Technische Universität München
Lehrstuhl für Mensch-Maschine-Kommunikation
Univ.-Prof. Dr.-Ing. habil. Gerhard Rigoll

Diplomarbeit

Erstellung einer Datenbank und Untersuchungen zur dynamischen Mimikerkennung

Verfasser:	Michael Hawellek Spitzwegstr.3, 94469 Deggendorf 2020187
Betreuer:	Frank Wallhoff
Laborzeit:	6 Monate
Unterbrechungen:	keine
Abgabetermin:	01.07.2004

Inhaltsverzeichnis

Kapitel 1

Einleitung

In der menschlichen Kommunikation spielt die mimische Information eine wichtige Rolle, um den Kommunikationspartner zu verstehen und auf die richtige Weise reagieren zu können. Der ausdruckstärkste Weg für Menschen ihre Emotionen zu zeigen sind die Gesichtsaudrücke. Diese stellen nicht nur Informationen über den Gefühlszustand eines Benutzers zur Verfügung, sondern lassen auch Rückschlüsse über die kognitive Aktiviät, das Temperament und die Persönlichkeit sowie die Aufrichtigkeit einer Person zu.

Wissenschatliche Untersuchenen ergaben dass 55% des Nachrichtengehalts bei der sozialen Interaktion zwischen Individuen über die Mimik übertragen wird. Deshalb wären funktionsfähige Klassifikations- und Erkennungssysteme wünschenswert, um einen wichtigen Kanal in der MMK zu schließen und eine verbesserte Interaktion zu ermöglichen. Die kommunikative Macht von Gesichtsaudrücken macht die Erkennung und das Verständnis von menschlichen Emotionen zu einer wichtigen Aufgabe in der Bildverarbeitung.

Computeranimierte Agenten und Roboter bringen eine soziale Komponente in die Mensch-Maschine Kommunikation und lassen uns auf eine neue Art und Weise über den Einsatz von Computern im täglichen Leben nachdenken. Dabei sind Einsatzbereiche wie künstliche Avatare, die mit menschlichen Benutzern interagieren, automatische Meeting-Manager, Flugzeugüberwachung, Verwendung in Automobilen sowie in allen Bereichen, wo Kommunikation zwischen Mensch und Maschine erfolgt, denkbar.

Menschliche Emotionen zu verstehen ist also eine wichtigste Fähigkeite für Computer, um auf intelligente Weise mit menschliche Benutzern interagieren zu können. Erst kürzlich gemachte Fortschritte in der Bildverarbeitung und im Bereich der neuronalen Netze eröffnen die Möglichkeit der automatischen Erkennung von Gesichtsaudrücken.

In den letzen Jahren hat sich das Interesse der Forschung von dem Hauptbereich der Gesichtserkennung auch auf die Emotionserkennung ausgeweitet. Immer mehr Wissenschaftler versuchen Systeme zur automatischen Erkennung von Gesichtsausdrücken zu entwickeln. Die Erkennung von Gesichtsaudrücken in Bildsequenzen mit signifikanter Kopfbewegung ist eine Herausforderung für viele Anwendungen in der MMK. Und obwohl die Kopfbewegung mit der Bewegung der Merkmale der Emotionen einhergeht, wurde der Bewegung als Teil der Mimik und damit der Dynamik der Emotionen bis jetzt nur wenig Aufmerksamkeit

geschenkt. Wichtig ist vor allem, natürliche und spontane Emotionen zu detektieren und die zugrunde liegende Information zu extrahieren, zu interpretiern und entsprechend darauf zu reagieren.

1.1 Motivation

In der folgenden Arbeit soll ein Versuchsaufbau für die Aufnahme von Blickrichtungen und Gesichtsaudrücken entwickelt und damit eine Datenbank für sieben Klassen von Emotionen erstellt werden. Dazu sind alle erforderlichen Hardware- und Softwarevoraussetzungen zu schaffen und damit eine möglichst automatisierte Aufnahme zu ermöglichen.

Im zweiten Teil sollen mit der erstellten Mimikdatenbank Untersuchungen zu dynamischen Verfahren der Erkennung von Gesichtsaudrücken durchgeführt werden. Als Abschluss der Arbeit ist ein Vergleich der Methoden, die Präsentation der Ergebnisse und eine Einordung des Systems anzustellen.

1.2 Gliederung

Die Diplomarbeit ist wie folgt strukturiert:

- **Kapitel 2:** Dieses Kapitel behandelt die Grundlagen zum Verständnis der Mimiken beim Menschen, geht auf die Eigenschaften dieser ein und gibt Anregungen zu deren Erfassung in einer Emotionsdatenbank.
- **Kapitel 3:** Hier wird auf wichtige Grundbegriffe einer Datenbank eingegangen und die Anforderungen festgelegt sowie auf den Aufbau der Emotionsdatenbank eingegangen.
- **Kapitel 4:** Dieser Abschnitt enthält alle wichtigen Entwicklungen und Daten zum Aufbau des Setups sowohl der Hardware als auch der Software. Es beschreibt insbesondere die Versuchsanordnung, die dafür entwickelte Software und den Ablauf der Aufnahme. Auch wird auf die Erstellung der Videos zur Verwirklichung der Mimikaufnahme eingegangen und die Probleme und Verbesserungen bei der Erstellung besprochen.
- **Kapitel 5:** Hier wird ein Überblick über bereits bestehende Systeme und Verfahren zur Erkennung von Gesichtsausdrücken gegeben.
- **Kapitel 6:** Dieser Abschnitt beinhaltet die Grundlagen und die Mimikerkennung mit Pseudo-3D Hidden Markov Modellen und vergleicht dieses Verfahren mit einem weiteren einfacheren Modell zur Merkmalsextraktion.
- **Kapitel 7:** Im letzten Abschnitt werden alle durchgeführten Trainings- und Testdaten aufgeführt, Ergebnisse präsentiert und Lösungen besprochen. Schließlich erfolgt die Zusammenfassung und das Resümee der Arbeit.

Kapitel 2

Grundlagen

Wichtig für die Erstellung einer Mimik-Datenbank ist es, sich im Vorfeld klarzumachen, welche Emotionen und Mimiken im natürlichsprachlichen Dialog beim Menschen überhaupt vorkommen. Welches sind die häufigsten Gesichtsausdrücke?

Dazu wurden beispielsweise nach [Fai98] bereits viele Untersuchungen angestellt. Sowohl Psychologen als auch Anthropologen lieferten interessante Ergebnisse. Auch im Internet kursieren viele Untersuchungen zu diesem Thema. Um herauszufinden wie Menschen auf bestimmte Gesichtsausdrücke reagieren, genügt es, Personen Fotografien von Gesichtsausdrücken anderer Menschen zu zeigen. Das Ergebnis ist, dass eine bestimmte Mimik bei fast allen zu einer identischen Interpretation führt. Auch bei sehr differenzierten Gesichtsausdrücken kommen die Personen zu einem übereinstimmenden Ergebnis. Selbst fremde Kulturen mit unterschiedlichem sozialen Umfeld zeigen die selben Übereinstimmungen.

Der Psychologe Paul Ekman hat nach [Fai98] eine Liste aller Studien zu diesem Thema zusammengestellt und als Übereinstimmung aller Untersuchungsergebnisse folgendes gefunden: Es gibt bestimmte allgemeingültige Gesichtsausdrücke, die sich in folgende sechs Basisemotionen einteilen lassen:

- Traurigkeit
- Freude
- Ärger
- Angst
- Ekel
- Überraschung

sowie den neutralen Zustand als Ergänzung. Allen Personen, denen Mimiken aus einer dieser Kategorien gezeigt wurden, kamen zu dem selben Ergebnis. Deshalb können diese sieben Gesichtsausdrücke als allgemein gültig angesehen werden. Anzumerken ist, dass es noch Erweiterungen und Unterkategorien dieser Gesichtsausdrücke gibt, wie beispielsweise Schmerzen, Leidenschaft oder körperliche Erschöpfung. Diese entstehen aber nicht durch einen see-

lischen, sondern aufgrund eines körperlichen Zustandes. Schließlich gibt es noch Gesichtsausdrücke, die nicht in eine der genannten Kategorien fallen, sondern entweder Mischformen oder subjektive, durch äußere Umstände bedingte Mimiken sind. Diese werden in der vorliegenden Arbeit nicht berücksichtigt.

2.1 Basisemotionen

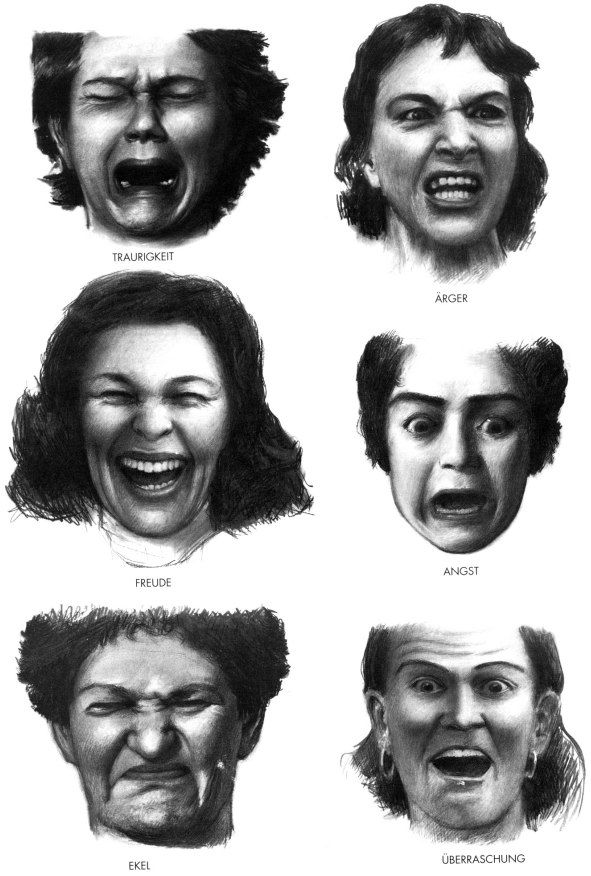


Abbildung 2.1: Die sechs Basisemotionen nach [Fai98]

Um Emotionen zu erkennen, ist es wichtig, sich zuerst klarzumachen, welche grundlegenden Charakteristika die einzelnen Mimiken besitzen. Erst mit dieser Information kann ein für die Aufnahme passendes Setup für die realistische Aufzeichnung entworfen werden. Zunächst wird deshalb näher auf die Entstehung und die Merkmale jeder Basisemotion eingegangen und zudem zu erklären versucht, welche Bedeutung die Mimik im sozialen Verhalten hat, um daraus Verfahrensweisen für deren Erzeugung sowie für die Aufnahme der Datenbank abzuleiten. Die im Folgenden beschriebenen Eigenschaften und Erkenntnisse sind überwiegend der Theorie und den Beobachtungen nach Faigin [Fai98] entnommen, welches besonders gut die wichtigen Gesichtspunkte der Emotionen für den Ausdruck und dessen Herkunft verdeutlicht.

2.1.1 Traurigkeit

Traurigkeit hat seine Wurzeln in der frühen Kindheit und ruft einen Gesichtsausdruck hervor, welcher dem allerersten Säuglingsschrei am nächsten kommt. Auch andere negative Ausrücke haben mimische Elemente, die bei Traurigkeit typisch sind. Diese äußert sich in vielen Formen: Von Kummer über Leid bis hin zu Weinen. Der Unterschied liegt nur in der Intensität begründet, seine Urform bleibt der Säuglingsschrei.

Abbildung 2.2: Weinendes Kind (entnommen aus [Fai98])

Die mimische Gesichtsmuskulatur für Traurigkeit ist somit angeboren und kann deshalb von allen Menschen unabhängig sozialer Klasse und Kulturund dechiffriert werden. Charakteristisch ist der quadratische Mund („Uahh" beim Säuglingsschrei), was viele Menschen auch durch zusammenpressen der Lippen zu unterdrücken versuchen, und die zusammengepressten Augen (siehe Bild 2.2) durch Kontraktion des Augenringmuskels.

Grundsätzlich liegen Lachen und Weinen nah beieinander. Beobachten kann man dies besonders gut bei Babies, die gerade lachen und im nächsten Moment zu weinen beginnen. Was sich aber im Unterschied zum Lachen in der Mimik Traurigkeit widerspiegelt, ist die Anstrengung die Weinen erfordert, wodurch der verkrampfte Ausdruck entsteht. Der Übergang vom neutralen Zustand zu kummervoller Mimik drückt sich durch wachsende Anspannung aus. So sind beim Weinen bis zu 9 verschiedene Gesichtsmuskeln beteiligt, mehr als bei jedem anderen Gesichtsausdruck.

2.1.2 Freude

Nach dem Schreien kommt das Lachen. Ein Baby beherrscht innerhalb kürzester Zeit die beiden wichtigsten fundamentalen Gesichtsausdrücke: Freude und Schmerz. Ein heranwachsender Mensch dagegen unterdrückt mehr und mehr negative Empfindungen, wie Weinen

oder Schreien. Das Lachen hingegen bleibt ihm sein ganzes Leben erhalten. Lachen ist zudem der allgemeingültigste Gesichtsausdruck, der Menschen untereinander verbindet und in allen Kulturen gleich verstanden wird. Gleichzeitig ist er aber der Gesichtsausdruck mit den feinsten Abstufungen. In ein Lachen kann Traurigkeit sowie Ärger hineinspielen.

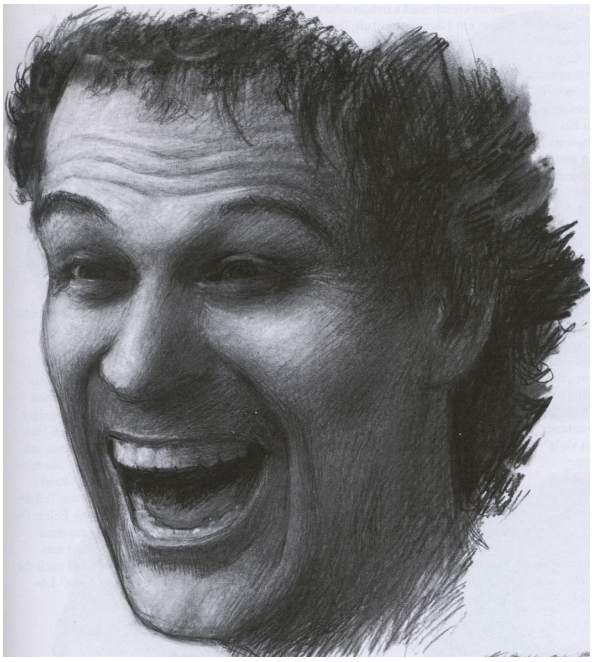


Abbildung 2.3: Freude [Fai98]

Zwei Muskeln sind für lächelnden Gesichtsausdruck verantwortlich: Der Jochbeinmuskel, der die Aufgabe hat, den Mund zu einem Lachen zu verziehen sowie der spiralförmige Augenringmuskel, der auch an anderen Aufgaben wie Weinen oder Ärger beteiligt ist. Beim Lachen spielt er die Hauptrolle: Die lachenden Augen. Ein Lächeln wirkt erst echt, wenn sich die zarte Haut um die Augen in viele kleine Knitterfältchen legt. Der Mund kann sich bewusst zu einem Lächeln verbiegen, aber so wirkt das Lächeln gekünstelt. Erst wenn der Augenringmuskel beteiligt ist, bilden sich Fältchen um die Augen und das Lachen wirkt echt.

Ein Lächeln zeigt sich also zum einen an den Augen. Sobald es sich ankündigt, verkleinert sich der Abstand zwischen den Augenliedern. Je stärker das Lächeln ausgeprägt ist, desto stärker zieht sich der Augenringmuskel zusammen. Im Unterschied zum Weinen ziehen sich beim Lachen nur der innere und untere Muskelteil des Augenlids zusammen. Zum anderen geht der Mund in die Breite und die oberen Zähne kommen zum Vorschein. Dadurch entsteht ein erheblicher Kontrast zwischen der dunklen Mundhöhle und den weißen Zähnen. So hat es die Natur eingerichtet, dass ein Lachen als einziger Gesichtsausdruck auf viele Meter Entfernung erkannt wird und mit keinem anderen verwechselt werden kann. Ein weiteres unübersehbares Merkmal sind die runderen Wangen. Je breiter das Lächeln ist, desto mehr drücken sich die Wangen zusammen.

Die ürsprüngliche Funktion des Lachens bestand darin, dass Menschen in der Steinzeit Freunde bzw. Feinde bereits von Weitem enttarnen konnten.

Ein Lächeln kann noch viele weitere Formen aufgrund seelischer Zustände annehmen, wie das schüchterne Lächeln am Beispiel der Mona Lisa. Bei der Mimikerkennung geht es aber wie erwähnt nur um die Erkennung des Grundzustandes, weswegen weitere Differenzierungen (Bespiel siehe Abbildung 2.8) nicht weiter beachtet werden sollen.

2.1.3 Ärger

Ärger ist ein heftiges Gefühl, das ganz plötzlich auftreten und auch gleich wieder verschwinden kann. Die Gesichtszüge ändern sich dabei unablässig. Ärger drückt sich aber nur in einem kleinen Detail der Mimik aus und zwar vor allem in der Augenpartie. Zum einen in der Veränderung der Augengröße durch Aufreißen der Augen, was Zorn vermittelt, zum anderen durch Senken der Augenbrauen, was bedrohlich wirkt.

In diesem Zusammenhang spricht man auch von dem bösen Blick. Wenn ein Mensch aus der Steinzeit in eine schwierige Situation kam, zog er die Augenbrauen herunter. Böse wirkt diese Mimik aber erst in Verbindung mit weit aufgerissenen Augen. Somit spiegelt sich der Ärger erst in Kombination aus beiden Merkmalen wieder. Das Auge weitet sich, obwohl das Oberlid nach unten drückt.

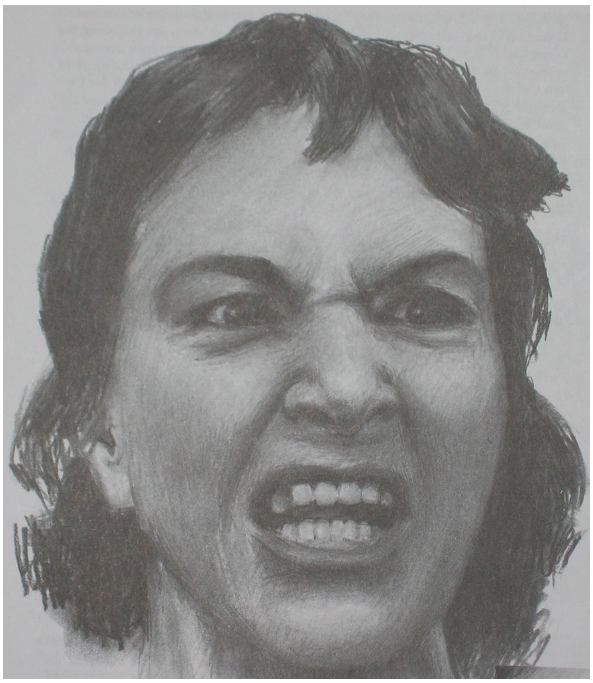


Abbildung 2.4: Ärgerlicher Mensch [Fai98]

Die Mundpartie eines zornigen Menschen nimmt die vielfältigsten Formen an und ist damit sehr dynamisch. Dieses äußert sich in fest zusammengepressten Lippen bis hin zu weit aufgerissenem Mund. Trotzdem gibt es grundlegende Übereinstimmungen: So ist die Form des Mundes im Großen und Ganzen quadratisch und die Mundwinkel bleiben unten. Egal welche Formen der Mund auch annimmt, erst in Zusammenhang mit dem bösen Blick und den heruntergezogenen Augenbrauen drückt sich Ärger aus, wie besonders gut aus Bild 2.4 hervorgeht.

2.1.4 Angst

Die Emotion Angst kann beim Menschen vielfältige Formen annehmen. Besonders zeigt sich Angst in lebensbedrohlichen Situationen, beispielsweise bei Katastrophen oder einem schrecklichen Erlebnis. Nicht nur in Extremsituationen, sondern auch im Alltag treten die verschiedensten Formen von Furcht auf: Angst vor Veränderungen, vor Versagen oder vor Ablehnung. Auch lässt sie sich bei Personen beobachten, die Angst vor dem Zahnarzt haben oder in Panik geraten. Wie auch bei allen anderen Gesichtsausdrücken offenbart sich Angst in vielen Abstufungen des mimischen Ausdrucks. Allerdings kann man drei wesentliche Merkmale festhalten, die immer auftreten und an denen immer die gleichen Muskeln beteiligt sind: Hochgezogene Augenbrauen, geweitete Augen und geöffneter Mund mit straffen Lippen.

