] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2002). A survey of socially interactive robots: concepts, design, and applications. Technical Report CMU-RI-TR-02-29, Carnegie Mellon University Robotics Institute.

[Fritsch et al., 2002] Fritsch, J., Lang, S., Kleinehagenbrock, M., Fink, G., and Sagerer, G. (2002). Improving adaptive skin color segmentation by incorporationg results from face detection. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN), Berlin, Germany*, Seiten 337–343.

[Fröba and C., 2001] Fröba, B. and C., K. (2001). Face detection and tracking using edge orientation information. *SPIE Visual Communications and Image Processing*, Seiten 583–594.

[Funt et al., 1998] Funt, B., K., B., and Martin, L. (1998). Is machine colour constancy good enough? In *Proc. 5th Eur. Conf. on Computer Vision (ECCV'98)*, Seiten 445–459.

[Féraud et al., 2001] Féraud, R., Bernier, O., Viallet, J.-E., and Collobert, M. (2001). A fast and accurate face detector based on neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(1):42–53.

[Golomb et al., 1991] Golomb, B., Lawrence, D., and Sejnowski, T. (1991). Sexnet: A neural network that recognizes sex from human faces. *Advances in Neural Information Processing Systems*, 3:572–577.

[Graf et al., 2004] Graf, B., Hans, M., and Schraft, R. (2004). Care-o-bot ii - development of a next generation robotic home assistant. *Autonomous Robot Journal*, 16:193–205.

[Gross and Boehme, 2000] Gross, H.-M. and Boehme, H.-J. (2000). PERSES - a Vision-based Interactive Mobile Shopping Assistant. In *Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics (SMC'00), Nashville*, Seiten 80–85. IEEE/Omnipress.

[Gross et al., 2000] Gross, H.-M., Boehme, H.-J., Key, J., and Wilhelm, T. (2000). The PERSES Project - a Vision-based Interactive Mobile Shopping Assistant. *Künstliche Intelligenz*, 4:34–36.

[Gross and König, 2004] Gross, H.-M. and König, A. (2004). Robust omniview-based probalistic self-localization for mobile robots in large maze-like environments. In *17th Int. Conf. on Pattern Recognition (ICPR'04), Cambridge, GB*, Seiten 266–269. IEEE Computer Society Press.

[Gross and Levenson, 1995] Gross, J. and Levenson, R. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9:87–108.

[Hong et al., 1998] Hong, H., Neven, H., and v. d. Malsburg, C. (1998). Online facial expression recognition based on personalized gallery. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Seiten 354–359. IEEE Comp. Soc.

[Hyvärinen and Karhunen, 2001] Hyvärinen, A. and Karhunen, J. (2001). *Independent Component Analysis*. John Wiley& Son, Inc.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430.

[Isard and Blake, 1996] Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density. In *Proc. Eur. Conf. on Computer Vision*, volume 1, Seiten 343–356.

[Isard and Blake, 1998] Isard, M. and Blake, A. (1998). Condensation – conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28.

[Jähne, 1997] Jähne, B. (1997). *Practical Handbook on Image Processing for Scientific Applications.* CRC Press LLC.

[Jang and Kweon, 2001] Jang, G.-J. and Kweon, I.-S. (2001). Robust object tracking using an adaptive color model. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'01)*, Seiten 1677–1682.

[Jebara et al., 1998] Jebara, T., Russell, K., and Pentland, A. (1998). Mixtures of eigenfeatures for real-time structure from texture. In *Proc. 6th Int. Conf. on Computer Vision (ICCV'98)*, Seiten 128–135.

[Kalman, 1960] Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transaction of the ASME - Journal of Basic Engineering*, Seiten 35–45.

[Kanade, 1977] Kanade, T. (1977). *Computer Recognition of Human Faces*, volume 47. Interdisciplinary Systems Research.

[Kanade et al., 2000] Kanade, T., Cohn, J., and Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'00), Grenoble, France*, Seiten 46–53.

[Kass et al., 1987] Kass, M., Witkin, A., and Terzopoulos, D. (1987). Snakes: Active contour models. *Int. J. Computer Vision*, 1(4):321–331.

[King and Weiman, 1990] King, S. and Weiman, C. (1990). Helpmate autonomous mobile robot navigation system. In *Proc. of the SPIE Conf. on Mobile Robots, Boston*, Seiten 190–198.

[Kobayashi and Hara, 1997] Kobayashi, H. and Hara, F. (1997). Facial interaction between animated 3d face robot and human beings. *Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics (SMC'97)*, Seiten 3732–3737.

[Kotropoulos and Pitas, 1997] Kotropoulos, C. and Pitas, I. (1997). Rule-based face detection in frontal views. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '97)*, volume 4, Seiten 2537–2540.

[Lades et al., 1993] Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311.

[Lanitis et al., 2004] Lanitis, A., Draganova, C., and Christodoulou, C. (2004). Comparing different classifiers for automatic age estimation. *SMC-B*, 34(1):621–628.

[Lanitis et al., 1995] Lanitis, A., Taylor, C., and Cootes, T. (1995). A unified approach to coding and interpreting face images. In *Proc. Int. Conf. on Computer Vision (ICCV'95)*, Seiten 368–373.

[Lien et al., 1998] Lien, J., Kanade, T., Cohn, J., and Li, C. (1998). Automated facial expression recognition based on facs action units. In *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'98)*, Seiten 390–395.

[Lisetti and Rumelhart, 1998] Lisetti, C. and Rumelhart, D. (1998). Facial expression recognition using a neural network. *Proc. 11th Int. Flairs Conference (FLAIRS'98)*, Seiten 328–332.

[Littlewort et al., 2003] Littlewort, G., Bartlett, M., Chenu, J., Fasel, I., Kanda, T., Ishiguro, H., and Movellan, J. (2003). Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification.

[Liu and Wechsler, 2003] Liu, C. and Wechsler, H. (2003). Independent component analysis of gabor features for face recognition. *IEEE Trans. Neural Networks*, 14(3):919–928.

[Lyons and Budynek, 1999] Lyons, M. and Budynek, J. (1999). Automatic classification of single facial image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362.

[Lyons et al., 2000] Lyons, M., Campbell, R., Plante, A., Coleman, M., Kamachi, M., and Akamatsu, S. (2000). The noh mask effect: Vertical viewpoint dependence of facial expression perception. *Proceedings of the Royal Society of London B*, 267:2239–2245.

[MacCormick and Blake, 1999] MacCormick, J. and Blake, A. (1999). A probabilistic exclusion principle for tracking multiple objects. In *Proc. Int. Conf. Computer Vision (ICCV'99)*, Seiten 572–578.

[Magnusson, 2005] Magnusson, M. (2005). Understanding social interaction: Discovering hidden structure with model and algorithms. In Anolli, L., Duncan JR., S., Magnusson, M., and Riva, G., editors, *The Hidden Structure of Interaction, From Neurons to Culture Patterns*. IOS Press.

[Mase and Pentland, 1991] Mase, K. and Pentland, A. (1991). Recognition of facial expression from optical flow. *Institute of electronics information and communication engineers Transactions*, 74(10):3474–3483.

[Matthews and Baker, 2003] Matthews, I. and Baker, S. (2003). Active appearance models revisited. Technical Report CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[Menser and Bräunig, 1999] Menser, B. and Bräunig, M. (1999). Segmentation of human faces in color images using connected operators. In *IEEE Int. Conf. on Image Processing*, volume 3, Seiten 632–636.

[Moghaddam and Yang, 2000] Moghaddam, B. and Yang, M.-H. (2000). Gender classification with support vector machines. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition (FGR'00), Grenoble, France*, Seiten 306–311.

[Okuno et al., 2002] Okuno, H. G., Nakadai, K., and Kitano, H. (2002). Social interaction of humanoid robot based on audio-visual tracking. In *Proc. 18th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'02), Cairns, Australia*, Seiten 725–735. Springer-Verlag.

[Otsuka and Ohya, 1998] Otsuka, T. and Ohya, J. (1998). Extracting facial motion parameters by tracking feature points. In *Proc. 1st Int. Conf. on Advanced Multimedia Content Processing*, Seiten 442–453.

[Padgett and Cottrell, 1997] Padgett, C. and Cottrell, G. (1997). Representing face images for emotion classification. *Advances in Neural Information Processing Systems*, 9:894–900.

[Phillips, 1999] Phillips, P. (1999). Support vector machines applied to face recognition. *Advances in Neural Information Processing Systems*, 11:803–809.

[Rasmussen and Hager, 2001] Rasmussen, C. and Hager, G. D. (2001). Probabilistic data association methods for tracking complex visual objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):560–576.

[Reithinger et al., 2003] Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., and Tschernomas, V. (2003). Smartkom - adaptive and flexible multimodal access to multiple applications. In *Proceedings of Int. Conf. of Multimodal Interfaces*, Seiten 101–108.

[Rowley et al., 1998] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.

[Sanger et al., 1997] Sanger, D., Miyake, Y., Haneishi, Y., and Tsumura, N. (1997). Algorithm for face extraction based on lip detection. *Journal of Imaging Science and Technology*, 41(1):71–80.

[Schiele and Waibel, 1995] Schiele, B. and Waibel, A. (1995). Gaze tracking based on face-color. In *Int. Workshop on Automatic Face- and Gesture-Recognition (FGR'95)*, Seiten 344–348.