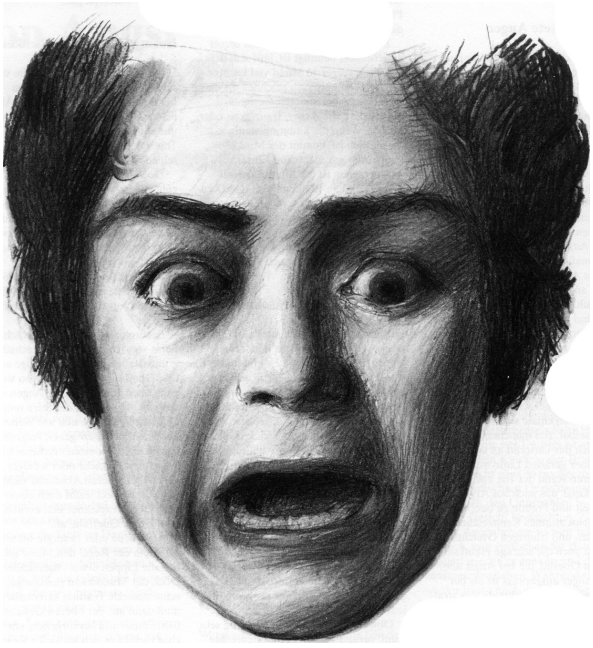


Abbildung 2.5: angsterfüllter Ausdruck [Fai98]

Angst zeigt sich zum einen bei Schreck durch Aufreißen der Augen und dem Hochziehen der Brauen. Als Unterschied zum erstaunten Gesichtsausdruck ist Angst durch eine zusätzliche Muskelkontraktion des Augenbrauenrunzlers charakterisiert, der Brauen zum einen nach unten und zum anderen nach oben zieht. Jeder Mensch, der Angst hat, befindet sich im Alarmzustand. Alle Sinne sind darauf ausgerichtet, eventuelle Gefahrenquellen zu lokalisieren. Diese erhöhte Aufmerksamkeit resultiert also in geweiteten Augen. Hinzu kommt das automatische Öffnen des Mundes in Gefahrensituationen. Darwin erklärt dies folgendermaßen: Die Emotion Angst erhöht die Atemfrequenz, somit ist das Öffnen des Mundes eine vorbereitende Maßnahmen für eine schnelle Flucht. Durch Zittern, das ebenso von Angst ausgelöst wird, wird der Plytysma-Muskel aktiviert, der eine Dehnung des Mundes bewirkt.

Eine weitere Form von Angst ist der Schock, der in einer Schrecksekunde auftritt. Dieser wird ausgelöst durch einen Moment, in dem etwas unerwartet entsetzliches passiert. Der Schock ist ein Extremzustand, der unvorhersagbare Folgen haben kann. Des Weiteren gibt es

verschiedene Ausdrücke und Grade von Angst. Schrecken, Furcht, aber auch Besorgnis oder Verdruss zählen dazu. Für Traurigkeit beispielsweise stehen nicht so viele unterschiedliche Wörter zur Verfügung.

2.1.5 Ekel

Bei Ekel fallen jedermann sofort viele Dinge ein, die als abstoßend empfunden werden. So reicht der Ekel von Ablehnung gegenüber einem Ex-Liebhaber bis hin zu Nasenrümpfen bei verdorbenen Lebensmitteln. Immer treten aber die gleichen Formen der Gesichtszüge auf. Die Oberlippe hebt sich in der Mitte und die Nase wird gerümpft. Die Mimik zeigt sich also bei echter Abscheu oder wird auch als kommunikative Geste eingesetzt.

Abbildung 2.6: Ekel [Fai98]

Hinter dem Nasenrümpfen steht eine heftige körperliche Reaktion: Das Erbrechen. Ekel ist der mimische Hinweis darauf, welche Reaktionen bestimmte Dinge auslösen können. Die Mimik wird bereits durch eine unbewusste Assotiation hervorgerufen. Die Parallelen zum tatsächlichen Erbrechen sind deutlich: Der Oberlippenheber übernimmt die Hauptaktion. Der innere Muskelstrang spannt die Oberlippe um die Kiefer, biegt die Lippe in eine quadratische Form und zieht die Nase an den Nasenflügeln nach oben. Bei diesem Gesichtsausdruck spielt die Nase eine wichtige Rolle. Das Nasenrümpfen ist ein unverwechselbarer Bestandteil der Mimik Ekel.

Ekel kann von allem durch Dinge hervorgerufen werden, die entweder den Geschmacks- oder den Geruchssinn empfindlich stören. Das bedeutet aber nicht, dass es unbedingt schlecht riechen muss. Vielmehr hängt es davon ab, was die Nase gewöhnt ist. Nach einer durchzechten Nacht beispielsweise kann einen Menschen schon allein der Geruch von Alkohol am nächsten Tag ekeln.

Die Besonderheit bei Ekel ist, dass er aufgrund einer körperlichen Empfindung ausgelöst wird und somit keinen emotionalen Zustand darstellt, aber trotzdem zu den Basisemotionen zählt.

2.1.6 Überraschung

Überraschung ist ein sehr spontaner und flüchtiger Gesichtsausdruck ist, der sich nur für Bruchteile von Sekunden abspielt. Es gibt wenige Gelegenheiten im täglichen Leben, bei denen sich dieser Ausdruck eingehend studieren ließe. Selbst für Schauspieler ist es schwierig, Überraschung zu zeigen. Das Problem dabei ist, dass, falls die Situation nicht wirklich überraschend kommt, der Gesichtsausdruck auch nicht echt wirkt. Am einfachsten lässt sich Überraschung pantomimisch darstellen. Alles was man tun muss, ist, die Augen weit zu öffnen und die Augenbrauen hochzuziehen. Die Bedeutung ist leicht zu verstehen: „Oh, wirklich?"

Die Überraschung kennzeichnen zwei Charakteristika: Die großen Augen und der plötzlich offen stehende Mund. Nach [Fai98] begründet Charles Darwin diese Vorgänge damit, dass man mit den Augen schnell alles ungewöhnlich erfassen möchte und sich zugleich wie bei der Angst auf eine eventuelle Flucht vorbereitet. Wenn jemand plötzlich von etwas überrascht wird, „setzen alle Muskeln einen Moment lang aus" und der Mund klappt durch sein eigenes Gewicht nach unten. Der geöffnete Mund schafft die Voraussetzung dafür, dass man schnell Luft holen kann.

Abbildung 2.7: Überraschung [Fai98]

2.2 Erzeugung von Emotionen

Im vorherigem Abschnitt wurden die Grundeigenschaften als auch Merkmale von Emotionen erläutert und die Einteilung in sechs Basisemotionen vorgenommen und begründet. Um aber Verfahrensweisen für die Erzeugung und die gewünschte Aufzeichnung ableiten zu können, sind Erkenntnisse über die Entstehung dieser beim Menschen sehr hilfreich. Aus diesem

Grund wird der Emotionsbegriff erläutert, kurz auf verschiedene Ansätze zur Erzeugung eingegangen und schließlich eine Umsetzung vorgestellt.

Abbildung 2.8: Beispiel: Stille Trauer nach [Fai98]

2.2.1 Definition

> ***Emotion*: Gemütsbewegung, seelische Erregung; Gefühlszustand; vgl. Affekt**

Eine Emotion ist nach Rainer [Rai04] definiert als ein Prozess. Emotionsepisoden sind dynamische Prozesse, die von konkreten Ereignissen ausgelöst werden und von relativ kurzer Dauer sind. Als Emotionen oder umgangssprachlich schlicht Gefühle werden spezielle psychische Phänomene bezeichnet, die ihren Ursprung sowohl extern, also außerhalb des Individuums haben können, wie auch internen Reizen folgen können. Emotionen wie auch ihre Auslösereize können selten oder häufiger auftreten. Auch ihre Intensität kann - abhängig von äußeren wie von inneren Gegebenheiten her - schwanken. Solche Zustände werden auch globaler und umfassender als die Stimmung des Menschen bezeichnet. Emotionen haben einen mehr oder weniger klaren Beginn und ein Ende, sind also keine (verhältnismäßig) überdauernden Erscheinungen wie beispielsweise der sog. Charakter. Vor und nach der Emotion befindet sich das Individuum in einem neutralen bzw. schon wieder veränderten Gefühlszustand.

Um Emotionen auszulösen, ist es hilfreich zu wissen, wie sie entstehen und woher sie kommen. Dazu gibt es eine Theorie von James-Lange (1885), der dies folgendermaßen erklärt:

> Zuerst erfolgt die Wahrnehmung eines Ereignisses, worauf der Körper mit einer spezifischen neurophysiologischen Reaktion antwortet. Aus dieser Reaktion ergibt sich schließlich eine physiologische Veränderung, die Emotion.

Ein weiterer, allerdings gegensätzlicher Ansatz für die Entstehung einer Emotion liefert z.B. Lazarus (1966):

> Aus der Wahrnehmung eines Ereignisses und der darauf folgenden unspezifischen Erregung entsteht die Emotion als kognitive Bewertung.

Obwohl diese beiden Ansätze gegensätzliche Theorien über die Ablaufursache der Emotion vertreten, ist der eigentliche Auslöser aber derselbe: Die Wahrnehmung eines Ereignisses.

2.2.2 Auslösereize

Auslöser für Emotionen finden sich besonders im alltäglichen Leben, allerdings können diese nur selten „vorhergesagt" bzw. erwartet werden. Die meisten dort gezeigten Emotionen treten zum größten Teil nur in Kontakt mit anderen Individuen auf. Die positiven oder negativen Ereignisse, die mit zu den entsprechenden Emotionen führen, können verschiedene Ausprägungen in ihrer Qualität wie Quantität aufweisen; sowohl Intensität als auch Häufigkeit können bei Alltagsbegebenheiten stark schwanken. Insofern und aus aufnahmespezifischen Gründen sind Alltagssituationen für die Aufnahme einer Emotionsdatenbank nicht geeignet. Große Ereignisse wie Kriege oder Katastrophen oder positive Life-Events wie Geburt, Heirat oder Beförderung erzeugen sehr realistische Emotionen. Diese treten aber im Allgemeinen selten auf und sind kaum berechenbar oder vorhersagbar und aus diesem Grund ebenso unpraktikabel für diese Anwendung.

Um die normalerweise zeitlich begrenzten und schwer vorhersagbaren Gefühle beim Menschen zu erfassen, ist die Emotionspsychologie bemüht, standardisierte und für den Probanden schwer durchschaubare Methoden zu deren Induktion zu finden. Möglichkeiten bestehen beispielsweise durch Darbietung von Bildern und Dias, Filmen, Hörbeispielen oder der Imagination von Situationen. Eine Vorgehensweise besteht darin, den Probanden aufzufordern, sich möglichst lebhaft an gefühlsauslösende Momente zu erinnern. Die andere versucht, eine Person durch Lesen von kurzen Aussagen aktiv in bestimmte Stimmungen zu versetzen. Dabei zeigen sich allerdings bei der Hälfte der Personen signifikante Unterschiede von Verhaltensmaßen, andere reagieren überhaupt nicht darauf. Für die hier gestellten Anforderungen an die Aufnahme der Datenbank ist dies unpraktikabel, weil die gewünschte Emotion oft falsch oder zu schwach ausgeprägt ist und eine Aufzeichnung und Kategorisierung oft nicht möglich wäre.

Die Möglichkeit durch Dias oder Bilder Emotionen wie beispielsweise Ekel hervorzurufen werden häufig eingesetzt. Allerdings sind Mimiken wie Ärger oder Angst nicht realisierbar. Für die Anforderungen dieser Arbeit weist das Medium Film das größte Potential auf den gewünschten Effekt und die besten Ergebnisse zu erzielen.

2.2.3 Medium Film als Auslöser

Filme dienen ebenfalls zur Erzeugung gewisser Gefühlslagen, wobei sich spezielle Darstellungen im Laufe verschiedener Studien als besonders geeignet erwiesen. Auch hat man durch

moderne Schnitt- und Manipulationstechniken die Möglichkeit, gewisse Sequenzen für die Induktion einzelner Emotionen zu adaptieren. Unterschiedliche Emotionen wie Traurigkeit, Heiterkeit und Freude, wie auch globalere positive wie negative Stimmungen können durch geeignetes Filmmaterial ausgelöst werden. In Studien hat es sich als praktisch erwiesen, Filme zur Angstauslösung an Personengruppen anzupassen. Beispielsweise wurde der Gruppe Raucher ein Film gezeigt, in dem einem Nikotinsüchtigen die Diagnose Lungenkrebs gestellt wurde. Durch die Identifikation mit dem Betroffenen und der für diese Zielgruppe realistische Diagnose konnte die Emotion Angst dadurch besonders gut erzeugt werden.

Für die vorliegende Anwendung und im Hinblick auf die Erstellung der Datenbank ist diese Spezialisierung weniger erwünscht, da von vielen unterschiedlichen Personengruppen eine möglichst umfassende Datenbank erstellt werden musste. Ebenso sollte das erstellte Filmmaterial für alle Personen verwendet werden können. Die negativen Auswirkungen dieser allgemeinen Anforderungen musste durch die im folgenden Abschnitt beschriebene Lösung kompensieret werden. Diese Lösung hat sich am praktikabelsten für die gegebenen Anforderungen herausgestellt.

2.2.4 Angepasste Emotionserzeugung

Aufgrund der vorher beschriebenen Erfahrungen wurde die Verfahrensweise zur Anregung der Emotionen weiterentwickelt. Dabei wurde versucht, möglichst allgemeine, nicht auf Personengruppen zugeschnittene Filme zu zeigen und damit trotzdem eine eindeutige Emotion anzuregen.

Ziel dieser Datenbank ist es u.a. möglichst realistische und natürliche Emotionen aufzuzeichnen, da erwiesen ist, dass bei gespielten Emotionen andere Muskelgruppen bewegt werden als bei den realen. Würde man zufällige Filmaufnahmen verwenden, müssten diese erst unter hohem Zeitaufwand vorverarbeitet und segmentiert werden. Außerdem ist dabei nicht sichergestellt, dass alle Aufnahmen unter den gleichen Lichtverhältnissen, Abständen und Blickrichtungen aufgenommen werden würden. Um also eine konsistente[1] Datenbank zu erstellen, die auch jederzeit erweiterbar ist, ist ein Kompromiss zwischen beiden Möglichkeiten zu finden. Zum einen muss die Testperson wissen, welche Emotion sie zeigen soll und zum anderen sollte diese Emotion dann von dieser nicht gespielt, sondern durch einen äußeren Reiz angeregt werden. Dabei entsteht das Problem, dass jeder Mensch auf bestimmte Situationen anderes reagiert. So findet einer lustig, was den anderen ärgert.

Die Schwierigkeit bei der dynamischen Emotionsaufnahme besteht darin, die einzelnen Emotionen erstens möglichst realistisch und spontan zu erzeugen, diese aber gleichzeitig im richtigen Moment aufzunehmen, und zweitens bei jedem Probanden auch die gleichen Emotionen hervorzurufen. Als Lösung dieses Problems wurde folgende Vorgehensweise entwickelt:

Der Versuchsperson wird zu Beginn jeder Emotion ein Video vorgespielt, welches die folgende Emotion ankündigt, dem Probanden die Gelegenheit gibt sich auf die folgende Emotion einzustellen und ihm zudem zeigt, wie die Mimik aussehen soll. Dabei kann sich die

[1] Begriff Konsistenz: siehe Kapitel 3.1, Seite 15

Person gedanklich damit identifizieren und darauf vorbereiten. Zudem wird sichergestellt, dass die Person richtig auf die folgenden Filmausschnitte reagiert und damit das Problem der Nichteindeutigkeit beseitigt.

Im Folgenden werden dann zu der jeweiligen Emotion angepasste Filmausschnitte gezeigt. Diese haben die Aufgabe die entsprechende Emotion hervorzurufen. Durch die Beschränkung der Aufzeichnung auf bestimmte Schlüsselstellen im Video und durch die Vorgabe der Emotion kann die jeweilige Mimik zeitlich genau in der Datenbank abgelegt werden.

Bevor der entwickelte Versuchsaufbau und das verwendete Filmmaterial eingehend erläutert wird, wird zunächst der Aufbau und die Anforderungen der zu erstellenden Datenbank definiert.

Kapitel 3

Datenbank

Ein wesentlicher Bestandteil der Arbeit ist die Erstellung einer Mimikdatenbank, die die in Kapitel 2.1 beschriebenen sechs Basisemotionen sowie den neutralen Zustand enthält. Die Datenbank soll als Grundlage für die dynamische Mimikerkennung in der Mensch-Maschine Kommunikation dienen sowie für weitere Anwendungsbereiche zur Verfügung stehen.

3.1 Grundlagen

Eine Datenbank ist nach Zehnder [Zeh98] „eine selbständige, auf Dauer und für flexiblen und sicheren Gebrauch ausgelegte Datenorganisation, die einen Datenbestand (Datenbasis) und die dazugehörige Datenverwaltung umfasst". Der Datenbestand oder auch die Nutzdaten sind „die Gesamtheit der gespeicherten Datenwerte, die den Inhalt der Datenbank bilden."

Die Datenbank muss für die Korrektheit gewisse Funktionen erfüllen. Eine davon ist die *Konsistenz*. Gemäß obiger Definintion ([Zeh98]) ist diese folgendermaßen definiert: „Konsistenz ist die Freiheit von Widersprüchen innerhalb der Datenbank. Diese ist gegeben, wenn der Inhalt der Datenbank alle vordefinierten Konsistenzbedingungen erfüllt". Im Fall der Emotionsdatenbank wäre das die korrekte Abspeicherung der jeweiligen Mimik im korrekten Verzeichnis, so dass keine Überschneidungen oder Fehlzuweisungen entstehen. Ein weiterer Begriff im Zusammenhang mit den Anforderungen an die Datenbank ist die *Datenqualität*. Laut [Zeh98] ist sie „ein mehrdimensionales Maß für die Eignung von Daten, den an ihre Erfassung/Generierung gebundenen Zweck zu erfüllen. Die Eignung kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern." Wichtig ist auch die Gewährleistung der *Integrität*. Nach [Zeh98] bedeutet Integrität, dass die bestehende Datenbank nicht mehr verändert wird, also Struktur und Inhalte gleich bleiben. Eine Erweiterung der bestehenden Datensätze ist möglich und erwünscht. Allerdings muss gewährleistet sein, dass weder die alten Daten überschrieben noch die Struktur verändert werden kann.

3.2 Existierende Datenbanken

Für dynamische Sequenzen existieren im Gegensatz zu statischen Bildern weltweit nur wenige Datenbanken, von denen die wichtigsten hier kurz beschrieben werden sollen:

Im Zusammenhang mit einem der verbreitesten Gesichtserkennungssysteme existiert eine Datenbank, erstellt von der Forschungsgruppe um Donato [Don99], in der 24 Personen Sequenzen von 150 unterschiedlichen Gesichtsaktionen der Dauer von je 6-8 Frames zeigen. Die Datenbank enthält sowohl schwache, stärkere als auch ausgeprägte Bewegungen der einzelnen Bereiche und diente dazu, alle möglichem Aktionsbereiche zu klassifizieren und diese zum Training sowie zur Erkennung zu verwenden. Allerdings zeigt diese Datenbank nur Ausschnitte der oberen und unteren Gesichtesbereiche. Aufnahmen mit Kopfbewegung wurden zudem ausgeschlossen. Auch Black und Yacoob haben für ihre Klassifikationen eine Datenbank aus 70 Sequenzen der Größe 540x420 Pixel erstellt, die insgesamt 128 Ausdrücke beinhalten. Es handelt sich dabei wiederum um gestellte Aufnahmen. Eine weitere Datenbank in diesem Bereich entwickelte Essa und Pentland für ihre Untersuchungen mit den Flow- und Muskelmodellen zur Erkennung der Gesichtsausdrücke. Dazu wurden 20 Leute aufgefordert die sechs Basisemotionen zu spielen und mit einer Auflösung von 450x380 Pixel gefilmt. Dabei wurde von Schwierigkeiten der Personen bei dem Ausdruck von Traurikeit berichtet und deswegen diese Emotion von den Untersuchungen ausgeschlossen. Genaueres zu ähnlichen Beobachtungen als Ergebnis dieser Arbeit siehe Kapitel 4.7 auf Seite 36.

Zwei weitere Datenbanken wurden nach Cohen [Coh02] von Chen und Huang sowie von Cohn-Kanade erstellt. Erstere besteht aus fünf Personen mit jeweils sechs Aufnahmen pro Emotion. Die Mimiken beginnen und enden dabei in dem neutralen Zustand mit einer Länge von 60 Bildern. In der zweiten wurden 104 Personen aufgenommen, allerdings zeigen nur die wenigsten alle Emotionen. Im Mittel sind die Mimiken sechs Frames lang und enden im Bild mit der größten mimischen Ausgeprägtheit.

In diesem Zusammenhang ist noch die Datenbank, erstellt von Hülsken, Müller und Wallhoff [FH01] zu nennen, welche 96 Sequenzen von sechs Personen mit der Auflösung 320x240 Pixel einer durchschnittlichen Dauer von 15 Frames enthält. Dabei wurden allerdings nur vier Emotionen aufgezeichnet. Diese „alte” Datenbank wurde in der vorliegenden Arbeit dazu verwendet, das vorhandene Erkennungssystem zu reaktivieren und zu überprüfen, was ausführlich in Kapitel 6.3, Seite 54 erläutert ist.

Die Motivation zur Erstellung einer neuen Emotionsdatenbank war erstens die Inkonsistenz bestehender Datenbanken (z.B das Fehlen entscheidender Klassen bzw. andere Anforderungen) sowie zweitens, dass zu wenig Trainingsdaten für eine robuste und signifikante Bewertung vorhanden waren.

Die Neuerung der in dieser Arbeit erstellten Datenbank ist die Aufzeichnung dynamischer Emotionssequenzen der vorher besprochenen Klassen, die natürliche und spontane Mimiken enthalten sollen. Außerdem sollte die Datenbank umfangreicher als die bisherigen sein, um mehr Trainings- und Testdaten zu Verfügung zu haben.

3.3 Anforderungen

Die Anforderungen den Aufbau betreffend werden zum einen durch die sechs Basisemotionen sowie den neutralen Zustand vorgegeben, zum anderen durch die aufgenommenen

Personen. Der genaue Aufbau wird im Anschluss (siehe Grafik 3.2) dargestellt. Zunächst wurden aber allgemeine Anforderungen definiert:

- **Struktur**: Aufnahme von sieben Mimiken verschiedener Personen; unbestimmte Anzahl von Personen[1]; Möglichkeit der Aufnahme einer Datenbank für die Blickwinkel- und Zeigerichtungserfassung; einheitliche Definition der Verzeichnis- und Bildnamen.