[Schröter et al., 2004] Schröter, C., Böhme, H.-J., and Gross, H.-M. (2004). Robust map building for an autonomous robot using low-cost sensors. In *IEEE Int. Conf. on Systems, Man and Cybernetics (SMC'04), Den Haag, Netherlands*, Seiten 5398–5403. IEEE/Omnipress.

[Schulte et al., 1999] Schulte, J., Rosenberg, C., and Thrun, S. (1999). Spontaneous, short-term interaction with mobile robots. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA'99)*, Seiten 658–663.

[Schulz and Burgard, 2001] Schulz, D. and Burgard, W. (2001). Probabilistic state estimation of dynamic objects with a moving mobile robot. *Robotics and Autonomous Systems*, 34:107–115.

[Schulz et al., 2001] Schulz, D., Burgard, W., Fox, D., and Cremers, A. (2001). Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'01)*, Seiten 1665–1670.

[Schulz et al., 2003] Schulz, D., Burgard, W., Fox, D., and Cremers, A. (2003). People tracking with mobile robots using sample-based joint probabilistic data association filters. *Int. Journal of Robotics Research (IJRR)*, 22(2):99–116.

[Shakhnarovich et al., 2002] Shakhnarovich, G., Viola, P., and Moghaddam, B. (2002). A unified learning framework for real time face detection and classification. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'02), Washington, D.C., USA*, Seiten 16–26.

[Shiomi et al., 2004] Shiomi, M., Kanda, T., MIralles, N., Miyashita, T., Fasel, I., Movellan, J., and Ishiguro, H. (2004). Face-to-face interactive humanoid robot. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'04)*, Seiten 1340–1346.

[Shligerskiy, 2002] Shligerskiy, M. (2002). Entwicklung von Verfahren zur Schätzung des Geschlechtes aus Gesichtsbilddaten. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

[Soriano et al., 2000] Soriano, M., Martinkauppi, B., Huovinen, S., and Laaksonen, M. (2000). Skin color modeling under varying illumination conditions using the skin locus for selecting training pixels. In *Workshop on Real-Time Image Sequence Analysis, Oulu, Finland*, Seiten 43–49.

[Soriano et al., 2003] Soriano, M., Martinkauppi, B., Huovinen, S., and Laaksonen, M. (2003). Adaptive skin color modeling using the skin locus for selecting training pixels. *Pattern Recognition*, 36(3):681–690.

[Störring, 2004] Störring, M. (2004). *Computer Vision and Human Skin Colour*. PhD thesis, Faculty of Engineering and Science, Aalborg University.

[Sung and Poggio, 1998] Sung, K.-K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):39–51.

[Tao et al., 1999] Tao, H., Sawhney, H., and Kumar, R. (1999). A sampling algorithm for tracking multiple objects. In *Proceedings of the International Workshop on Vision Algorithms*, Seiten 53–68.

[ten Bosch et al., 2004] ten Bosch, L., Oostdijk, N., and de Ruiter, J. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Proc. Conf. of Text, Speech and Dialogue (TSD'04), Brno, Czech Republic*, Seiten 563–570.

[Terrillon et al., 2000] Terrillon, J., Shirazi, M., Fukamachi, H., and Akamatsu, S. (2000). Comarative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *4th IEEE Int. Conf. on Automatic Face- and Gesture-Recognition (FGR'00), Grenoble, France*, Seiten 54–61.

[Thrun et al., 1999] Thrun, S., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Haehnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. (1999). Minerva: A second generation mobile tour-guide robot. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'99), Detroit, Michigan, USA*, Seiten 1999–2005.

[Tojo et al., 2000] Tojo, T., Matsusaka, Y., Ishii, T., and Kobayashi, T. (2000). A conversational robot utilizing facial and body expressions. In *in Proc. Int. Conf. on System, Man and Cybernetics (SMC'2000)*, Seiten 858–863.

[Trapp, 2005] Trapp, N. (2005). Active appearance models. Diplomarbeit, Fachgebiet Neuroinformatik, TU Ilmenau.

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

[Tweed and Calway, 2002] Tweed, D. and Calway, A. (2002). Tracking many objects using subordinated condensation. In *Proc. British Machine Vision Conference (BMVC'02), Cardiff, UK*, Seiten 283–292.

[Varshavskaya, 2002] Varshavskaya, P. (2002). Behavior-based early language development on a humanoid robot. In *2nd Int. Workshop on Epigenetic Robotics, Edinburgh, Scotland*, Seiten 149–158.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Robust real-time face detection. In *Proc. 8th Int. Conf. On Computer Vision (ICCV-01), Vancouver, Canada*, volume 2, page 747. IEEE Computer Society.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

[Wahlster, 2002] Wahlster, W. (2002). Smartkom: Fusion and fission of speech, gestures, and facial expressions. In *Proc. 1st Int. Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan*, Seiten 213–225.