- **Spezifikationen:** Farbige Einzelbilder abgespeichert in einem verlustfreien Bildformat (Farbtiefe 24 bit) der Größe 640x480, ganannt PPM (Portable Pixel Map); Bildwiederholrate: 25 Frames/sec; gleiche Beleuchtungs- und Aufnahmebedingungen für jede Person, insbesondere gleicher Abstand der Aufnahme und identisches Setup[2]; ausreichende Beleuchtung; Aufzeichnung der für die Mimik charakteristischen Bereiche[3].

- **Sonstiges:** Aufnahme von verschiedenen Personen (männl. und weibl.); möglichst natürliche und spontane Emotionen; Vorsegmentierung und Integrität der Daten; leichte Erweiterbarkeit.

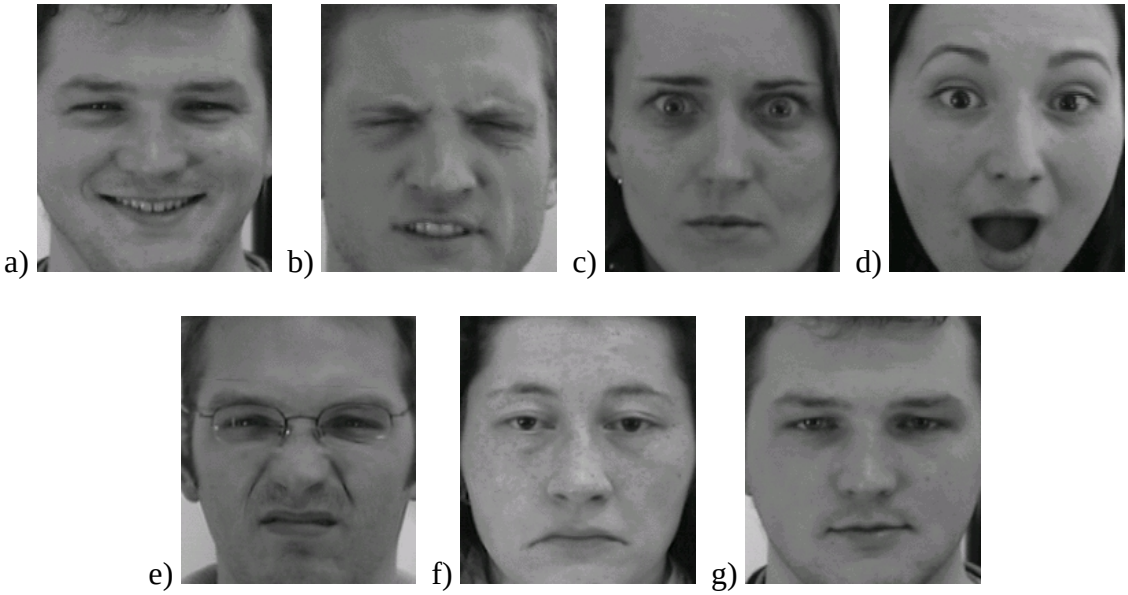


Abbildung 3.1: Die sieben Emotionen der Datenbank: a) Freude, b) Ärger, c) Angst, d) Überraschnung, e) Ekel, f) Traurigkeit und g) Neutral

3.4 Aufbau und Datenstruktur

Die Grundstruktur kann aus Abbildung 3.2 abgeleitet werden.

[1] maximale Anzahl wird durch Wahl des Dateinamens auf 9999 Personen beschränkt
[2] Aufbau und Spezifikationen des Setups siehe Kapitel 4
[3] Maximaler Zoom: vom Unterkiefer bis oberer Stirn, minimal: Kopfgröße gleich halbes Bild

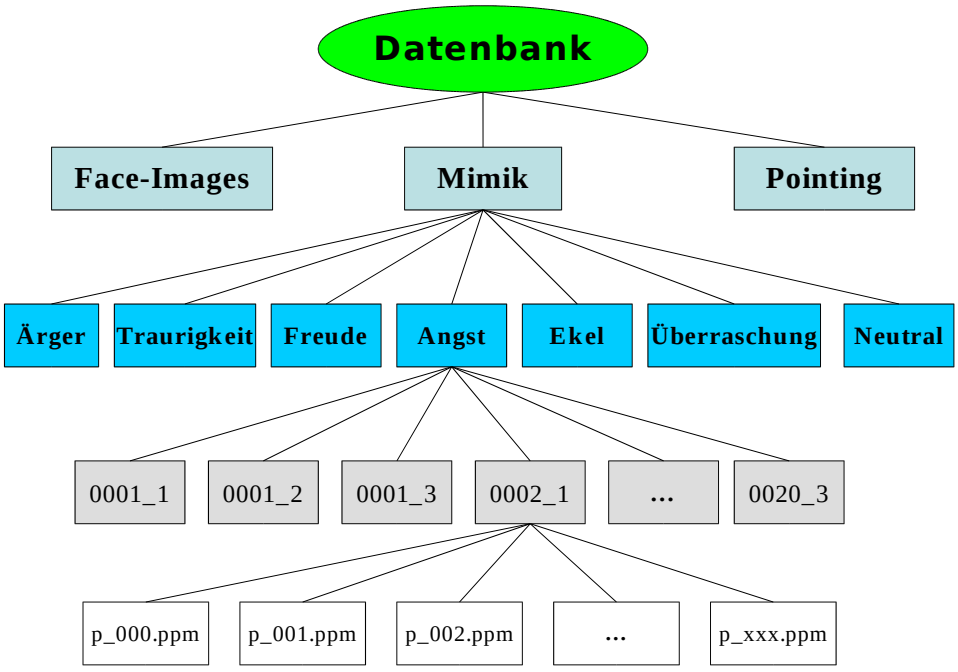


Abbildung 3.2: Struktur der Datenbank

An oberster Ebene stehen hier zunächst die drei existierenden Hauptdatenbanken, die einmal für die Blickwinkelerfassung, des weiteren für das Pointing[4] und schließlich die erstellte Mimikdatenbank. Letztere wird weiter unterteilt nach den einzelnen Emotionen, welche wiederum in Personenverzeichnisse und auf unterster Ebene in die aufgenommenen Einzelbilder unterschieden werden. Die beiden anderen enthalten jeweils die Personenverzeichnisse, in denen die durch die Positionsdaten gekennzeichneten Bilder abgespeichert werden. Ausführliche Erläuterungen stehen hierzu in Abschnitt 4.2.2. Die folgenden Beschreibungen beziehen sich allerdings nur auf die erstellte Mimikdatenbank.

Die unbearbeiteten Rohdaten bestehen aus vorsegmentierten Einzelnbildern im „PPM" Format. Der Datenbestand enthält 20 verschiedene Personen, wobei pro Mimik für jede Person drei verschiedene Aufnahmereihen gemacht wurden. Insgesamt besteht er somit pro Emotion aus 60 Mimikaufzeichnungen, die die Änderung vom neutralen Zustand zur jeweiligen Emotion enthalten. Somit umfasst der Datensatz 420 Aufzeichnungen mit 100-300 Bildern je Aufnahme. Die Größe der unkomprimierten Bilder variert je nach Anzahl der gespeicherten Bilder von 50MB bis zu 170 MB je Emotion und Person.

3.5 Verzeichnisstruktur

Die Namensgebung für die Datenbankeinträge sieht wie folgt aus:

[4] für Blickwinkelerkennung und Pointing wurden die Grundlagen entwickelt, jedoch keine Datensätze aufgezeichnet

Mimikverzeichnisse: [EMOTION]
wobei EMOTION = {anger,happy,disgs,sadns,fears,surpr,neutr}[5]

Personenverzeichnisse: [PERSON]_[SET]
wobei PERSON = {0001, ..., 0020}
und SET[6] = {1,2,3}

Bilderbezeichnungen: p_[ZAHL].[FORMAT]
wobei ZAHL = {000, ..., 015}
und FORMAT = Bildformat (z.B ppm)

Werden beispielsweise alle Bilder der ersten Aufnahme von Testperson 3 der Mimik *happy* benötigt, sind sie unter folgendem Pfad zu finden:

$MIMIKDB[7]/happy/0003_1/p_*.ppm[8]

Die Datenbank ist auf der beiliegender DVD unter dem Verzeichnis

/Database/MIMIK

zu finden. In diesem Verzeichnis liegen zum einen die obig erwähnten Datensätze der einzelnen Mimiken, zum anderen finden sich dort alle von mir nachbearbeiteten und erstellten Bilder und Scripten zur Emotionserkennung.

So existiert dort ebensfalls eine speziell für die Mimikerkennung vorverarbeitete Datenbank mit bereits erstellten Emotionssequenzen, die exakt 16 Bilder pro Mimik und Person enthält und die genau das Auftreten einer Emotion zeigen. Diese liegt im Unterverzeichnis „Database/Preprocessed" und hat die gleiche Struktur wie die Originaldatenbank. Allerdings sind dort nicht mehr alle Personen vorhanden, sondern nur noch die für den Trainingsprozess ausgewählte Bilder. Eine Liste dieser findet sich unter „Database/list/" in der Datei *picdir* (Aufbau siehe Anhang **??**). Im folgenden sind neben der Hauptdatenbank die weiteren wichtigen Verzeichnisse erklärt, alle bezogen auf „/Database":

- /list: Listen der Trainings- und Testdaten
- /scripts: Perl- und Bash-Scripts für die gesamte Mimierkennung
- /ppm: vorverarbeitete Bilder in ppm-Format
- /feat: bearbeitete Bilder in pgm-Format und Merkmalsdateien
- /train: 2D- und 3D-Trainingsdateien der Bilder in HTK-Format
- /test: 3D-Testdateien in HTK-Fomat

Der genaue Aufbau des Setups mit Skizzen und Abständen sowohl zur Blickwinkelerfassung als auch zu Mimikaufzeichnung finden sich detailliert in Kapitel 4. Auch die verwendete Hardware, Konfigurationen der einzelnen Geräte und die Bedienung der der Software sind dort erklärt.

[5] bestehend aus je fünf Zeichen
[6] Aufnahmeversuch je Person
[7] $MIMIKDB = Heimatpfad der gesamten Datenbank
[8] * steht für alle Bilder in diesem Verzeichnis

Kapitel 4

Versuchsaufbau und Aufnahme

Ziel des Aufbaus ist es sein, die Aufnahme von Mimiken sowie die Blickwinkelerfassung von Testpersonen zu ermöglichen und diese in einer Datenbank gemäß Kapitel 3 abzuspeichern. Die Aufgabenstellung gliedert sich demnach in zwei Bereiche:

1. Entwurf eines Setups zur Erfassung verschiedener Blickwinkel und -richtungen von Testpersonen (Gaze-Detection), sowie Mimiken in einer Datenbank

2. Aufnahme einer Datenbank für die Mimikerkennung

Anfangs erstellte ich das Setup für die Erfassung der Blickwinkel, welches gleichzeitig für die Mimikaufnahme gedacht war. Aufgrund unterschiedlicher Anforderungen und auftretenden Problemen bei gleichem Setup für die Mimikaufzeichnung, wurde ein zusätzlicher Aufbau, speziell abgestimmt auf die Aufnahme von Emotionen, entworfen. Als erstes wird das Setup für die Blickrichtungserfassung und das Pointing vorgestellt.

4.1 Erster Versuchsaufbau

Die allererste Anforderung an das Setup war die Aufnahme von Kopfbewegungen für die sog. Gaze-Detection (Blickrichtungserfassung). Diese zielt darauf ab, verschiedene Kopfansichten von der Front- bis zur Seitenansicht sowie Bewegungen nach oben und unten aufzuzeichnen.

Dazu muss gewährleistet sein, dass ein möglichst großer Bereich der Kopfbewegung aufgefangen wird. Es gibt entweder die Möglichkeit, die Versuchsperson an vorher definierte Punkte im Raum blicken zu lassen oder eine Kamera um die Person zu bewegen. Da letzteres mit einigem Aufwand verbunden ist und sich nicht so leicht realisieren lässt, habe ich mich für die erste Variante entschieden.

Ziel ist es, möglichst viele Kopfpositionen aufzuzeichnen. Die Realisierung, die Person an festgelegte Punkte blicken zu lassen, hat den Nachteil, dass nur diskrete Positionen abgespeichert werden können. Deswegen wird als Lösung folgendes Setup eingeführt:

Ein Beamer projeziert ein sich bewegendes Objekt auf einen halbdurchlässigen Spiegel, der als Leinwand dient. Die Testperson bekommt die Anweisung, das Objekt mit dem

Kopf zu verfolgen. Hinter dem Spiegel befindet sich eine Kamera, die während der gesamten Versuchsdauer die Bewegungen des Kopfes erfasst. Diese werden als Einzelbilder mit 25 Frames/sec der Größe 640x480 Pixel in der Datenbank abgelegt. Die erfassten Positionen werden im Header der Bilddateien als Kommentar abgelegt. Der Vorteil dieser Anordnung ist, dass während der Bewegung kontinuierlich Positionen aufgezeichnet werden und der Versuchsperson keine expliziten Anweisungen über die Blickrichtung gegeben werden müssen. Die Steuerung erfolgt über das Objekt. Der Nachteil ist allerdings, dass erstens eine große Projektionsfläche benötigt wird und zweitens, dass die Person, um eine möglichst große Kopfdrehung zu erreichen, ziemlich nah vor dem Spiegel sitzen muss. Für den Abstand ergibt sich aber aufgrund der Projektion des Beamers und der Behinderung durch die Testperson selbst ein Minimum (siehe Bild 4.1).

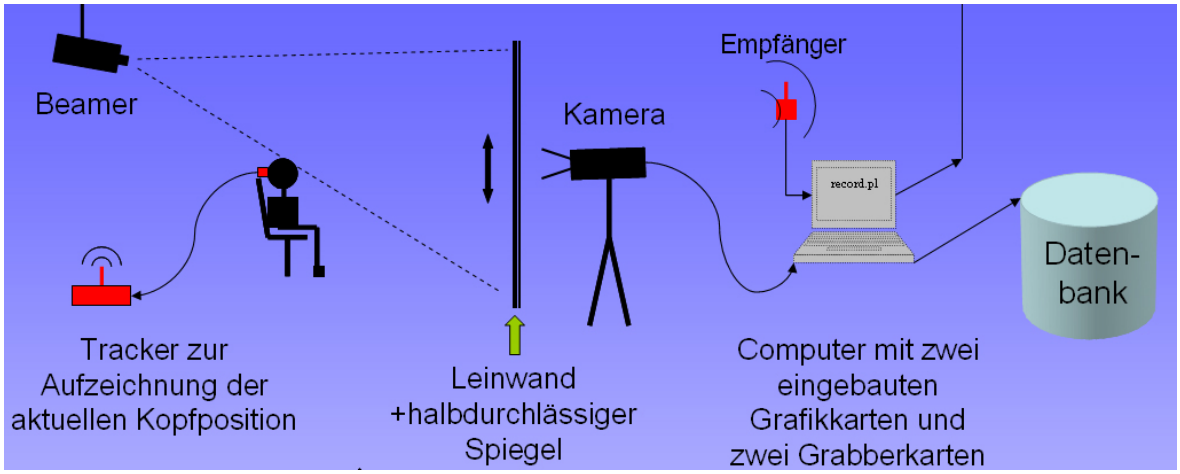


Abbildung 4.1: Schmatische Seitenansicht des Versuchsaufbaus für Blickwinkelerfassung und Pointing

Die Übermittlung der Kopfposition und -drehung erfolgt über den Iso-Tracker (geaue Gerätebeschreibung in Kapitel 4.2.2 auf Seite 24), einem Gerät, welches mit Hilfe elektromagnetischer Felder die Position eines Messkopfes in x, y und z-Richtung sowie deren Ausrichtung über die drei Raumdrehwinkel (Roll, Anzimuth, Elevation) ermittelt und über eine Schnittstelle an den Computer ausgibt. Dazu wird die Sonde am Hinterkopf der Testperson mit Hilfe einer Klammer befestigt.

Der Iso-Tracker übermittelt die Absolutpositionen und Winkel eingestellt auf die Ausrichtung und Position des Empfängers. Da nicht gewährleistet werden kann, dass die Sonde immer unter gleichen Winkeln und Ausrichtung an der Versuchsperson befestigt wird und eine kleine Abweichung schon zu großen Winkeländerungen führt, gibt es die Möglichkeit, die Winkel bei der gewünschten Ausrichtung auf Null zu setzen und den Tracker zu kalibrieren. Dadurch wird es realisierbar, vergleichbare und konsistente Daten zu gewinnen. Zur Kalibrierung ist es notwendig, dass die Blickrichtung der Personen vor der Aufnahme auf einen Ausgangspunkt ausgerichtet wird und danach die Nullpunkteinstellung erfolgt. Jede Testperson muss dazu auf ein Objekt in der Mitte der Leinwand blicken, welches gleichzeitig der Position der Kamera hinter der Leinwand entspricht. Die Kalibrierung erfolgt nun

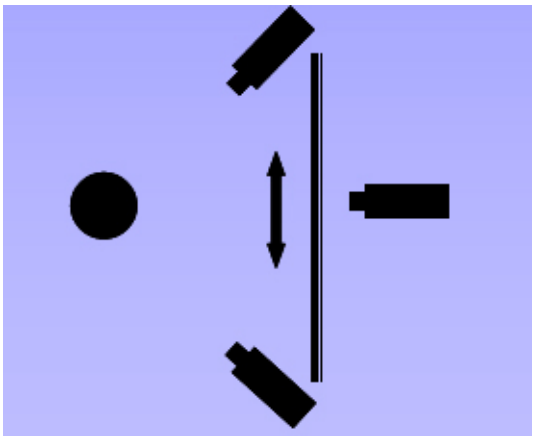


Abbildung 4.2: Draufsichts des Blickwinkelsetups

durch Tastendruck mit Hilfe der Aufnahmesteuerung, beschrieben in Kapitel 4.3 auf Seite 25).

Realisiert wurde dieser Aufbau im „Usability Labor" am Lehrstuhl für Mensch-Maschine Kommunikation der Technischen Universität München (TUM). Dort existieren bereits zwei Räume, getrennt durch einen halbdurchlässigen Spiegel. Hinter diesem lässt sich eine Kamera positionieren. Da der Betrachter aufgrund der Spiegelwirkung kein Bild sehen würde, ist der halbdurchlässige Spiegel bis auf den Ausschnitt, hinter der sich die Kamera befindet, mit weißem Papier abgeklebt. Der Beamer ist an der gegenüberliegenden Wand des Raumes unterhalb der Decke installiert, um ein möglichst großes Bild zu erzeugen. Die Person wird so vor den Spiegel positioniert, dass sie dem Bild nicht im Weg sitzt. So ergibt sich für die Entfernung der Sitzposition von dem Spiegel aufgrund der räumlichen Anordnung ein Abstand von ca. 1,40m (siehe auch Bild 4.1).

Da die Person bei gegebenem Abstand auf den äußeren Rand der Projektion blickend nur einen maximalen Winkel unterhalb von 90° erreichen kann, ist es nicht möglich die Seitenansicht der Person mit bestehender Anordnung aufzuzeichnen. Deswegen werden zusätzlich zu der Frontkamera noch zwei weitere Kameras genutzt, die von beiden Seiten gemäß der Draufsicht in Bild 4.2 angebracht sind. Somit ist es möglich in einem Durchgang alle Ansichten in der Datenbank abzulegen. Die Bilder der beiden Kameras werden über einen Analogmischer zu einem Bild zusammengefasst, gleichzeitig mit der Frontansicht aufgezeichnet und zusammen mit den Positionsdaten in der Datenbank abspeichert.

4.2 Anforderung und Ausstattung der technischen Geräte

Um den Anforderungen der Steuerung, Projektion sowie der gleichzeitigen Aufnahme von drei Kameras gerecht zu werden, ist ein leistungsfähiger Computer, ausgestattet mit nachstehend beschriebener Hardware, notwendig.

4.2.1 Computersystem

Als Betriebssystem wird ein unter Linux laufendes System mit Intel Pentium Prozessor (2,6 Ghz) und 300GB Festplattenspeicher für die Datenbank verwendet, der in einem doppeltem Raid System gegen Datenverlust gesichert ist.

Die Steuerung erfolgt über ein modulares, selbst entwickeltes Aufnahmeprogramm, beschrieben in 4.3 auf Seite 25. Angezeigt wird das Videosignal auf einem Bildschirm. Da der Beamer mit einem zweiten unabhängigen Videosignal angesteuert werden muss, ist eine zweite Grafikkarte erforderlich, die unter Linux problemlos eingebunden werden kann.

Für die Aufnahme von Front- und den beiden Seitenkameras wurden zwei „Grabber"-Karten eingebaut, die es ermöglichen, zeitgleich Bilder auf der Festplatte abzuspeichern. Die Seitenkameras lassen sich mit Hilfe einer elektrischen Steuerung auf die Versuchsperson ausrichten. Die Frontkamera muss per Stativ auf die Frontansicht ausgerichtet werden. Des Weiteren ist die Überwachung und Anzeige der Kamerabilder auf Monitoren möglich.

Die Person kann also vor und während der Aufnahme über Monitore überwacht werden. Der Zeitpunkt der Kalibrierung und das Einstellen der Kameras sowie der Sitzhöhe geschehen nicht automatisch und müssen manuell auf jede Person angepasst werden.

4.2.2 Isotracker

Um später eine Zuordnung der Blickrichtung des Kopfes mit der aufgezeichneten Ansicht in der Datenbank vornehmen zu können, muss diese gleichzeitig ermittelt und gespeichert werden. Dies erfolgt über die Positions- und Winkeldaten erfasst durch den Isotracker.

Der Isotracker ermittelt über elektromagnetische Felder die Abosolutposition und die Drehung einer Messsonde unter Angabe der drei Raumwinkel (Roll, Anzimuth und Elevation) gegenüber einem definierten Ursprung. Die Positionen der x,y und z Koordinaten werden entweder in *inch* oder *cm* ausgegeben, die Raumdrehwinkel in Grad.

Angeschlossen wird dieses Messgerät über die serielle Schnittestelle des PC's, die auf eine bestimmte Übertragungsrate eingestellt sein muss. Des Weiteren ist die Kalibrierung einiger Werte der Schnittstelle nötig. Die Kommandozeile mit den genauen Einstellungen befindet sich im Anhang A.

Zum Aufbau des Isotrackers sollte sich der Empfänger aufgrund von Störquellen möglichst nahe bei der Messsonde und in einem gewissen Abstand von Metallgegenständen befinden, um die Messung der Position nicht zu verfälschen.

Die Ansteuerung und das Auslesen der Werte wird vom entwickelten Steuerprogramm übernommen. Um das Messgerät zu initialisieren, sind Voreinstellungen des Gerätes über Kommandozeilenbefehle nötig. Dazu wird das Gerät vom Programm neu gestartet, die Maßeinheiten auf metrisches System und auf Einzelausgabe der Positionen umgestellt. Die Kalibrierung geschieht ebenso durch das Aufnahmeprogramm. Genaueres dazu siehe Programmtext auf beiliegender CD.

Nachdem die technischen Voraussetzungen geklärt sind, wird nun auf die Funktionsweise der Aufnahmesteuerung sowie anschließend auf den Ablauf eingegangen.

4.3 Steuerungssoftware

Das im Rahmen dieser Diplomarbeit entwickelte Programm hat die Aufgabe, die Benutzerführung, die Aufnahmesteuerung sowie das Ablegen der Daten in der Datenbank zu übernehmen. Gleichzeitig soll es für die Blickwinkelerfassung, die Gestik- sowie für die Emotionsaufzeichnung und eventuelle Erweiterungen ausgelegt sein. Zudem soll es übersichtlich, für jeden verständlich und leicht zu bedienen sein. Statt eines kommandozeilen-basierten Programms ist daher eine grafische Oberfläche (GUI = Graphical User Interface) sehr von Vorteil.

Leicht lassen sich grafische Benutzeroberflächen in Perl/Tk programmieren, welches die relativ einfache Programmiersprache Perl mit dem relativ einfach zu verstehenden Toolkit Tk verbindet. Damit können Perl-Programme statt über die Kommandozeile nun auch fensterbasiert bedient werden - mit Buttons, Eingabefeldern, Listboxen, Menüs und Scrollbalken. Tk ermöglicht dem Perl-Programmierern neben Kommandozeilenoptionen sowie der Standardein- und ausgabe auch grafische und ereignisgesteuerte Applikationen zu entwickeln.

4.3.1 Funktionen und Merkmale

Die Aufnahmesteuerung wird durch Start des Tools mit dem Namen „record.pl" zu finden im Verzeichnis „/Program" auf beiliegender CD in einer Shell unter Linux gestartet. Für das korrekte Funktionieren müssen zusätzlich die Dateien *Personencounter.index* und *Info.txt* (siehe 4.3.3). Die Oberfläche sieht folgendermaßen aus: Das Programm beinhaltet nachstehende

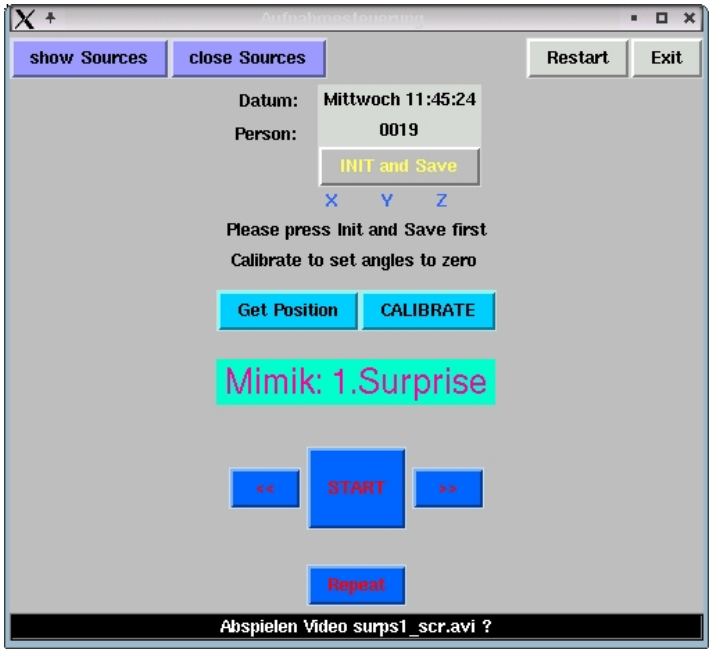


Abbildung 4.3: Aufnahmesteuerung record.pl

Funktionen:

- Anzeige des Datums
- Anzeige der laufenden Personennummer
- Anzeigen der Videosignale zum Monitoring der Eingänge
- Anzeige der Positionsdaten des Iso-Trackers
- Kalibrierung des Iso-Trackers
- Anzeige der aktuellen Aufzeichnung innerhalb des Ablaufprotokolls
- Start/Wiederholung der Aufnahme
- Statusleiste mit Zusatzinformationen
- Beenden/Neustart des Programms

4.3.2 Ablauf und Funktionsweise

Nach dem Aufruf des Programms wird die aktuelle Personennummer (wichtig für die Aufzeichnung und Benennung der Datenbankeinträge) aus der Datei *Personencounter.index* geladen. In dieser Datei wird die letzte Nummer gespeichert und bei Neustart ausgelesen, so dass die Verzeichnisse in der Datenbank in aufsteigender Reihenfolge angelegt werden können und keine alten überschrieben werden. Die Dateinamen und Verzeichnisse werden wie in Abschnitt 3.5 beschrieben, erzeugt. Vor Beginn der Aufnahme muss über die Taste „Init" die Initialisierung des Positionsmessgerätes durchgeführt werden. Dabei werden einige Einstellungen (Reset, metrische und nicht kontinuierliche Ausgabe, ASCII Format sowie Umstellung auf die positive Hemisphere) vorgenommen.

Für die *Blickwinkelerfassung* und die *Zeigebewegungsaufnahme* werden die Positionsdaten des Iso-Trackers benötigt, die mit der Funktionstaste „Get position" zur Überprüfung angezeigt werden können. Die Kalibrierung muss vor der Aufnahme mit „Calibrate" durchgeführt werden, ansonsten ist die Aufnahme blockiert. Für die Mimikaufzeichnung sind die Positionsdaten nicht notwendig.

Mit der Taste „Show sources" können die anliegenden Videosignale zur Überprüfung angezeigt werden. Diese werden allerdings bei Start der Aufzeichnung aus technischen Gründen von selbst beendet.

Unterhalb der Kalibrierung wird angezeigt, um welche Aufnahme es sich gerade handelt. Bei der Mimikaufzeichnung wird zudem eingeblendet, um welchen Versuch es sich handelt, da es sich in der Praxis als sinnvoll erwiesen hat, pro Emotion mehrere Aufzeichnungen[1] vorzunehmen.

Die Regelung der Aufnahme erfolgt mit den Steuertasten. Mit den Vor-/Zurück Tasten (»/«) können Aufnahmen übersprungen bzw. wiederholt werden. Dabei wird ein interner Zähler herauf- bzw. heruntergezählt. Mit dem „START" Button werden die Videos und die

[1] in disem Fall wurden 3 Versuche pro Mimik aufgezeichnet

Aufnahme gestartet. Der Ablauf ist Folgender: Das Programm prüft als erstes, ob die Initialisierung stattgefunden hat, und schließt dann eventuell geöffnete Videofenster, um die Aufnahme mit dem gängigen externen Bilderfassungsprogramm durchführen zu können. Als nächstes wird die Textdatei *Info.txt* geladen, in der sich alle für die Aufnahme nötigen Informationen (Pfade, Dateinamen der Videos, Anfang und Dauer der Aufzeichnung, Art der Aufnahme) befinden. Der Aufbau der Datei, die eine feste Formatierung aufweisen muss, wird in Kapitel 4.3.3 und im Anhang A.1 genauer ausgeführt. Von dieser Datei wird dann die entsprechende, durch den Zähler festgelegte Zeile ausgelesen und die Informationen in einem Vektor gespeichert. Das Programm überprüft als nächstes anhand eines Markers in diesem Vektor, um welche Aufnahme es sich handelt. Bei Blickwinkel- und Zeigerichtungserfassung wird nur ein Video gestartet und die Aufnahme (in diesem Fall von zwei Quellen) beginnt zeitgleich mit dem Video. Bei der Mimikaufzeichnung wird erst geprüft, ob es sich um den Anfang einer neuen Emotion handelt. In diesem Fall wird als erstes das Einführungsvideo abgespielt, nach zwei Sekunden Pause beginnt das eigentliche Emotionsvideo. Die Aufzeichnung startet aber erst nach einer bestimmten Anzahl von Frames, die ebenso in dem Vektor gespeichert ist, genauso wie die Dauer der Aufnahme in Sekunden. Bei Start des Videos erfolgt noch die Umschaltung auf den zweiten Bildschirm bzw. den Beamer. Schließlich wird der Zähler um eins erhöht, überprüft ob das Ende erreicht ist und die nächste Emotion angezeigt.

Falls die Aufnahme nicht geglückt sein sollte, kann die letzte Aufnahme mit der „Repeat" Taste wiederholt werden. Dabei wird, falls es sich um die erste Aufnahme einer Mimik handelt, das Einführungsvideo nicht mehr abgespielt. Eine mögliche Erweiterung wäre beispielsweise bei Wiederholung ein anderes für die Person passenderes Video abzuspielen, um eine bessere Reaktion zu erzielen. Bevor die Wiederholung startet, werden die vorher erstellten Bilder in dem Verzeichnis gelöscht, um Überschneidungen zwischen alten und neuen Aufnahmen zu vermeiden.

Der „New" Button beginnt die Aufnahmeserie einer neuen Person. Dabei wird der Zähler auf Null gesetzt, das System und eventuelle Parameter neu initialisiert und die Personennummer automatisch erhöht, um sofort mit der nächsten Aufnahmeserie beginnen zu können. Mit „Exit" wird das Programm verlassen.

4.3.3 Aufbau der Ablaufprotokoll-Datei

Die Datei *Info.txt* muss im selben Verzeichnis wie das Aufnahme-Programm liegen und einen exakten Aufbau aufweisen. Die Liste besteht aus Einträgen, die durch Tabulatoren bzw. Leerzeichen getrennt sein müssen. Eine Zeile steht für die Aufnahme einer beliebigen Mimik, der Zeigerichtung, der Blickwinkel oder eventuellen anderen Erweiterungen. Eine Beispielzeile der speziell auf die Aufnahme festgelegten Parameter sieht wie folgt aus:

Mimik	1.angry	anger_road.avi	/elmo0/misc	anger_new.avi	seq	0	100	MIMIK/anger	angry_s.jpg
Arg[0]	Arg[1]	Arg[2]	Arg[3]	Arg[4]	Arg[5]	Arg[6]	Arg[7]	Arg[8]	Arg[9]

Eine Zeile ist in 10 Spalten (Arg[0]-Arg[9]) unterteilt, wovon jedes Argument einen bestimmten Zweck (Parameter) erfüllt. Im einzelnen haben die Einträge folgende Bedeutung:

Erster Eintrag (Arg[0]) ist die Art der Aufzeichnung (also Mimik- oder Blickwinkelerfassung), der zweite (Args[1]) spezifiziert die Unterart. Im Fall von Mimik bedeutet dies zum einen um welche Emotion und zum anderen um welchen Versuch es sich handelt. Der Eintrag muss bei Mimikaufzeichnung auf jeden Fall [Zahl].[Emotion] enthalten, da das erste Zeichen der zweiten Spalte für den Dateinamen der Datenbank verwendet wird. Für die Blickwinkelerfassung gibt es nur einen Versuch, deswegen kann hier auf die Nummer verzichtet werden. Im dritten Feld (Arg[2]) steht der Dateiname des Videos, das vor der Aufzeichnung abgespielt werden soll. Das Aufnahmeprogramm sucht dabei den Dateinamen im festgelegten Heimatverzeichnis[2] unter „/VIDEO". In Feld Arg[3] wird der Pfad angegeben, auf dem das Dateisystem der Datenbank aufgebaut ist und auf den das Steuerprogramm zugreift. Im Eintrag Arg[4] steht die optional aufzurufende Intro-Video Datei, welche bei Beginn jeder neuen Mimik angegeben wird. Im zweiten und dritten Versuch, wie in meiner Ausführung implementiert, muss kein weiteres Einführungsvideo gezeigt werden. In diesem Fall steht ein [*] an dieser Stelle. Arg[5] beinhaltet die Parameter *seq*, *pic* oder *sim*. Diese werden als Flags benutzt, die intern vom Programm abgefragt werden. Steht dort *sim* (= Simulation), handelt es sich um die Blickwinkel- oder Gestikerfassung. Dabei wird zuerst die „Smiley"-Simulation, danach beide Aufnahmen gestartet. Steht in dem Eintrag *pic* (= Picture), weiß das Programm, dass es sich um die Aufzeichnung des neutralen Zustands handelt. Hierbei wird nur ein Bild gezeigt, danach die Aufnahme der Frontkamera gestartet; *seq* (= Sequence) steht für die Aufzeichnung der Mimiksequenzen. Das Feld Arg[6] gibt den Beginn der Aufzeichnung der Mimik nach Start des Videos in Sekunden an. Arg[7] bestimmt die Dauer der Aufnahme, allerdings muss die Angabe in der Anzahl der Einzelbilder [3] gemacht werden. Mit diesen beiden Werten ist es möglich, die Aufzeichnung der Mimik auf den entscheidenden Moment einzuschränken und dem jeweiligen Video anzupassen. Dies erleichtert somit erheblich die Nachbearbeitung und reduziert gleichzeitig die anfallende Datenmenge. Im Eintrag Arg[8] steht der Datenbankpfad für die jeweilige Aufzeichnung relativ zum Heimatverzeichnis in Feld Arg[3]. Im letzten Eintrag werden Dateinamen von Bildern, die der Versuchsperson zwischen den einzelnen Aufzeichnungen gezeigt werden, angegeben. Am Ende der Datei (in der letzten Zeile) muss im ersten Feld das Schlüsselwort „Ende" stehen. Somit erkennt das Programm das Ende der Textdatei und startet keine weitere Aufnahme mehr.

Die Parameterdatei erlaubt einfache Anpassungen (z.B. neue Videos) durch Abänderung der Einträge, ohne in das eigentliche Aufnahmeprogramm eingreifen zu müssen. Auch Erweiterungen können problemlos in der strukturierten Textdatei implementiert werden, wie beispielsweise mehr Versuche pro Mimik oder eventuell die Aufzeichnung von anderen Emotionen. Durch die Einspielung von Videos kann die Erzeugung möglichst spontaner Emotionen wenigstens in erster Näherung gewährleistet werden.

[2] Anpassung über Ablaufprotokolldatei *Info.txt*

[3] es werden 25 Frames pro Sekunde aufgezeichnet - für eine Aufnahmedauer von 3 Sek. ergibt sich also ein Wert von 75 Bildern

4.4 Ablauf der Blickwinkel- und Zeigerichtungserfassung

Aufgabe für die Testperson bei der Blickwinkelerfassung ist es, ein Objekt, welches sich auf einer vordefinierten Bahn auf der Leinwand bewegt, mit dem Kopf zu verfolgen. Programmiert wurde die Trajektorie eines Objektes (in diesem Fall ein Smiley) sowie dessen Bewegung in *Perl* mit der Bibliothek *ImageMagick* und kann jederzeit rekonstruiert, in Teilen verändert und angepasst werden. Die derzeitige Bewegung ist so gesteuert, dass möglichst alle Bereiche auf der Leinwand abgedeckt werden. So werden Geraden, quadratische Funktionen sowie Kreisbahnen durchlaufen. Ausgang der Bewegung ist der Mittelpunkt der Leinwand, auf den der Isotracker zuallererst kalibriert werden muss. Die Kalibrierung erfolgt wie in ... beschrieben.

Dem Probanden wird gesagt, er müsse mit dem Kopf, die Augen gerade gerichtet, der Bewegung des Smileys folgen. Gleichzeitig erfolgt die Aufnahme der Bewegung sowie der Positionsdaten (Koordinaten und Winkel) des Kopfes in Form von durchnummerierten und eindeutig zu identifizierenden Bildern (verlustfreies ppm-Format), in deren Header die Positionsdaten geschrieben werden, um sie dann in der Datenbank eindeutig identifizieren zu können. Durch die drei Kameras werden während der Aufnahme alle geforderten Ansichten und Positionen des Gesichts abgedeckt und dadurch eine lückenlose Datenbank ermöglicht.

Bei der Aufnahme des Zeigerichtung wird genauso verfahren wie bei der Blickwinkelerfassung, nur dass der Sensor für den Isotracker am Handgelenk befestigt wird und zwar so, dass er während des Pointings seine Ausrichtung relativ zum Arm nicht verändert. Die Kalibrierung wird danach durch Deuten auf den Center-Smiley durchgeführt. Die Einstellung der drei Kameras muss dabei auf die relevanten Teile des Körpers angepasst werden.

Im Anschluss wird die gleiche Bahn wie bei der Blickwinkelerfassung durchlaufen. Die Testperson muss diesmal mit ausgestrecktem rechtem Arm auf den Smiley zeigen und der Bewegung des Objekts folgen. Gleichzeitig erfolgt wieder die Aufzeichnung durch die drei Kameras. Die Daten werden ebenso in einem eigenen Ordner mit der Personennummer abgespeichert. Erweiterungen auf mehr Aufzeichnungen oder andere Abläufe sind jederzeit möglich.

Im Weiteren wird nicht mehr auf die Blickwinkel- und Zeigerichtungserfassung eingegangen, da der Schwerpunkt der vorliegenden Diplomarbeit bei der Erstellung der Mimik-Datenbank und der Mimikerkennung liegt.

4.5 Unterschiedliche Anforderungen an die verschiedenen Setups

Bei dem für die Blickwinkelerfassung optimierte Setup ergeben sich für die Mimikaufzeichnung einige Probleme:

Während bei der Blickwinkelerfassung eine Kopfdrehung erwünscht und sogar notwendig war, ist dies bei der Mimikaufnahme sehr von Nachteil, da von der Testperson in diesem

Fall nur die Frontansicht gefilmt wird. Dreht diese aber aufgrund des großen Bildes den Kopf, wirkt sich dies nachteilig auf die spätere Emotionserkennung aus. In den Anforderungen ist aber festgelegt, dynamische Mimiken der Frontansicht aufzuzeichnen. Auch würden die Grundsätze der Datenbank, also die Allgemeingültigkeit und Einheitlichkeit, verletzt werden. Zudem fällt es dem Betrachter bei gegebenem Abstand des Setups und dem großen Bild sehr schwer, die Details auf der Leinwand zu erkennen. Was bei einem Objekt, welches verfolgt werden soll, kein Problem ist, wirkt sich jedoch bei einer Filmszene welche Emotionen auslösen soll, sehr nachteilig auf den gewünschten Effekt aus.

Der halbdurchlässige Spiegel ist optimal, um der Testperson die Kamera vorzuenthalten und dessen Aufmerksamkeit auf das bewegende Objekt zu konzentrieren. Allerdings dämpft dieser Licht, welches der Kamera hinter dem Spiegel fehlt und wodurch damit die Aufnahmequalität verschlechtert wird. Als Lösung kann hierzu die Beleuchtung der Testperson erhöht werden, um den Lichtverlust auszugleichen.

Allerdings ergibt sich als Folge ein weiteres Problem: Die Diskrepanz zwischen der Beleuchtung und der Lichtstärke der Filmprojektion. Wie vorher beschrieben, ist eine ausreichende Beleuchtung notwendig, um die Kamera mit genügend Licht zu versorgen. Bei heller Umgebung sinkt aber aufgrund der begrenzten Lichtstärke des Beamers die Sichtbarkeit der Projektion. Bei dem sich bewegendem hellen Objekt für die Gaze-Detection stört die Umgebungshelligkeit nur bedingt, da der Rest des Bildes schwarz ist und das Objekt noch relativ gut wahrzunehmen ist. Handelt es sich aber um eine etwas dunklere Filmszene, ist es nahezu unmöglich, noch etwas auf der Leinwand zu erkennen. Senkt man die Ausleuchtung der Testperson, sinkt auch die Aufnahmequalität, die aber besonders für die Emotionsdatenbank und spätere Anwendungen sehr wichtig ist.

Die einzige Lösung dieser Probleme ist eine Trennung der Aufnahmekonditionen und ein Neuentwurf, abgestimmt auf die Mimikerkennung.

4.6 Zweiter Versuchsaufbau

4.6.1 Geändertes Setup für Mimikaufzeichnung

Das Setup (siehe Abbildung 4.4) für die Aufzeichnung der Mimiken ist aufgrund der einfacheren Anforderung leicht zu realisieren:

Erstens werden die fest installierten Seitenkameras sowie große Leinwand nicht benötigt. Für das Abspielen der Videos für die Emotionsdarstellung reicht ein einfacher Fernsehmonitor oder Bildschirm vollkommen aus. Die Kamera wird direkt auf dem Bildschirm befestigt. Die geringe Abweichung der Kamera vom Zentrum des Bildschirms, auf das die Aufmerksamkeit der Testperson gerichtet ist, kann leicht durch einen gewissen Abstand der Versuchsperson vom Bildschirm kompensiert werden, wie in Abbildung 4.4 veranschaulicht.