[Watzlawick et al., 1996] Watzlawick, P., Beavin, J., and Jackson, D. (1996). *Menschliche Kommunikation - Formen, Störungen, Paradoxien.* Verlag Hans Huber.

[Wilhelm and Backhaus, 2004] Wilhelm, T. and Backhaus, A. (2004). Statistical and neural methods for vision-based analysis of facial expressions and gender. In *Proc. IEEE Int. Conf. on System Man and Cybernetics*, Seiten 2203–2208.

[Wilhelm et al., 2003a] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. (2003a). Automatischer Weissabgleich für eine omnidirektionale Kamera. In *Proc. 9. Workshop für Farbbildverarbeitung, Esslingen*, Seiten 43–50. Schriftenreihe ZBS.

[Wilhelm et al., 2003b] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. (2003b). Looking closer. In *Proc. 1st Eur. Conf. on Mobile Robots (ECMR'03)*, Seiten 65–70. ZTUREK Research-Scientific Institute.

[Wilhelm et al., 2003c] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. (2003c). Towards an attentative robotic dialog partner. In *Proc. 4th Int. Conf. on Multimodal Interfaces (ICMI'03), Vancouver*, Seiten 297–300. ACM press.

[Wilhelm and Martin, 2004] Wilhelm, T. and Martin, C. (2004). Vergleich von hautfarbbasierten Multi-Target-Trackern. In *Proc. 3rd Workshop on Self-Organization of AdaptiVE Behavior (SOAVE'04), Ilmenau*, Seiten 27–36. VDI-Verlag.

[Wiskott et al., 1995] Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1995). Face recognition and gender determination. In *Int. Workshop on Automatic Face- and Gesture-Recognition (FGR'95), Zürich*, Seiten 92–97.

[Wiskott et al., 1997a] Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997a). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis Machine Intelligence*, 19(7):775–779.

[Wiskott et al., 1997b] Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997b). Face recognition by elastic bunch graph matching. In Gerald Sommer, Kostas Daniilidis, and Josef Pauli, editors, *Proc. 7th Int. Conf. on Computer Analysis of Images and Patterns, Kiel*, number 1296 in ?, Seiten 456–463, Heidelberg. Springer-Verlag.

[Wu et al., 2003] Wu, C., Liu, C., Shum, H.-Y., Y.-Q., X., and Zhang, Z. (2003). Automatic eyeglasses removal from face images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(3):322–336.

[Wullschleger and Brega, 2001] Wullschleger, F. and Brega, R. (2001). The mops has grown-up: From a research platform to a high-availability service robot. In *Proc. Eur. Workshop on Advanced Mobile Robots (EUROBOT'01), Lund, Sweden*.

[Yacoob and Black, 1996] Yacoob, Y. and Black, M. (1996). Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. Journal of Computer Vision*, 25(1):23–48.

[Yacoob and Davis, 1996] Yacoob, Y. and Davis, L. (1996). Recognizing human facial expression from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642.

[Yang and Waibel, 1996] Yang, J. and Waibel, A. (1996). A real-time face tracker. *Proc. of the 3rd Workshop on Applications of Computer Vision*, Seiten 142–147.

[Yang and Waibel, 1998] Yang, J. and Waibel, A. (1998). Skin-color modeling and adaptation. In *Proc. 3rd Asian Conf. on Computer Vision (ACCV'98) (2), Hong Kong, China*, Seiten 687–694.

[Yang et al., 2002] Yang, M.-H., Kriegman, D., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

[Yang et al., 2000] Yang, M.-H., Roth, D., and Ahuja, N. (2000). A snow-based face detector. *Advances in Neural Information Processing Systems*, 12:855–861.

[Yoneyama and Iwano, 1997] Yoneyama, M. and Iwano, Y. (1997). Facial expression recognition using discrete hopfield neural networks. *Proc. Int. Conf. on Image Processing (ICIP'97)*, 3:117–120.

[Yuille, 1991] Yuille, A. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70.

[Zarit et al., 1999] Zarit, B., Super, B., and Quek, F. (1999). Comparison of five color models in skin pixel classification. In *Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Seiten 58–63.

[Zell, 1994] Zell, A. (1994). *Simulation neuronaler Netze*. Addison-Wesley.

[Zimmermann et al., 2003] Zimmermann, P., Guttormsen, S., Danuser, B., and Gomez, P. (2003). Affective computing - a rationale for measuring mood with mouse and keyboard. *Int. Journal of Occupational Safety and Ergonomics (JOSE)*, 9:539–551.