Der Winkelunterschied beim direkten Blick in die Kamera zum Blick auf die Bildschirmmitte verringert sich zunehmend mit größerer Entfernung vom Bildschirm. Bei einem Abstand von einem Meter ist kaum noch ein Unterschied feststellbar. Außerdem ist der Abstand

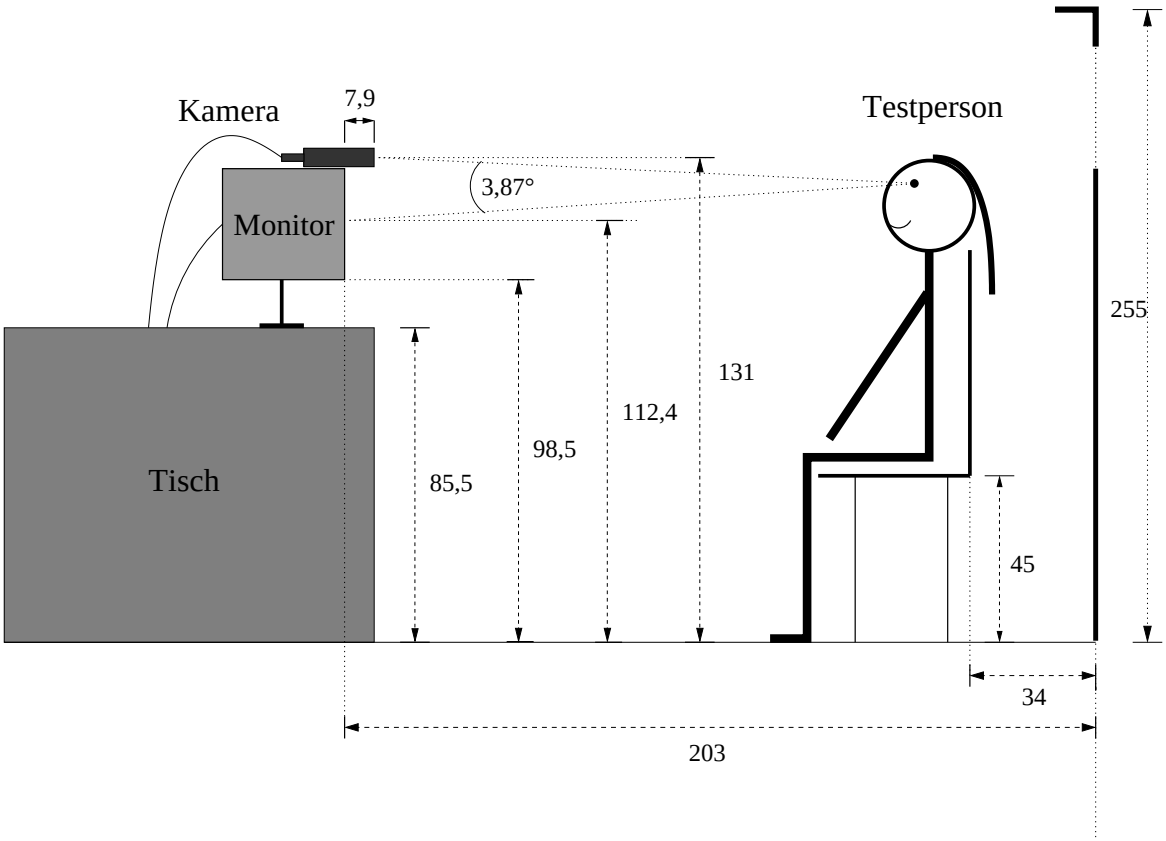


Abbildung 4.4: Vermaßte Seitenansicht des *Mimiksetups*; alle Maßangaben in cm

in Hinblick auf eine mögliche Emotionserkennung durch eine Webcam viel geringer, weil der Anwender im Normalfall direkt vor dem Bildschirm sitzt. Daher ist es durchaus sinnvoll, die Erstellung der Datenbank und damit das Setup diesen Umständen und Kriterien der späteren Anwendung anzupassen und die Emotionserkennung auf dieser Grundlage durchzuführen. Zudem hätte ein zu großer Abstand vom Bildschirm den Nachteil, dass das Bild sehr klein werden und der gewünschte Effekt der Emotionsanregung verloren geht. Deswegen wurde ein Abstand der Testperson von $1,7m$, wie in der Seitenansicht in Abbildung 4.4 ersichtlich, gewählt.

Wichtig für die Aufnahme der Datenbank bei diesem Setup ist die ausreichende Ausleuchtung der Testperson. Dazu werden zuerst alle Deckenlichter des Aufnahmeraums eingeschaltet und zusätzlich noch drei weitere diffuse Lampen aufgestellt, um eine möglichst gleichmäßige Beleuchtung des Gesichts zu erreichen. Sie müssen zum einen so aufgestellt werden, dass die Ausleuchtung der für die Mimik wichtigen Bereiche gleichmäßig geschieht (keine Schatten im Bereich des Gesichts) und zum anderen, dass die Testperson nicht geblendet wird. Dies wird dadurch erreicht, dass die Person nicht direkt angeleuchtet wird, sondern durch Reflexion an der Decke oder der Wand die allgemeine Raumhelligkeit erhöht wird. Generell gilt die Vermeidung von Spitzenlichtern.

Vorteil des neuen Setups ist die bessere Sichtbarkeit der Videos, eine eingeschränkte Kopfbewegung, kein Lichtverlust durch den Spiegel sowie realistischere Bedingungen.

Abbildung 4.5: Beispiel der Kameraansicht einer Testperson

4.7 Mimikaufnahme

Die Mimikaufzeichnung läuft aus bereits besprochenen Gründen (siehe 4.6) folgendermaßen ab: Zuerst erfolgen zwei Aufnahmen des neutralen Zustandes. Dem Probanden wird dabei jeweils ein Bild gezeigt, welches möglichst verhindern soll, überhaupt eine Emotion zu zeigen. Währenddessen wird für jeweils 75 Frames die Aufnahme gestartet und die Bilder entsprechend in der Datenbank abgelegt. Danach beginnt die erste Mimik Überraschung, zu deren Beginn ein besonders ausführliches Einführungsvideo gezeigt wird, welches den Ablauf der Aufzeichnung sowie Sinn und Zweck der Videos erklärt und dem Probanden hilfreiche Informationen zu seinem Verhalten gibt. Danach startet das eigentliche Emotionsvideo und an der entsprechenden Stelle die Aufzeichnung für den festgelegten Zeitraum[4] (Beispiel siehe Bild 4.6).

Hierzu ein Überblick zum zeitlichen Ablauf der Videos der Aufnahmeserie einer Person:

[4]genaueres zu Beginn und Dauer jeder Aufzeichnung siehe Anhang A.1, Datei *Info.txt*

Gezeigtes Video	Aufzeichnung Emotion
Bild 1	Neutral 1
Bild 2	Neutral 2
Einführungsvideo	/
Vorzeigevideo Überraschung	/
Video 1: Überraschung	Überraschung 1
Video 2: Überraschung	Überraschung 2
Video 3: Überraschung	Überraschung 3
Vorzeigevideo Ärger	/
Video 1: Ärger	Ärger 1
...	...
Video 3: Freude	Freude 3
Bild 3	Neutral 3

Der folgende Abschnitt beschreibt die Erstellung der unter Kapitel 2.2.4 erwähnten Videos zur Emotionserzeugung, die entscheidend für die erfolgreiche Aufzeichnung der richtigen Emotionen in der Datenbank sind und somit einen wichtigen Teil des Setups darstellen.

Die Bearbeitung der Videos wurde unter Windows XP vorgenommen, da nur dort die erforderlichen Programme und die Unterstützung des *Windows Media Video*-Formats (WMV) gegeben war, sowie weitere benötigte Codecs zur Verfügung standen. Die Aufnahme der Einführungsvideos wurde mit einer Digitalkamera des Lehrstuhls aufgezeichnet und im WMV-Format importiert.

Das Verzeichnis, in dem sämtliche fertigen Videos liegen und auf das die Aufnahmesteuerung *record.pl* zugreift, befindet sich unter:

Videoverzeichnis auf Elmo: /home/elmo0/home/haw/VIDEO/

4.7.1 Aufnahme und Schnitt der Intro's

Für das Einführungsvideo, in dem ich den Ablauf der Aufnahme erkläre, habe ich mich selbst mit der Digitalkamera vor schwarzem Hintergrund gefilmt. Für die sechs Videos, die die jeweilige Emotion vorzeigen, habe ich auch andere Personen gefilmt. Im Zusammenschnitt erfolgt jeweils von mir die Ankündigung der Emotion, danach wird eine Person gezeigt, die sie vorspielt. Daraufhin folgen dann die eigentlichen Emotionsvideos wie in Kapitel 4.7 beschrieben.

Die aufgenommen Videos wurde als WMV-Dateien abgespeichert, die ich unter Windows mit dem *Movie Maker* bearbeitet und geschnitten habe. Die entstehenden Dateien müssen, um sie unter Linux abspielen zu können, in *avi*-Dateien umgewandelt werden. Hierzu werden drei Tools benötigt:

Zur Umwandlung gibt es nur ein kostenloses Tool namens *Graphedit*, welches den Ton der WMV-Files extrahieren kann. Dazu müssen aber erst bestimmte Filter installiert werden, die mit dem Programm geliefert werden. Die noch benötigten Codecs sollten unter Standard-Windows XP installiert sein. Die einzelnen Filter werden im Baukastensystem mit Hilfe der

Abbildung 4.6: links: gezeigtes Videos, rechts: Reaktion der Person

grafischen Oberfläche zusammengesetzt. Mit dem *Tpm-encoder* wird im zweiten Schritt das Video extrahiert und in MPEG codiert. Dieses Format lässt sich problemlos unter Linux abspielen. Zuletzt werden mit *Virtual Dub* das Video und die Audio-Spur wieder zusammengefügt und zudem noch komprimiert und als *avi* abgespeichert. Das Codierverfahren kann dabei frei gewählt werden.

In den fertig geschnittenen Videos kündige ich nach der Einführung die erste Emotion[5] an und spiele sie dem Probanden auf dem Monitor vor. Nach zwei Sekunden Pause beginnt im Anschluss daran das erste Emotionsvideo.

4.7.2 Schnitt der Emotionsvideos

Für die Mimikanregung wichtigen Videos, die die Personen zum Zeigen der Emotion bewegen sollen, habe ich Filmausschnitte verwendet.

Für die Emotion Freude fanden sich viele passende Beispiele, die fast bei jedem den gewünschten Effekt erzielten. Schwieriger war es bei Überraschung sowie Trauer. Auch wirkliche Angst bei einer Person zu erzeugen, erwies sich als sehr schwierig. Ebenso ist es nicht möglich, innerhalb einer kurzen Zeit jemand ehrlich wütend zu machen. Vielmehr übernehmen die Videos die Aufgabe, die Person beim Ausdruck der Emotion positiv zu beeinflussen und zu unterstützen.

Die passenden Szenen aus den Filmen habe ich mit *VirtualDub* zusammen geschnitten und abgespeichert. Dabei handelt es ich um kurze Filmausschnitte von 5 bis 40 Sekunden Dauer. Bei Angst und Trauer wurden längere Szenen verwendet, da sich Trauer nicht so schnell einstellt. Ebenso muss die Angst erst aufgebaut werden. Bei den lustigen Szenen dagegen reicht eine kurze Sequenz mit einer Pointe um eine Testperson zum Lachen zu bringen.

Pro Emotion hab ich drei Videos ausgewählt, um erstens auch wirklich sicherzugehen, falls ein Ausschnitt nicht die gewünschte Emotion auslöst und zweitens um mit geringem

[5] in meinem Setup: Überraschung

Aufwand mehr Trainingsdaten zu gewinnen. Zudem bleibt die Option, die beste Aufzeichnung aus den drei Aufnahmen auszuwählen.

4.7.3 Evaluierung

In diesem Abschnitt möchte ich zusammenfassend die Wirkung der Videos bei der Aufnahme der einzelnen Emotionen beschreiben und Verbesserungen angeben. Die meisten der folgenden Erkenntnisse habe sich erst im Laufe der Aufnahmen gezeigt. Der Großteil der Personen reagiert unterschiedlich ausgeprägt auf die einzelnen Videos, allerdings ist die Grundrichtung dieselbe.

1. **Freude:** Diese Emotion ist am leichtesten zu erzeugen, da es viele Situationen gibt, die Menschen zum Lachen bewegen. So genügt schon ein kurzer Filmauschnitt mit einer lustigen Szene, um die Testperson zur Mimik Freude zu bewegen. Da die Pointe zudem genau festlegbar ist, ist das Timing der Aufnahme problemlos möglich. Ausserdem stellt sich die Emotion bei den Probanden relativ schnell ein und es ergeben sich auch bei der Eindeutigkeit keinerlei Probleme. Für diese Emotion habe ich ein Video mit lustigen Szenen und 2 Filmausschnitte verwendet.

   Für Freude gibt es keine Verbesserungsnotwendigkeit. So haben sich zum einen das Verfahren als auch die verwendeten Videos als optimal für die Erzeugung dieser Emotion herausgestellt.

2. **Überraschung:** Realistische Überraschung ist im Allgemeinen schwer zu erzeugen, da für jede Person eine spezielle Situation, angepasst auf das Charakterprofil des Menschen, entworfen werden müsste. Dies ist aber zu aufwendig und entspricht auch nicht den Anforderungen der Datenbank.

   Mit Hilfe von Videos gibt es zwei Vorgehensweisen: Z.B. ein Video, welches eine überraschende Situation herstellt. Hierbei ergibt sich aber das Problem, dass der Ausschnitt nur bei wenigen Personen den gewünschten Effekt als auch die notwendige Ausgeprägtheit erzeugt, weil jeder Mensch auf andere Situationen überrascht reagiert. Die andere Möglichkeit ist, Schauspieler in Situationen zu zeigen, in denen sie überrascht reagieren. Aus der Situation heraus und durch das Mimikspiel des Schauspielers ist es viel leichter möglich, auch den Probanden zum Zeigen der Emotion zu bewegen. Es lässt sich hierzu nicht vermeiden von der Testperson einen gewissen schauspielerischen Einsatz zu verlangen und an sie zu appelieren, sich in die Situation und den Ausschnitt hineinzuversetzen. Deshalb habe ich letztere Variante in den drei Filmschauschnitten benutzt.

   Die verwendeten Videos wurden nach einer Testphase nocheinmal angepasst. Es wurde zusätzlich versucht, durch Abwechslung in der Situation möglichst alle Bereiche von Überraschung abzudecken. Der Start der Aufzeichnung lässt sich hier relativ gut bestimmen. Die Ergebnisse mit Unterstützung der Probanden haben sich als gut herausgestellt.

3. **Ärger:** Eine Person innerhalb einer kurzen Zeit ohne ein Gespräch wütend zu machen, ist nahezu unmöglich. Somit habe ich mich für die Anregung dieser Emotion für dieselbe Vorgehensweise wie bei Überraschung entschieden. In den Filmausschnitten werden ärgerliche Personen mit wütenden Gesichtern gezeigt. Die Testpersonen sollen sich mit der Situation identifizieren und in einen ähnlichen Zustand versetzt werden.

   Mit schauspielerischer Unterstützung der Testpersonen sind die Ergebnisse bei Ärger trotz der schwierigen Emotion relativ gut. Änderungen waren nur bedingt nötig.

4. **Angst:** Angst exitstiert nach 2.1 in vielen verschiedenen Varianten, allerdings ist authentische Angst bei Menschen nur in Extremsituationen zu beobachten, die für die Aufnahme aber nicht in Frage kommen. Selbst Menschen, die sich bei einem Film fürchten, verbergen diese Furcht innerlich und zeigen keinerlei Mimikspiel. Nur bei spannungsentladenden Momenten schrecken die Zuschauer zusammen, ein kurzer Augenblick in dem sich vielleicht sogar wahre Furcht zeigt.

   Um diesen Moment einzufangen, habe ich etwas längere Filmszenen herausgesucht, in denen sich eine Angstsituation langsam aufbaut und in einem Schreckmoment entlädt. Die Aufnahme startet kurz vor dem entscheidenden Augenblick und kann so ziemlich genau festgelegt werden.

   Die Aufnahmen haben gezeigt, dass die Personen sehr gut auf die gezeigten Videos reagieren, lediglich das letzte verwendete induziert die Angst weniger gut als die ersten beiden. Im letzten wird ein Mensch gezeigt, auf den etwas Unerwartetes zukommt. Diese Szene erfordert eine gewisse Identifikation mit dem Schauspieler, was in dem kurzen Zeitraum nicht optimal funktioniert.

5. **Traurigkeit:** Hierbei handelt es sich um die schwierigste auzulösende Emotion beim Menschen. Schon im Alltag fällt es Erwachsenen schwer, diese Emotion zu zeigen, da sie früh lernen, diese zu unterdrücken. Nur schwerwiegende traurige Ereignisse können die Hemmschwelle brechen. Umso schwerer ist es während eines Versuchs und innerhalb kurzer Zeit diese Emotion auszulösen. Zum einen ist es nötig eine traurige Athmosphäre zu schaffen, zum anderen muss den Versuchspersonen genügend Zeit gegeben werden, sich darauf einzustellen.

   Gelöst wird dies durch längere Szenen von bis zu einer Minute Dauer und durch ein längeres Vorstellen der Emotion. Dies Filmszenen zeigen traurige Situationen von Hilflosigkeit, Abschied und Trennung. Dabei ist es sehr schwer festzustellen, zu welchem Zeitpunkt die Aufnahme starten soll und die Mimik beginnt. Deswegen wird hier für einen längeren Zeitraum aufgezeichnet, um mit Sicherheit eine traurige Mimik einzufangen. Allerdings erschwert dies die Nachbearbeitung und erhöht die Datenmenge[6].

   Als Ergebnis für diese Mimik hat sich gezeigt, dass die Personen zum einen die Emotion nur schwer und wenn, dann nur schwach ausdrücken, zum anderen, dass sich

[6] hierzu wird der meiste Platz von 160MB pro Aufnahme in Anspruch genommen

die Emotion nur sehr langsam einstellt. Im Hinblick auf die dynamische Emotionserkennung ist dies ein großer Nachteil, da die Änderung innerhalb eines bestimmten Zeitraums stattfinden muss. Nur bei wenigen Personen[7] ergaben sich aus genannten Gründen brauchbare Daten für die Mimikerkennung. Zudem sollte bei dieser Emotion über ein anderes Verfahren der Erzeugung nachgedacht werden.

6. **Ekel:** Um diese Mimik bei Personen hervorzurufen gibt es mehrere Möglichkeiten, die auch relativ hohe Erfolgschancen aufweisen. So genügt es schon, Testpersonen Bilder von ekligen Dingen wie verdorbenen Lebensmitteln zu zeigen. Diese Emotion ist demnach leicht auszulösen.

   In meinem Setup habe ich mich allerdings des Mediums Film bedient, das in der Lage ist, noch bessere Reaktionen hervorzurufen. Die Mimik stellt sich in der Regel auch sofort ein.

   Diese Videos sind zum Auslösen dieser Emotion ideal und führen bei fast 100% der Testpersonen zum gewünschten Erfolgt. Auch die Ausgeprägtheit ist sehr gut. Neben Freude ist Ekel die am einfachsten zu erzeugende Mimik.

7. **Neutral:** Der Vollständigkeit halber sei hier noch die Mimik Neutral aufgeführt, die sich leicht aufzeichnen lässt. Hierzu werden den Probanden Bilder, die neutrale Dinge zeigen[8] auf dem Monitor präsentiert und währenddessen für 3 Sekunden die Aufnahme gestartet. Zwei Aufnahmen davon werden ganz zu Beginn gemacht, da die Testperson zu diesem Zeitpunkt noch keine Ahnung hat, was zu tun ist, und dementsprechend entspannte Gesichtszüge aufweist. Die Letzte erfolgt ganz am Ende.

   Die Erfolgschancen bei diesem Vorgehen sind sehr gut und zudem entfällt in den meisten Fällen die Nachbearbeitung, da sich hier ja keine Emotion einstellen muss, sondern ein unveränderter bereits vorhandener Zustand aufgenommen wird.

[7] ungefähr bei 60% der Probanden
[8] hier: ein neutraler Smiley

Kapitel 5

Verfahren zu Mimikerkennung

Die ersten Grundlagen im Bereich dynamischer Emotionserkennung wurden bereits 1991 von Mase entwickelt. Die Mimikerkennungsverfahren beschäftigen sich hauptsächlich mit der Extraktion der Merkmale und der Klassifikation dieser. Bei der Merkmalsextraktion gibt es zwei Vorgehensweisen: Zum einen werden beim geometrisch basierten Ansatz wichtige Merkmalspunkte im Gesicht extrahiert (Augen, Nase, Mund, etc.) und diese anhand deren Positionen und Abstände zu einem Merkmalsvektor zusammengefasst, der die Gesichtsgeometrie widerspiegelt. Zum anderen gibt es den ganzheitlichen Ansatz, wobei Bildfilter (wie beispielsweise Gabor-Wavelets) auf das gesamte Bild angewendet werden und daraus der Merkmalsvektor erstellt wird. Ein weiterer Punkt ist die Vorverarbeitung der Emotionsdaten, die neben der Detektion der Gesichter auch die Erkennung des Auftretens einer Gesichtsbewegung durchführt. Diese Verfahren sind aber für die Untersuchungen dieser Diplomarbeit nicht relevant.

Unterschieden werden muss auch zwischen statischen und dynamischen Verfahren. Ein typischer Vertreter für statische Analyse ist das Facial Action Coding System (FACS) von Ekmann und Friesen. Bei dem in dieser Diplomarbeit verwendeten System handelt es sich um ein dynamisches Verfahren, dass den zeitlichen Aspekt mit berücksichtigt. Des Weiteren ist wichtig, ob das System Kopfbewegungen erlaubt und ob es sich dabei um ein Verfahren handelt, welches die Berechnungen in Echtzeit durchführt. Neuere Forschungsgruppen entwickeln gerade Systeme, die die ganze Bandbreite von der Gesichtsfindung bis zur Erkennung der Emotion beinhalten. Die meisten Verfahren stützen sich wie auch meine Arbeit auf die Einteilung in sechs Basisemotionen. Für den Vergleich der Erkennungsraten der verschiedenen Verfahren ist auch zu beachten, ob die Erkennung personenabhängig oder -unabhängig durchgeführt wurde. Letztere ergibt in den meisten Fällen deutlich schlechtere Ergebnisse.

Hier soll nur ein Überblick über bereits bestehende Verfahren gegeben werden, um das in dieser Diplomarbeit untersuchte System einordnen und mit diesen vergleichen zu können.

5.1 FACS

Ekman und Friesen haben ein System entworfen, um alle visuell unterscheidbaren Gesichtsbewegungen zu beschreiben, welches sich *Factial Action Coding System* (FACS) nennt. Es

basiert auf der Nummerierung aller Bewegungseinheiten (AU's[1]) eines Gesichts, die Bewegung erzeugen. Insgesamt gibt es 46 AU's im FACS (Beispiele siehe Grafik 5.1), die eine Änderung des Ausdrucks beschreiben. Die Kombination dieser AU-Einheiten ergibt eine große Zusammenstellung von möglichen Gesichtsausdrücken. FACS werden weiterhin in der Medizin verwendet, um beispielsweise zwischen simuliertem und realem Schmerz zu unterscheiden oder um herauszufinden, welche Personen die Wahrheit sagen und welche nicht.

Abbildung 5.1: links: AU0 = neutral, rechts: AU4 = brow lowerer

Auch wurden Aspekte der FACS's in Computersystemen, die künstlich erzeugte Emotionen von grafischen Systemen darstellen lassen sowie für Parametrisierung von Gesichtsbewegungen in Muskelmodellen angewendet.

5.1.1 FACS-basierte Erkenner

Einige Ansätze kombinieren die AU's mit anderen Methoden (z.B. optical flow) zur Mimikerkennung oder benutzen Weiterentwicklungen:

So erkennt Tian [lT02] 16 AU's und deren Kombinationen. Die Beschreibung der Gesichtsmerkmale wie Augen, Augenbrauen, Mund und Mundwinkel erfolgt dabei durch „multistate templates", deren Parameter von einem neuronalen Netzwerk Klassifikator zur Erkennung der AU's benutzt werden. Allerdings benötigt das System eine Initialisierung und ist somit nicht automatisiert.

Donato [Don99] kombiniert verschiedene Techniken wie „optical flow", „principle and indipendent component analysis", „local feature analysis" und eine Repräsentation durch Gabor wavelets. Er erkennt acht einzelne AU's und vier AU-Kombinationen, indem er Bildsequenzen verwendet, die manuell ausgerichtet wurden und frei von Kopfbewegung waren. Er erzielte damit Erkennungsraten von 95% mit den Gabor wavelets.

Bartlett [Bar99] erreichte sogar 99% bei der Erkennung von sechs AU's durch Kombination des ganzheitlichen Ansatzes mit der „optical flow" Analyse.

5.1.2 Gabor-Wavelet Analyse

Tian und Kanade [lT02] benutzen die Gabor-Wavelet basierte Methoden zur Erkennung von Bewegung im Gesicht. Dabei erreichen diese Transformationen hohe Sensibilität und Ausprägung der Erkennung von Emotionen und einzelnen Bewegungsfeldern des FACS.

[1] Action Units

Bisherige Arbeiten zeigen, dass die ganzheitlichen Methoden hohe Sensibiliät und Ausgeprägtheit bei emotionsspezifischen Ausdrücken (Freude, Traurigkeit, usw.) und Action Units (AU) unter folgenden vier Randbedingungen aufweisen:

1. Personen sind entweder anglo-europäischen oder asiatischen Ursprungs
2. Kopfbewegung ist ausgeschlossen
3. Bilddaten wurden ausgerichtet, zugeschnitten und auf Standardgröße angepasst
4. spezifische Ausdrucksänderungen oder Action Units sind erkennbar

Nach [lT02] ergaben sich bei der AU Erkennung für homogene Bilder ohne Kopfbewegung sehr gute Erkennungsraten, die sich bei Objekten mit kleinen Kopfbewegungen allerdings rapide verschlechterten. Zusammen mit dem geometrischen Ansatz war die darauf angewandten Gabor Koeffizienten robust gegen Kopfbewegung. Dies lässt die Schlussfolgerung zu, dass der Erfolg der Gabor Koeffizienten im Wesentlichen von der Vorverarbeitung abhängt. Eine Synthese aus Gabor-Wavlet und geometrischen Merkmalen brachte das beste Ergebnis.

5.1.3 Nachteile

Als Nachteil des FACS ergibt sich aufgrund der großen Anzahl von AU's eine hohe Komplexität für deren Kombination zu einer Emotion. Des Weiteren wird die Dynamik der Mimik bei dem reinen FACS nicht berücksichtigt, da nur die statische Information einer ausgedrückten Emotion betrachtet wird. Auch stufen wachsende psychologische Untersuchungen den zeitlichen Verlauf einer Mimik, welcher bei dem FACS komplett fehlt, als sehr wichtig für die Emotionserkennung ein.

5.2 Bayes Klassifikation

Besonderes Interesse der Forschungen in den 90'ern galt auch der Konstruktion automatischer Methoden zur Erkennung von Gesichtsausdrücken aus Videosequenzen. Arbeiten dazu gab es unter anderem von Chen und De Silva. Sebe entwickelte ein Erkennungsmethode mit Hilfe eines naiven Bayes (NB) Modells. Sie zeigten mit ihrem Aufbau, dass eine Cauchy Modellannahme eine bessere Klassifikation als eine Gaussmodell ergibt.

Eine Weiterentwicklung ist der Tree-Augmented-Naive Bayes (TAN) Klassifikator entwickelt von Cohen [Coh02], der Abhängigkeiten zwischen Gesichtsausdrücken erlernt. Zudem liefert er noch einen Algorithmus, der die beste TAN Struktur findet. Die personenabhängigen sowie die -unabhängigen Untersuchungen dieser Forschungsgruppe zeigen die signifikant besseren Ergebnisse der TAN Stuktur gegenüber den einfacheren NB Klassifikatoren. Die Konfusionsmatix für personenunabhängige Versuche ergab eine durchschnittliche Erkennungsrate von 65,5%. Allerdings handelt es sich dabei um eine selbst erstellte Datenbank, bei denen die Personen gespielte Emotionen zeigten.

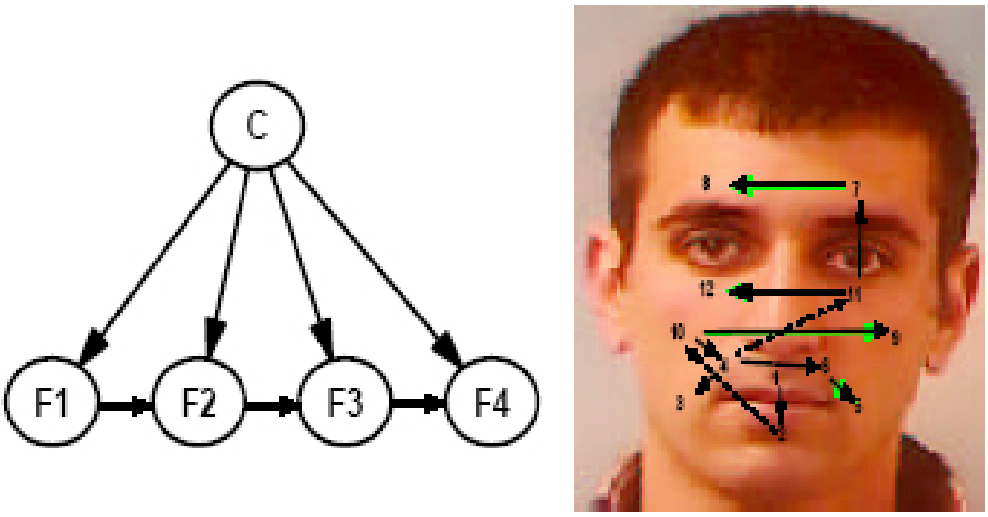


Abbildung 5.2: links: Beispiel eines TAN-Klassifikators, rechts: erlernte TAN Struktur für die Gesichtsmerkmale

5.3 Analyse der „Facial Motion”

Der Großteil der Forschungen im Bereich Bildverarbeitung zur Erkennung von Gesichtsausdrücken hat sich nach [Bar99] auf Bewegungsanalyse durch „optical flow estimation” konzentriert. Wenn die Gewebe- und Muskelstruktur bei verschiedenen Personen ähnlich ist, dann sollten auch die Bewegungen unabhängig von den Oberflächenunterschieden der Gesichter ähnlich sein. Im frühen Stadium der Emotionserkennung hat Mase (1991) den „optical flow” durch Definition eines Fensters für jeden Muskel dazu benutzt, die Aktivität von zwölf der 44 Gesichtsmuskeln zu bestimmen.

Laut [Bar99] entwickelten Yacoob & Davis (1994) eine „mid-level” Repräsentation der Bewegung aus dem „optical flow”, die in Beschreibungen wie „rechter Mundwinkel hebt sich” resultierte. Rosenblum, Yacoob & Davis erweiterten ihre Analyse durch ein komplettes zeitliches Profil der Ausdrücke von Beginn der Emotion über den Höhepunkt zur Entspannungsphase. Dazu trainierten sie radiale Basisfunktionen neuronaler Netzwerke um die Phase der Emotion aus einer Bewegungsbeschreibung zu bestimmen und konstruierten dazu für jede Emotion ein eigenes Netzwerk. Radiale Basisfunktionen nähern nichtlineare Abbildungen durch Gauss-Interpolation von Beispielen an. Auch Beymer, Shashua und Poggio (1993) benutzten die radialen Basisfunktionen.

Ein Ansatz in diesem Bereich stammt noch von Cohn et al. (1997), der ein System gebaut hat, welches Gesichtsbewegungen durch Zuordnung von Merkmalspunkten im Gesicht klassifiziert. Dabei wurden über 40 Punkte im Ausgangsbild manuell lokalisiert und die Verschiebung der Merkmalspunkte mit Hilfe des optischen Flusses bestimmt.

5.4 Modell-basierte Techniken

Nach [Bar99] haben viele Erkennungssysteme detaillierte physische Modelle des Gesichts verwendet (Mase, 1991; Terzopoulos & Waters, 1993; Essa & Pentland, 1997). Essa & Pent-

land erweiterten das anatomisch physische Modell von Tezopoulos & Waters (1993) und wendeten es zu Erkennung und Synthese von Gesichtsausdrücken in zwei verschiedenen Methoden an.

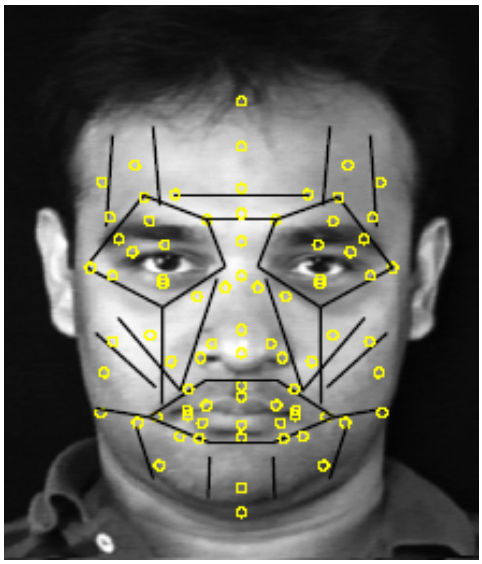
a)  b) 

Abbildung 5.3: a) Wireframe-Modell eines Gesichts, b)Muskelmodell [Ess95]

Das Modell besteht nach [Ess95] und [Bar99] aus einem geometrischem Netz mit 44 Gesichtsmuskeln, dessen Verbindungspunkte und den elastischen Eigenschaften der Haut (siehe Grafik 5.3). Bilder von Gesichtern wurden durch Verformung, die auf sechs Punkten des Gesichts basiert, auf das physische Modell abgebildet. Bewegungsschätzungen von dem optischen Fluss wurden durch das physische Modell in einem rekursiven Schätzungs-Kontroll Netzwerk verfeinert. Die geschätzten Kräfte wurden schließlich benutzt, um die Gesichtsaudrücke zu klassifizieren. Die zweite Methode benutzt nach [Ess95] das physische Modell, um eine „spatio-temporal motion energy template" (siehe Abbildung 5.4) des gesamten Gesichts für jede einzelne Emotion zu generieren. Diese einfachen, biologisch plausiblen Energie „Vorlagen" wurden dann zur Erkennung verwendet.

In einem Modell-basierten System ist die Klassifikationsgenauigkeit allerdings durch die Gültigkeit des Modells beschränkt. Es gibt viele Faktoren, die die Bewegung der Haut, verursacht durch die Muskelkontraktion beeinflussen und somit wäre es schwierig, sie alle exakt in einem deterministischen Modell zu beschreiben.

5.5 Merkmal-basierte Ansätze

Die ersten Annäherungen der Gesichtserkennung basierten laut [Bar99] auf der Messung von Nasenlänge, Kinngestalt und Abstand der Augen (Kanade, 1977). Lantis, Taylor & Cootes (1997) erkannten Identität, Geschlecht und Gesichtsaudrücke indem sie mit einem flexiblen Gesichtsmodell die Formen und räumliche Beziehungen einer Gruppe von Gesichtsmerkmalen maßen.

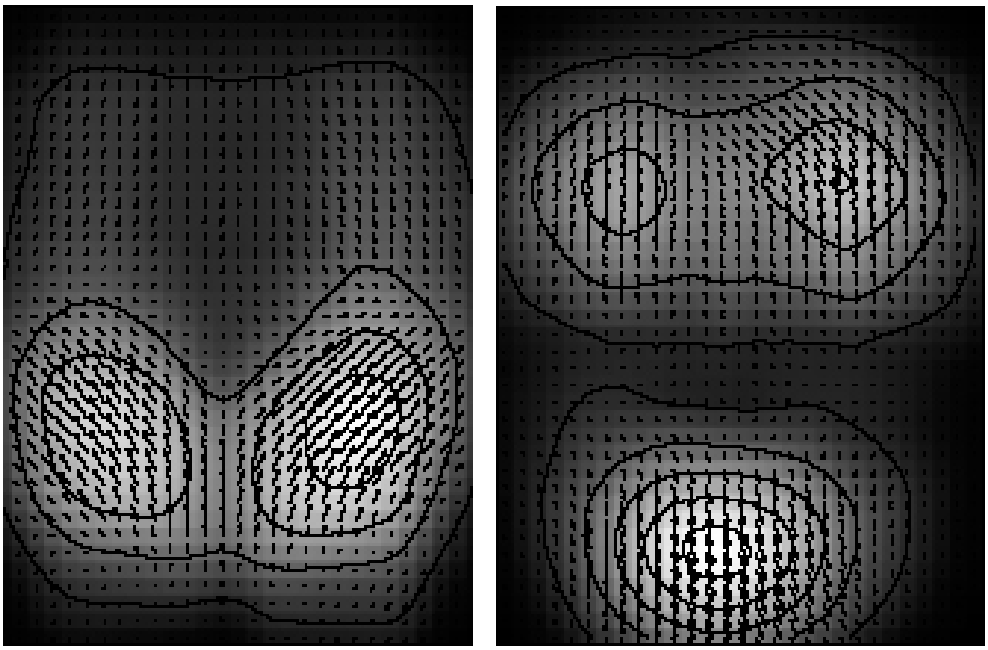


Abbildung 5.4: Motion-energy template für Freude (links) und Überraschung (rechts) [Bar99]

Der Vorteil eines solchen geometrischen Modells ist die drastische Reduzierung der Dimensionen der eingehenden Merkmale. Als Nachteil ist zu nennen, dass die Bildmerkmale die zur Klassifikation benötigt werden, nicht im Voraus bekannt sein dürften. Außerdem könnten dabei entscheidende Informationen durch Komprimierung des Bildes auf eine begrenzte Anzahl von Merkmalen verloren gehen (siehe dazu auch Kapitel 6.2, Seite 51).

5.6 Holistische Systeme

Eine Alternative zu den Merkmal-basierten Ansätzen bietet laut [Bar99] die holistische Analyse, bei der ein hoher Wert darauf gelegt wird, das Originalbild so gut wie möglich zu erhalten und es den Klassifikatoren zu überlassen, die wichtigen Merkmale eines Bildes zu entdecken (Movellan, 1994). Ein Vertreter dieses Ansatzes ist das „template matching”. In ähnlichen Ansätzen mit neuronalen Netzen, müssen die zur Klassifikation nötigen physischen Eigenschaften nicht im Voraus spezifiziert werden, sondern können aus der Statistik der Bildsequenz erlernt werden.

Eine räumliche holistische Darstellung basiert auf den Hauptbestandteilen eines Bildes (Cottrell & Fleming, 1990; Turk & Pentland, 1991). „Pricipal component analysis” (PCA) findet eine orthogonale Menge von Dimensionen, die für die Hauptrichtungen der Variabilität in dem Datensatz stehen. Die Achsen sind sog. „template images”, die „Holons” sowie „Eigenfaces” genannt werden. PCA wurde erfolgreich in der Gesichts- und Ausdruckserkennung angewendet (Bartlett, Viola & Ekmann, 1996; Padgett & Cottrell, 1997).

Ein weitere räumlich holistische Repräsentation wurde durch eine klassenspezifische lineare Projektion der Bildpunkte (Belhumeur, Hespanha & Kriegman, 1997). Genaue Aus-

richtung der Gesichter ist entscheidend für den Erfolg solcher Bild-basierten Annäherungen. „Feature" und „template" basierte Methoden müssen sich nach [Bar99] nicht gegenseitig ausschließen. So haben Lantis, Taylor & Cootes, 1997 beide Methoden verwendet.

5.7 Neuere Forschungen

Neuere Systeme versuchen wie eingangs beschrieben den Prozess der Ausdruckserkennung möglichst zu automatisieren. So erstellten Bartlett, Littlewort, Fasel und Movellan nach [Bar02] ein System, welches automatisch Frontalansichten in Videoaufzeichnungen in Echtzeit detektiert und diese unter Berücksichtigung von sieben Klassen codiert. Der Gesichtsfinder verwendet dabei einen stufenförmigen Merkmalsdetektor, der mit „boosting" Techniken [2] trainiert wurde. Der Erkenner erhält die Bildstellen, die von dem Detektor gefunden wurden. Von diesen Stellen wird eine Gabor Repräsentation erzeugt und dann von einer Reihe von SVM (Support Vector Machines)-Klassifikatoren verarbeitet. Dabei verbessert eine Kombination aus „Adaboost" und SVM'en die Leistung des Systems.

Ein weiterer Ansatz zur Emotionserkennung in Echtzeit liefert Philipp Michel und Rana El Kaliouby [Mic03]. Dabei benutzen sie einen automatischen Merkmalsfinder, der die Gesichter in einem live-Video lokalisiert und die Position von 22 Merkmalspunkten aus dem Gesicht extrahiert. Die Abstände im Video werden dann als Eingang für die Klassifikation mit SVM's verwendet. Die personenunabhängigen Erkennungsraten für dieses System lagen bei 71,8%.

Als letzter Ansatz zur Ergänzung der Systeme zur Emotionserkennung untersuchten Zhang, Lyons, Schuster und Akamatsu [Zha98] zwei Möglichkeiten der Merkmalsextraktion, einmal über geometrische Positionen und das andere Mal über Gabor wavelet Koeffizienten. Diese beiden Ansätze wurden dann gemeinsam oder unabhängig voneinander auf ein multi-layer Perceptron angewendet. Die besseren Ergebnisse zeigten sich wie in anderen Untersuchung bereits festgestellt mit den Gabor wavlets.

5.8 Ausblick

Bei dynamischen Erkennungssystemen für Gesichtsausdrücke wurden in den letzen Jahren erstaunliche Fortschritte gemacht, vor allem im Bereich der automatischen Gesichtsfindung. So wurden auf dem FACS basierende Systeme vollkommen abgelöst. Auch existieren vollautomatische Systeme, die Erkennungen direkt vom Video in Echtzeit durchführen. Allerdings stützen sich diese auf den geometrischen Ansatz der Gefahr läuft, wichtige Merkmale zu unterschlagen. Ebenso unterliegen die Erkennungsraten hohen Schwankungen bei Änderung der Lichtverhältnisse oder bei Kopfbewegung. Auch muss hierbei die Personenunabhängigkeit berücksichtigt werden, da viele gute Ergebnisse nur auf personenabhängig trainierte Systeme zutreffen.

[2] „Adaboost"- und „Gentleboost" Algorithmus

In dieser Diplomarbeit wird im Gegensatz dazu der holistische Ansatz verfolgt, der alle entscheidenden Merkmale im Bild mit in das trainierte System aufnimmt. Auch berücksichtigt die Modellierung mit den im folgenden Kapitel vorgestellten P3DHMM'en neben der statistischen Information der Merkmale die zeitliche Dynamik der Emotion. Deshalb birgt das nun vorgestellte Erkennungssystem ein großes Potential für zukünftige Mimikerkennung.

Kapitel 6

Mimikerkennung

Im letzten Teil der Arbeit sollten mit Hilfe der neu erstellten Datenbank zwei Verfahren zur Mimikerkennung getestet und verglichen werden. Dazu war es notwendig, die bisher erstellten Verfahren zu rekonstruieren und auf die neuen Anforderungen anzupassen. Zuvor werden noch die benötigten mathematischen Modelle vorgestellt und die grundsätzliche Funktionsweise der Erkennung mit diesen erläutert. Abschließend erfolgt in Kapitel 7 die Präsentation der Ergebnisse, der Vergleich dieser mit den bisherigen Experimenten und einem einfacheren Verfahren der Merkmalsextraktion.

6.1 Hidden Markov Modelle

Die Theorie der Hidden Markov Modelle (HMM) wurde bereits Ende der sechziger Jahre von Baum entwickelt. So sind diese nach [Rab89] und [Oer96] „stochastische Modelle, die zur Beschreibung und Analyse realer Signalquellen verwendet werden". Für die Signalmodelle bieten sich insbesondere Zeitreihen an, also Messergebnisse, die als Funktion der Zeit aufgezeichnet wurden. Die Modellbildung ermöglicht das „Verstehen" der Signalquelle und laut [Oer96] „die Simulation und die Vorhersage zukünftiger Daten". Durch Auswählen desjenigen Signalmodells, welches die Messergebnisse am besten beschreibt, ist eine Klassifizierung und Identifizierung möglich. Daher finden die Modelle besonders in der Spracherkennung Anwendung.

Hidden Markov Modelle sind Zustandsautomaten vom Moore Typ, bei dem ein Zustand mit einer gewissen Wahrscheinlichkeit in einen anderen Zustand übergeht. Ein Markov Modell besitzt nach [Rus03] also mehrere Zustände, die durchlaufen werden können und durch Übergangswahrscheinlichkeiten bewertet werden. „Zusätzlich besitzt jeder Zustand eine Wahrscheinlichkeitsdichtefunktion (WDF), die die Wahrscheinlichkeit für das Erzeugen („Emittiern") eines Mustervektors angibt". Die Markov Bedingung besagt, dass der Automat nur ein Gedächtnis von 1 hat, also die Übergangswahrscheinlichkeit nur vom aktuellen Zustand und nicht von der Vergangenheit abhängt. Der zugrunde liegende stochastische Prozess ist dabei verdeckt (Hidden).

6.1.1 Definition

Dem Markov Modell liegen eine endliche Menge (Zustandsmenge)

$$Q = \{s_1, s_2, s_3, ..., s_N\}$$

und eine Folge von Zufallsvariablen

$$q = \{q_1, q_2, ..., q_T\} : \text{Folge von Zufallsvariablen}$$

zugrunde, die Werte aus dieser Menge Q annehmen. Die Übergangswahrscheinlichkeiten für einen einfachen, stationären und kausalen Prozess haben die Form

$$P(q_t|q_1...q_{t-1}) = P(q_t|q_{t-1})$$

Dies gilt, da nur der unmittelbar vorhergehende Zustand einen Einfluss hat (Makov Bedingung). Die Parametermatrix ergibt sich dann zu

$$A = [a_{ij}]_{N\times N} \text{ mit } a_{ij} = P(q_t = s_j|q_{t-1} = s_j)$$

Für die Einträge der Matrix gilt die Stochastizitätsbedingung $a_{ij} \geq 0$ und $\sum_j a_{ij} = 1$. Die Startwahrscheinlichkeiten

$$\pi_i = P(q_1 = s_i),\ \sum_{i=1}^{N} \pi_i = 1$$

für die Einnahme eines Anfangszustandes werden in dem N-dimensionalen Vektor π vereinigt (siehe auch [Sam94]). Solche Prozesse heissen *Markov Ketten* und sind vollständig durch die Parameter π und **A** definiert.

Des weiteren wird nun ein zweiter Prozess angenommen, der abhängig vom eingenommenen Zustand ein Zeichen aus einem endlichen Ausgabealphabet

$$\kappa = v_1, ..., v_k$$

erzeugt. Ein Beobachter dieses Vorgangs sieht allerdings nur die diskrete Symbolfolge

$$O = O_1...O_T.$$

Die Folge der inneren Zustände bleibt jedoch verborgen. Für die Zeichenproduktion gilt

$$P(O_t|O_1...O_{t-1}, q_1...q_t) = P(O_t|q_t)$$

und hängt dabei nach [St98] nur vom aktuellen Zustand ab. Die diskrete Ausgabeverteilung lässt sich dann nach

$$b_{jk} = b_j(v_k) = P(O_t = v_k|q_t = s_j) \text{ und } B = [b_{jk}]_{N\times K}$$

bestimmen. Der gesamte stochastische Prozess heißt Hidden Markov Modell und nach 6.1 vollständig spezifiziert:

$$\lambda = (A, B, \pi) \tag{6.1}$$

Besonders in der Spracherkennung können die Modelle durch bestimmte Annahmen vereinfacht werden. So werden Modelltopologien mit ausgezeichneten Anfangs- und Endzuständen sowie einer zeitlichen Ordnung zur Wortmodellierung verwendet. Typische Vertreter sind *links-rechts-Modelle,* bei denen nur Zustände von links nach rechts durchlaufen werden können. Übergangswahrscheinlichkeiten a_{ij} für $j < i$ verschwinden. Zudem wird nach [St98] mit $\pi = (1, 0, ..., 0)^T$ sichergestellt, dass das Modell zwingend im Zustand s_1 startet, während das Ende nur im letzten Zustand erreicht werden kann. Bei *linearen* Modellen wird laut [St98] zusätzlich noch das Überspringen des $s_i \rightarrow s_{i+2}$ des Nachfolgezustandes s_{i+1} verboten.

Neben der geeigneten Wahl der Topologie und der Zustände des Hidden Markov Modells ist dieses noch der Beobachtungssequenz anzupassen, was dem Vorgang des Trainings bei Neuronalen Netzen entspricht. Nach [Oer96] und [Rab89] stellen sich damit folgende Probleme:

1. Wie wahrscheinlich ist eine Beobachtungssequenz bei gegebenen Hidden Markov Modell?
   $P(O|\lambda) = ?$

2. Wie bestimmt man bei geeigneter Beobachtungssequenz und Modell die a posteriori Wahrscheinlichkeiten einzelner Zustände und welche Optimalitäskriterien gibt es?

3. Wie optimiert man die Parameter (Übergangs- und Ausgangswahrscheinlichkeiten) eines HMM's?

6.1.2 Bestimmung der Beobachtungssequenz

Zum ersten Problem lässt sich die Wahrscheinlichkeit, dass eine konkrete Zustandsfolge durchlaufen wurde als Produkt

$$P(q|\lambda) = P(q_1...q_T|\lambda) = \pi_{q_1} \cdot \prod_{t=2}^{T} a_{q_{t-1}q_t} \tag{6.2}$$

von Anfangs- und Übergangswahrscheinlichkeiten ausdrücken. Da die Zeichenproduktion nur vom aktuellen Zustand abhängt lautet die Produktionswahrscheinlichkeit

$$P(O|q, \lambda) = P[O_1...O_T|q_1...q_T, \lambda) = \prod_{t=1}^{T} b_{q_t}(0_t). \tag{6.3}$$

Die Verbundwahrscheinlichkeit für ein gemeinsames Eintreten beider Folgen **q** und **O** ist demnach

$$P(O, q|\lambda) = P(O|q, \lambda) \cdot P(q|\lambda) = \pi_{q_1} b_{q_1}(O_1) \cdot \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t}(O_t) \tag{6.4}$$

Die gesuchte *Produktionswahrscheinlichkeit* ergibt sich dann nach [St98] zu

$$P(O|\lambda) = \sum_{q \in Q^T} P(O, q|\lambda) = \sum_{q \in Q^T} \pi_{q_1} b_{q_1}(O_1) \cdot \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(O_t) \tag{6.5}$$

Allerdings liegt der Berechnungsaufwand nach Gleichung 6.5 bei etwa $2T \cdot N^T$, da es für N Zustände N^T mögliche Zustandsabfolgen gibt und für jede dieser Zustandsabfolgen ungefähr $2T$ Berechnungen nötig sind. Beispielsweise sind für $N = 5$ (Zustände) und $T = 100$ (Beobachtungen) $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$ Kalkulationen durchzuführen, was selbst bei kleinen Werten für N und T ein zu hoher Aufwand ist.

Es existiert glücklicherweise eine effizientere Methode, um die Berechnung der gewünschten Wahrscheinlichkeiten zu vereinfachen. Diese Vorgehensweise heißt *Vorwärts-Rückwärts*-Verfahren. Dabei sind die *Vorwärtswahrscheinlichkeiten* durch

$$\alpha_t(j) = P(O_1 ... O_t, q_t = j|\lambda) \tag{6.6}$$

oder die *Rückwärtswahrscheinlichkeiten* durch

$$\beta_t(i) = P(O_{t+1} ... O_T|q_t = i, \lambda) \tag{6.7}$$

definiert. Nach [Sam94] bilden die $\alpha_t(j)$ als auch die $\beta_t(i)$ eine $T \times N$-Matrix, die „mit Hilfe von $2N^2T$ Multiplikationen gefüllt wird". Zur Bestimmung kann man eine der beiden Werte verwenden. Für die genaue Berechnungsvorschrift des Algorithmus siehe [St98](S.131).

6.1.3 Erkennung

Problem zwei beschäftigt sich damit, die Folge der verdeckten Zustände eines HMM's aufzudecken, also die richtige Zustandsabfolge zu finden. Die Kenntnis der Ausgabe O und der Modellparameter λ lassen einige statistische Rückschlüsse zu. Das Optimalitätskriterium als Teil dieses Problems beschäftigt sich damit, den wahrscheinlichsten Gesamtpfad auszuwählen, also die optimale Zustandsfolge q^* aus allen möglichen $q \in Q^T$ der Länge T zu finden. Gesucht wird bei gegebenem Modell λ und Beobachtungssequenz O eine optimale Zustandsfolge $q^* = q_1^* ... q_T^*$, so dass

$$P(O, q^*|\lambda) = \max_{q \in Q^T} =: P^*(O|\lambda); \tag{6.8}$$

Realisiert wird dieses Optimierungsproblem mit Hilfe des *Viterby-Algorithmus*, der nach [St98] „eine Variante des Verfahrens zur Berechnung der Vorwärtsmatrix" ist. Für den Ablauf dieses Verfahrens siehe [St98](S.133). Der *Viterby-Algorithmus* wird für die meisten Erkennungsaufgaben verwendet, so auch hier (siehe Abschnitt 6.1.3).

6.1.4 Training

Das dritte Problem der Optimierung entspricht dem Vorgang des *Trainings* eines Hidden Markov Modells. Dabei geschieht eine Anpassung der Modellparameter an eine gegebene Beobachtungssequenz: Optimiere $\lambda = (A, B, \pi)$ um $P[O|\lambda]$ zu maximieren. Dies entspricht

einer *Maximum-Likelihood*-Schätzung der Modellparameter des Modells λ unter Zuhilfenahme einer Lernstichprobe. Dazu gibt es mehrere Optimierungsverfahren, doch findet hier der *Baum-Welch-Algorithmus* Anwendung.

Die Baum-Welch Methode überführt ein Modell $\lambda = (A, B, \pi)$ in ein verbessertes Modell $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ mit adaptierten Parametern. Die genaue Vorgehensweise ist beschrieben in [St98](S.136/137) bzw. in [Rab89](S.264-266) und soll hier nicht weiter erläutert werden.

Nach diesem kurzen Einblick in die Hidden Markov Modelle ist noch anzumerken, dass ein wesentlicher Teil der Konstruktion eines HMM's in der Auswahl eines geeigneten Modells liegt, wozu die Bestimmung der Zustände, die Auswahl der Topologie und die Belegung der initialen Parameter zählt. Dazu lässt sich nach [Rab89] kein allgemeines Verfahren anwenden:

> „Es existiert kein einfaches, theoretisch korrektes Verfahren"[Rabiner]

Somit sind Erfahrungswerte bei der Wahl eines Modells sehr hilfreich.

6.2 Erkennungssystem

Ein automatisches Mimikerkennungssystem kann nach [Mül02] im allgemeinen in drei Abschnitte eingeteilt werden:

1. Gesichtsdetektion in einer Szene
2. Merkmalsextraktion
3. Statistische Klassifikation der mimischen Merkmale

Zum ersten Teil gibt es in der Wissenschaft schon viele erfolgreiche Ansätze zur Lösung des Problems mit Hilfe neuronaler Netze. Diese Technik ist bereits ausgereift und soll in dieser Arbeit nicht weiter betrachtet werden. Auch gibt es Möglichkeiten mit Hilfe des Condensation Algorithmus sowie aktuellen Weiterentwicklungen (siehe [Ple04]) Gesichter in Filmmaterial zu detektieren, zu tracken und schließlich zu extrahieren.

Die Extraktion der Merkmale versucht die für den Gesichtsausdruck wichtigen Eigenschaften herauszuholen und von den unwichtigen Informationen zu trennen. Im Idealfall entsteht eine Repräsentation aller für die Mimik entscheidenden Elemente. Es gibt zwei verschiedene Ansätze, um dies zu erreichen. Der erste versucht mit Hilfe eines geometrisch basierten Ansatzes (Geometric-Feature-Based Systems) wichtige Merkmalspunkte im Gesicht zu finden, deren Abstände und Bewegungsrichtungen in einem Merkmalsvektor gespeichert werden. So entsteht ein vektorbasiertes Abbild des Gesichts. Diese gewonnen Merkmale sind sehr robust, benötigen allerdings sehr viel Zeit zu deren Erstellung und Berechnung. Der zweite Ansatz behandelt das Bild als Ganzes (Holistic Systems), um mit Hilfe von Signalverarbeitungsmethoden, wie Gabor- oder diskreter Cosinus Transformation, die entscheidenden Merkmale zu extrahieren. Letzteres Verfahren findet in dieser Arbeit Anwendung. Der große Vorteil dieser Vorgehensweise ist, dass das Problem der Detektion der für die Mimik entscheidenden Punkte im Gesicht wegfällt, indem das ganze Bild durch wichtige Merkmale

beschrieben wird. So entgeht man dem Fehler, wichtige Merkmalspunkte zu vergessen. Allerdings besteht die Gefahr, dass das Modell zu einem Gesichtserkennungs- anstatt eines Mimikerkennungssystems degeneriert. Der Gesichtsaudruck kann nur korrekt erkannt werden, wenn die Änderung mit einbezogen wird. Der entscheidende Vorteil ist, dass mit Hilfe dieses Verfahrens sowohl die statischen Merkmale eines Bildes als auch die Änderung über die Zeit zur Berechnung des Modells verwendet werden.

Die abschließende Klassifikation weist den Merkmalen Wahrscheinlichkeitsverteilungen zu und berechnet ein Modell, welches die Mimik repräsentiert. Mit diesem kann dann die Erkennung durchgeführt werden. In vielen Erkennungssystemen werden neuronale Netze verwendet, in dieser Arbeit geschieht die Klassifikation mit Hilfe von Pseudo-3D Hidden Markov Modellen (P3DHMM).

6.2.1 Pseudo 3D Hidden Markov Modelle (P3DHMM)

Die Grundlagen eindimensionaler HMM'e wurden bereits eingangs dieses Kapitels erklärt. Für Gesichtserkennung haben schon Pseudo-2D Hidden Markov Modelle (P2DHMM) sehr gute Ergebnisse geliefert (siehe [Mül02]). Das P2DHMM ist eine Erweiterung des eindimensionalen HMM's, welches entwickelt wurde, um zweidimensionale Daten zu beschreiben. Der Begriff Pseudo ergibt sich aus der Tatsache, dass die Zustandsbestimmung angrenzender Spalten unabhängig voneinander vorgenommen wird. Die Startzustände in horizontaler Richtung werden als Superstates bezeichnet.

Da es sich bei der Erkennung an Stelle von Einzelbildern um Emotionssequenzen handelt, also eine Dimension mehr hinzukommt, hat sich der Einsatz von nach [FH01]P3DHHM'en als adäquat erwiesen. Die dreidimensionale Struktur entsteht durch nochmaliges Anwenden der obigen Erweiterung auf die P2DHMM's, nur werden neben dem Einfügen der Startzustände für die Spalten noch Startzustände für die Bilder eingeführt. Diese Hyperstates der P3DHMM'e können als Anfangszustände der Bilder interpretiert werden.

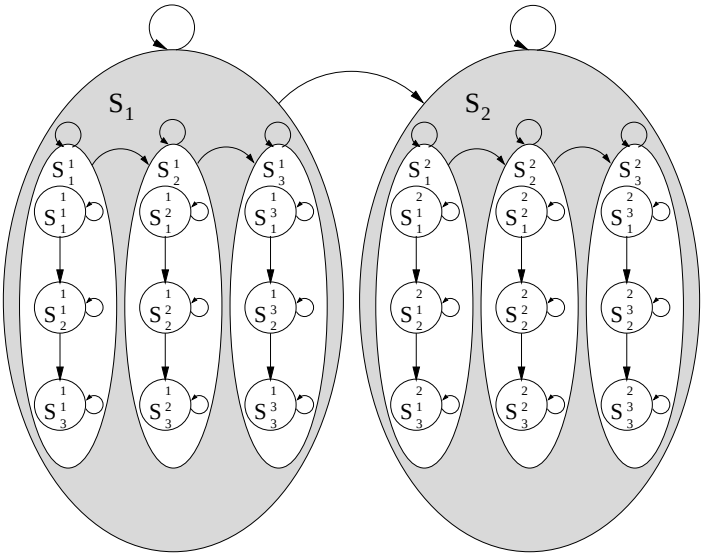


Abbildung 6.1: P3DHMM

Die Grafik 6.1 veranschaulicht ein P3DHMM mit $2x3x3$ Zuständen. Es besteht aus zwei sog. *Hyperstates* (S_1, S_2), die je ein P2DHMM einschließen. Jedes P2DHMM enthält wiederum drei *Superstates* (z.B. S_1^1), die je drei Zustände besitzen.

Nach [Sam94] ist es möglich, durch Einfügen spezieller Startzustände und -merkmale, das dreidimensionale Modell in ein lineares eindimensionales HMM umzuwandeln. Durch Rückführung der 3D- auf eine äquivalente 1D-Struktur wird die Anwendung des Standard Baum-Welch-Verfahrens und des Viterby-Algorithmus zur Erkennung möglich. Diese Rückführung lässt sich nach Rabiner (siehe [Rab89]) einfach durchführen. Das äquivalente eindimensionale HMM ist in Abbildung 6.2 dargestellt.

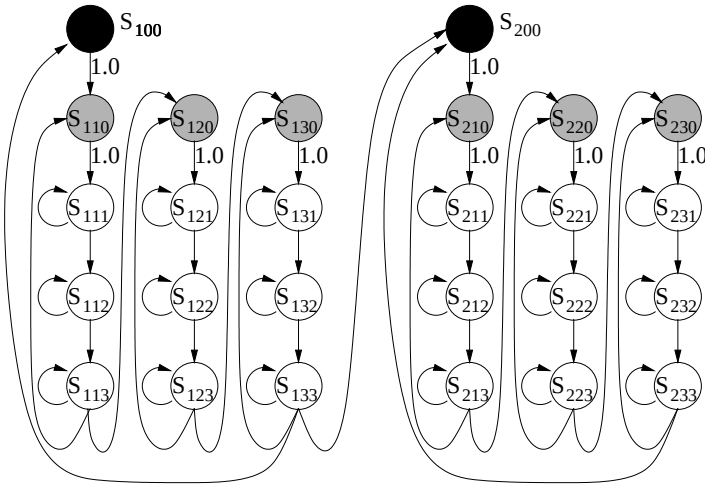


Abbildung 6.2: Äquivalentes 1D-HMM mit Startzuständen

6.2.2 Äquivalente 1D-Modelle

Zum Training werden nun die zu den P3DHMM'en äquivalenten 1D-HHM'e benötigt. Dies geschieht durch Einfügen spezieller Start-der-Bilder und Start-der-Spalten Zustände. In Grafik 6.3 sind die Startbedingungen der Spalten grau und die Bilder schwarz hinterlegt. Im äquivalenten 1D-Modell kann von dem Anfangszustand des Bildes (schwarz) also nur in das folgende Bild übergegangen werden. Dort muss wiederum vom Startzustand der Spalte in die Zustände dieser Spalte gesprungen werden. Erst wenn diese durchlaufen wurden, erfolgt der Übergang in die nächste Spalte bzw. das nächste Bild.

Zum Training und zur Erkennung werden die in der Sprachverarbeitung häufig angewandten HTK-Tools benutzt. Mit diesen können die Modelle erstellt, der Viterby- und der Vorwärts-Rückwärts Algorithmus sowie sonstige Standard-Operationen in der Sprachanalyse durchgeführt werden. Um mit diesen arbeiten zu können, müssen die Merkmale in eine HTK-übliche Form gebracht werden (siehe dazu Abschnitt 6.4.2.3, Seite 60 und Abschnitt 6.3.2, Seite 56). Zur Markierung der Startzustände in dieser Datei müssen diese zur Unterscheidung komplett andere Werte annehmen, um sie von den anderen Merkmalen unterscheiden zu können. Dabei wird den Anfangszuständen der Spalten ein viel höherer Wert zugewiesen als die Wahrscheinlichkeiten der Merkmalseinträge. Gleiches gilt für die Startzustände

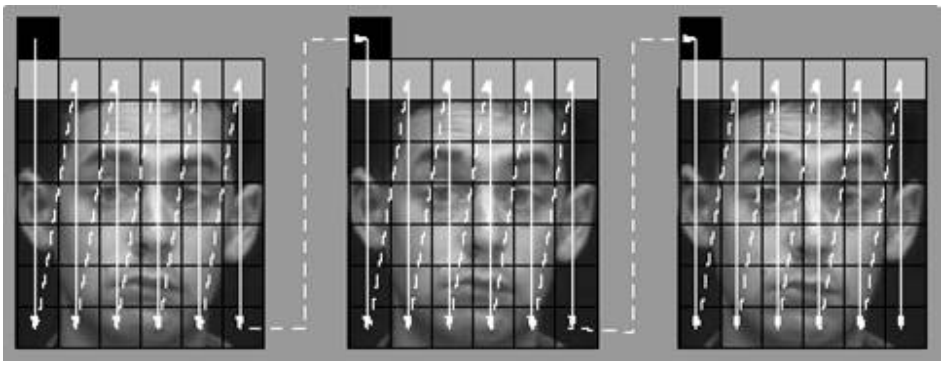


Abbildung 6.3: Bildsequenz mit Startzuständen für die Spalten (grau) und der Bilder (schwarz) [FH01]

der Bilder, bei denen dieser hohe Wert nur doppelt (über zwei Zeilen) auftritt. Dadurch ist die Merkmalsdatei nun vollständig spezifiziert.

6.2.3 Ablauf des Trainings

Das Training der äquivalenten 1D HMM'e ist aufgrund des *Vorwärts-Rückwärts-Algorithmus* ein zeitraubender Prozess[1]. Da jedes P3DHMM mehrere P2DHMM'e enthält, kann das Training in zwei Schritte aufgeteilt werden:

Im ersten Schritt werden alle Bilder, die eine Emotionssequenz beschreiben, in vier Abschnitte unterteilt, deren Länge möglichst gleichverteilt ist. Da das Modell aus vier Superstates[2] besteht, wird somit jedes Viertel der 15 Bilder einem P2DHMM zugewiesen. Mit diesen werden dann vier 1DHMM'e trainiert, die äquivalent zu den P2DHMM'en sind. Die trainierten P2DHMM's werden im Anschluß zu einem P3DHMM zusammengeklebt, indem Startzustände für die Bilder eingefügt werden. Insbesondere müssen die Übergänge der Endzustände auf den Beginn des nächsten Startzustandes und damit auf das nächste P2DHMM umgeleitet werden.

Im zweiten Schritt wird nun das gesamte P3DHMM[3] mit der kompletten Bildersequenz trainiert. Die Zeitersparnis liegt bei dieser Methode gegenüber dem normalen Training bei 50%.

6.3 Rekonstruktion des Systems

Für die Trainingsdaten des alten Systems wurde nach [FH01] eine Videosequenz mit 25 frames/sec und einer Auflösung von 320x240 Pixel verwendet. Die Kamera war wie in meinen Aufnahmen auf den Kopf in Frontalansicht ausgerichtet. Die Originalbilder wurden auf den Kopf zugeschnitten, in der Größe angepasst[4] und auf Punkte definiert durch Augen, Nase

[1] proportional zu $T \cdot N^2$

[2] jeder Superstate gehört zu einem P2DHMM

[3] beim Training handelt es sich wieder um ein 1DHMM welches dem äquivalenten P3DHMM entspricht

[4] eine Auflösung von 196x172 Pixel

und Mund ausgerichtet. Die verwendete Datenbasis besteht aus insgesamt 96 Aufnahmen von 6 Personen mit 15 Bildern je Aufnahme. Die Aufnahmen beginnen auch mit dem neutralen Ausdruck und ändern sich zu der jeweiligen Emotion. In diesem Fall wurden nur vier Emotionen berücksichtigt: Ärger, Überraschung, Ekel und Freude.

6.3.1 Reaktivierung der bestehenden Systems

Als erstes war es nötig, die bereits existierenden unkommentierten Skripten zur Mimikerkennung durchzugehen und zu reaktivieren. Dazu kopierte ich alle wichtigen Dateien nach und nach in das separates Verzeichnis $HOME[5]/MIMIK, um die für mich wichtigen von den unwichtigen Programmen zu trennen. Auch für die spätere Nachvollziehbarkeit und Übersichtlichkeit ist dies von großem Vorteil.

Im nächsten Schritt mussten die Pfade der Skripte, die entweder in Perl, Bash oder C++ programmiert waren, auf die neuen Verzeichnisse angepasst werden. Außerdem ist es wichtig, alle erforderlichen Tools, wie beispielsweise HTK, in die Systemumgebung durch Angabe der Pfade oder durch Eintrag in die $PATH[6] Variable einzubinden und alle von den einzelnen Skripten aufgerufene Programme bereitzustellen. Zum Abschluss wurde dann der gesamte Trainingsprozess der HMM's mit den angepassten Programmen durchgeführt und die damals erreichte Erkennungsrate überprüft als auch verifiziert. So konnte sicher gestellt werden, dass das System wieder einsatzbereit und funktionsfähig ist und ich, darauf aufbauend, die Modelle mit meiner neu erstellten Datenbank trainieren konnte.

6.3.2 Rekonstruktion des Erkennungssystems

Für die Reaktivierung habe ich die bereits zugeschnittenen und in pgm-Format umgewandelten Bilder benutzt, da zunächst die Funktionsweise der Trainings- und Testalgorithmen überprüft werden sollte.

Bei den alten Daten existieren zwei Verzeichnisse, die Bilddaten enthalten. Einmal in „$HOME/pgm", welches pro Person und Emotion die ausgesuchten 16 Bilder, die die jeweilige Mimikänderung zeigen, beinhaltet. Diese werden dann in *p_000.pgm* bis *p_015.pgm* umbenannt, und in ein Verzeichnis „$HOME/feat" mit der Emotion und dann der jeweiligen Person als Unterverzeichnis kopiert. In diesem Verzeichnis befinden sich somit alle zugeschnittenen relevanten Bilder. Ebenso waren dort bereits die zugehörigen Differenzbilder vorhanden. Da bis hier alle Vorarbeiten nachvollziehbar waren, habe ich mit der Nachstellung des Trainings bei der Merkmalsextraktion ausgehend von diesen Differenzbildern begonnen.

Im Verzeichnis „$HOME/scripts" liegen alle jemals erstellten Programme, die für die einzelnen Schritte nötig sind; unter anderem auch das Programm *complete_run.pl*, mit dessen Hilfe ich den Ablauf rekonstruieren konnte.

[5] $HOME = Varibale, die den Heimatpfad angibt; hier: $HOME = /home/elmo0/home/haw

[6] $PATH = Systemvariable unter Linux; enthält alle Verzeichnisse, in denen ausführbare Programme gesucht werden

Als erstes wird das Skript *mkfeat.pl* aufgerufen, welches mit Hilfe einer Liste namens *traindir*, die alle Pfade der zu trainierenden Bilder enthält, das C++ Programm *dct_trans* startet. Dieses erzeugt mit Diskreter Cosinus Transformation die DCT-Koeffizienten der Differenzbilder, die in ASCII-Files gespeichert werden. Diese müssen dann noch in eine HTK-übliche Form gebracht werden, was das Skript *featdiff.pl* ebenso durch Aufruf von *dct_trans* erledigt, nur mit den Parametern *m=2* und *t=2*, die mit den vorher erstellten ein zweidimensionales P2D-File in HTK-Form erzeugen. Diese P2D Modelle der Bilder liegen nun als **.f0* Dateien in den jeweiligen Verzeichnissen. Zum Training werden diese erzeugten Merkmalsdateien mit dem Programm *cpfeatdiff.pl* in ein eigenes Trainingsverzeichnis „/train" kopiert, die dort nun in der Form

[Emotion][Person]_]Versuch]_[001015].f0[7]

vorliegen.

Im nächsten Schritt generiert das Hauptprogramm mit Hilfe von *Proto2D* einen leeren, allgemeinen 2D-Prototypen *proto2D*. Dieser wird dann mit *HRest* und den Merkmalen aller Bilder vortrainiert. Es entsteht die ein Mittelwert-Modell aller Bilder namens *proto_2D*.

Das Skript *train2D* nimmt im Folgenden eine Aufteilung der 15 Bilder in 4 Abschnitte (Bilder 1-4, 5-8, usw.) vor und trainiert diese dann im einzelnen. Es entstehen also 4 nacheinander trainierte äquivalente 2D-Modelle. Der Vorteil dieses Vorgehens ist, dass das Training wesentlich schneller abläuft. Die vier einzelnen 2D-HTK-Files werden im Anschluß mit dem Programm *changeproto.pl* zu einem 3D-Modell zusammengeklebt.

Zum Abschluss erfolgt das Training der dreidimensionalen HTK-Prototypen mit den zusammengefassten 3D-Merkmalsdateien, die im Verzeichnis „/train/all" abgespeichert werden und die Form [Emotion][Person].f0 haben. Diese wurden bereits vorher von dem Skript *mk3dHtk.pl* erstellt. Dazu wird das Programm *train3Dall* aufgerufen, welches die vortrainierten HTK-Modelle auf die einzelnen Emotionen nachtrainiert. Der Vorgang dauert, verteilt auf mehrere Rechner ungefähr einen Tag. Als Ausgabe entsteht für jede Emotion (hier vier) ein Modell, abgelegt in „/scripts/hmm0".

Als Abschluss des Trainings wird mit Hilfe des Emotionsverzeichnisses (*model.list)* und der Datei *hed*.script noch ein sog. „Master-Model-File" (MMF) erstellt. Dieses enthält alle wichtigen Modelle der einzelnen Emotionen und wird im Folgenden für die Erkennung verwendet.

Zur Erkennung müssen zuerst die dreidimensionalen Merkmale der Testbilder erzeugt werden, welches das Programm *mk3Dhtk.pl* übernimmt. Diese Dateien werden dann in ein das „/test" Verzeichnis kopiert. Das Hauptprogramm startet nun mit dem HTK-Tool *HVite*, welchem der Viterby-Algorithmus zur Erkennung zugrunde liegt. Es entstehen **.rec* Dateien, die die erkannten Modelle enthalten. Ein Skript *top3.pl* zählt die erkannten Emotionen und gibt die Erkennungsraten aus.

Da die Merkmale neu generiert wurden und als Ergebnis dieser Erkennung im Vergleich mit den alten Daten die gleichen Ergebnisse erzielt wurden, kann das System nun als funktionsfähig und reaktiviert angenommen werden. Im Folgenden wird nun auf die Erkennung mit den neuen Daten sowie auf Veränderungen eingegangen.

[7] Beispiel: *anger0002_1_004.f0*

6.4 Erstellung der P3D-Modelle

Das Training der Modelle, inklusive der Vorverarbeitung und Merkmalsextraktion, kann mit Hilfe des Hauptscripts *run_all.pl* in $HOME/scripts gestartet werden. Für detaillierte Informationen zu den Scripten siehe Anhang A.2.

6.4.1 Vorverarbeitung

Die neuen Mimikdaten liegen in der in Kapitel 3 beschriebenen Form in den jeweiligen Verzeichnissen mit bis zu 300 Bildern je Mimik und Person vor. Allerdings zeigt nur ein Bruchteil der Bilder die für das Training benötigte Änderung der Emotion. Aus vergangenen Beobachtungen hat sich gezeigt, dass eine solche Änderung innerhalb einer halben Sekunde auftritt und sich ein Wert von 16 Bildern, die schließlich 15 Differenzbilder ergeben, als optimal herausgestellt hat. Die erste Aufgabe, die sich nun vor dem Beginn des eigentlichen Trainings stellt, ist die Vorverarbeitung der Daten.

6.4.1.1 Erstellung der Bilderlisten

Als erstes müssen aus den Bildern in der Datenbank diejenigen 16 herausgesucht werden, bei denen die Änderung zur jeweiligen Emotion auftritt. Dazu werden alle Bildersequenzen durchgesehen und die Startbilder, bei denen die Emotion beginnt, herausgesucht. Momentan gibt es noch keine einfache Möglichkeit, dies mit Hilfe eines Programms zu verwirklichen. Zwar existieren schon Ansätze, diese waren zum Zeitpunkt dieser Arbeit allerdings noch im Entwicklungsstadium.

Deswegen werden die Bilder der Reihe nach durchgesehen und das Startbild mit Angabe des Pfades in eine Datei namens *traindir*[8] geschrieben. Eine Zeile dieser Datei beinhaltet folgende Informationen:

anger/0002_1	034	+
[Pfad]	[Bildnummer]	[Bewertung]

dabei ist

[Pfad]=[Mimik]/[Person]_[Aufnahme]

[Bildnummer] = Nummer des Startbildes p_[Bildnummer].pgm

[Bewertung] = $\{*, +, -, \sim\}$[9]

Anhand des ersten Eintrags der Datei wird der Pfad spezifiziert, auf den sich die folgende Bildnummer bezieht. Diese kennzeichnet den Beginn der 16 Bilder der Emotionssequenz. Auf beide Einträge[10] greifen später die für das Training verwendete Scripten zu. Der Eintrag in der dritten Spalte bewertet die Emotion, um eventuell schlechtere Mimiken für das

[8] der Pfad der Datei befindet sich in $HOME/list

[9] „$*$" = normal, „+" = optimal, „$-$" = nicht brauchbar, „$\sim$" = fragwürdig

[10] besonders auf ersteren

Training auszuschließen oder um beispielsweise nur die ausdrucksstärksten (mit „+” gekennzeichneten) zu selektieren. Der Eintrag muss aber nicht vorhanden sein. In der Datei *picdir* im Verzeichnis „$HOME/list” sind alle brauchbaren aufgenommenen Startbilder der Datenbank eingetragen. Aus den Einträgen dieser Datei wurden im Laufe des Trainings weitere Listen mit reduzierten Personen erstellt und verwendet.

6.4.1.2 Anpassung der Bilder

Mit dem Programm *Cpppm.pl* und mit Hilfe der Listen werden die ausgewählten 16 Bilder pro Emotion in ein neues Verzeichnis „/ppm” kopiert, welches die gleiche Struktur wie die Originaldatenbank aufweist (siehe Kapitel 3). Nur beinhaltet diese nun genau die Bilder, die die für das Training und die Erkennung wichtige Änderung der Mimik zeigen.

Als nächstes wird das Programm *runxv+.pl* gestartet, welches die ersten Bilder jeder Person und Mimik mit dem Tool *xv+* aufruft. Dieses speichert die Position von Augen, Nase und Mund in einer **.ilab* Datei ab. Dazu muss der Mauszeiger nacheinander auf linkes Auge, rechtes Auge, Nasenspitze und Mundmittelpunkt bewegt werden und zur Abspeicherung der Position jeweils entsprechend die Tasten 5, 6, 7 und 8 gedrückt werden. Anhand dieser Informationen schneidet nun das Script *Prepfeat.pl*[11] den für die Mimik wichtigen Teil des Kopfes aus, dreht das Bild, so dass die beiden Augen auf gleicher Höhe liegen und skaliert es auf eine Größe von 196x172 Pixeln. Im übrigen ist die Farbinformation für die Erkennung der Mimik unwichtig, deshalb speichert das Script die Dateien im Grauwertformat mit 255 Graustufen als **.pgm* ab.

An dieser Stelle sind noch zwei verwendete Ansätze zu nennen. Für den Großteil der Untersuchungen wurde nur im jeweils ersten Bild einer Sequenz die Position von Augen, Nase und Mund manuell bestimmt und alle weiteren Bilder anhand dieser Koordinaten ausgerichtet. Bewegt die Testperson allerdings den Kopf, erscheint diese Bewegung in Form von unerwünschten Maximas in den im Anschluß berechneten Differenzbildern. Das Ziel dieses Versuchs sollte sein, dass die Bewegung als Bestandteil der Emotion mit in die Modelle aufgenommen werden würde und die HMM’e mit Hilfe einer großen Trainingsdatenmenge fähig wären, eine eindeutige Erkennung durchzuführen.

In zweiten Ansatz wurden alle 16 Bilder einer Sequenz gelabelt. Dadurch verschwindet zwar die Bewegung, allerdings zeigt ein Wackeln der Bildsequenz, die von ungenauer händischer Markierung der Punkte herrührt. Dieses „Zittern” zeigt sich durch noch stärkere Maxima der Differenzbilder als im ersten Fall. Der große Nachteil ist der erhöhte zeitliche Aufwand der Vorverarbeitung.

6.4.2 Merkmalsextraktion

Der entscheidende Prozess für die Mimikerkennung ist die Extraktion der für die Emotion charakteristischen Merkmale. Je mehr unnötige Information herausgefiltert wird, umso weniger aufwändig werden die Modelle und Berechnungen. Und je genauer die Merkma-

[11] aufgerufen vom Skript runPrepfeat.pl

le spezifiziert werden umso besser wird die Erkennung. Die Merkmalsextraktion trägt also einen bedeutenden Teil für den späteren Erfolg der Erkennung bei.

6.4.2.1 Berechnung der Differenzbilder

Im Falle der dynamischen Mimikerkennung stecken die wesentlichen Informationen in der Änderung der Merkmale eines Gesichts. Um diese zu extrahieren, gibt es die einfache Möglichkeit der Differenzbildung zweier aufeinander folgender Bilder, welches entsteht, indem die Grauwerte benachbarter Frames voneinander abgezogen werden. Die Berechnungsvorschrift (siehe auch [FH01]) ist folgende:

$$D'(x, y, t) = P(x, y, t) - P(x, y, t - 1) \tag{6.9}$$

Da die Differenzen nur minimal sind, ist eine Verstärkung mit einem linearen Verstärkungsfaktor notwendig. Beispielsweise wird ein Grauwert von 25 über eine lineare Funktion auf einen Wert von 100 abgebildet, was den Vorteil hat, dass dieser nun viel deutlich sichtbarer wird[12]. Damit nicht interessierenden Änderungen (Rauschen) herausgefiltert werden, wird diese Verstärkungskurve noch mit einem Threshold überlagert, der Werte unter einem bestimmten Wert nach Gleichung 6.10 auf Null abbildet.

$$D(x, y, t) = \begin{cases} 0 & : \|D'(x, y, t) < S\| \\ D'(x, y, t) & : \|D'(x, y, t) \geq S\| \end{cases} \tag{6.10}$$

So wird gewährleistet, dass nur die entscheidenden Änderungen verstärkt werden. Das optimal verstärkte Differenzbild sieht dann so aus:

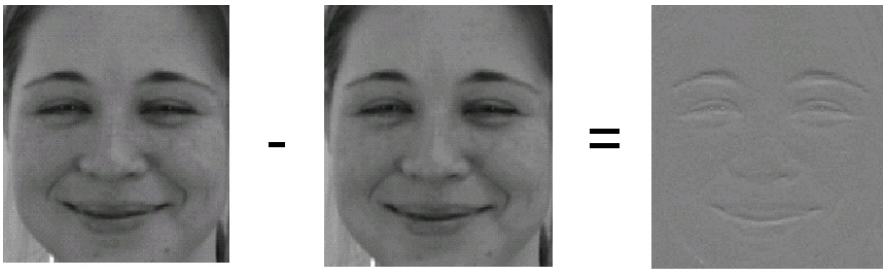


Abbildung 6.4: Differenzbild

Die hellen (=hohen) Werte sind die Bereiche, in denen Bewegung stattfindet. Die Größe der Grauwerte in einem Frame entsprechen also der Dynamik der Bewegung. Je heller die Bereiche, desto größer ist die Änderung. Wie zu erkennen ist, lässt sich allein anhand der Differenzbilder die Art der Emotion erschließen, was der Beweis für die erfolgreiche Extraktion der für die Mimik entscheidenden Merkmale ist.

Die Differenzbildung wird durch starten des Scripts *differences.pl* vorgenommen. Dieses führt nun für alle 16 Bilder pro Emotion die Differenzbildung durch Aufruf des eigentlichen Programms *differencedog* im */bin* Verzeichnis aus. Somit entstehen 15 Differenzbilder im *diff*.pgm* Format, mit denen die folgende DCT durchgeführt werden kann.

[12] je größer desto heller; 0 = schwarz, 255 = weiss

6.4.2.2 Diskrete Cosinus Transformation

Eine Merkmalsextraktion basiert in diesem Fall auf einer Diskreten Cosinus Transformation, wobei jedes Bild einer Sequenz mit einem Abtastfenster von oben nach unten und von links nach rechts analysiert wird. Für ein Fenster der Größe NxN wird die Transformation gemäß Gleichung 6.11 durchgeführt.

$$C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{x=0}^{N-1} f(x,y) \cdot \cos \frac{(2x+1)u\pi}{2N} \tag{6.11}$$

Das bisherige System (siehe Abschnitt 6.3) wurde bereits nach [FH01] mit verschiedenen Parametern für die Merkmalsextraktion getestet. Dabei hat ein DCT Block von 16x16 Pixel mit jeweils 75% Überlappung das beste Ergebnis geliefert. Um die Größe des Merkmalsvektors zu reduzieren, enthält dieser lediglich die ersten 10 Koeffizienten des oberen Dreiecks der DCT Block-Matrix.

Die Durchführung der DCT für jedes Differenzbild erfolgt durch Aufruf von *mkfeat.pl*. Die dabei berechneten Merkmalsvektoren werden als *diff*.f1* Textdateien in */feat* unter der jeweiligen Emotion/Person abgelegt. Die eigentliche Transformation wird von dem Programm *dct_trans* mit der Angabe der Fenstergröße von 16 Pixel und einer Überlappung von 12 Pixeln[13] vorgenommen. Um mit diesen nun den Trainingsprozess durchzuführen, müssen sie in eine eine HTK-Form gebracht werden.

6.4.2.3 Erstellung der Feature Dateien

Dazu wird das Script *featdiff.pl* aufgerufen, welches mit dem Programm *dct_trans* die vorher berechneten Merkmalsvektoren, durch Einfügen spezieller Startzustände zu Beginn jeder neuen Spalte, in ein 1D-HMM umwandelt, welches äquivalent zu dem entsprechendem P2D-HMM ist. Dies geschieht durch Angabe der Parameter *m=2* [14] und *st=2* [15]. Das Ausgabeformat sind *diff*.f0* Dateien.

Für den abschließenden Trainingsprozess werden die P3D-Modelle benötigt. Dies übernimmt das Script *mk3dHtk.pl*, welches mit *dct_trans* und dem Parameter *„m=3"* äquivalente P3D-Modelle durch Einfügen spezieller Startzustände für die Spalten (wie oben) sowie zusätzlich für die Bilder als 1D-HMM in HTK-Form erzeugt. Anschließend werden für jede Emotion/Person die Modelle der 15 Bilder zu einem einzigen äquivalenten P3D-HMM mit der Bezeichnung *all_htk.f1* zusammengefügt. Dieses enthält demnach die Beschreibung einer kompletten Emotionssequenz und wird im späteren Training benötigt.

Die vorher erstellten 2D-Merkmalsdateien werden mit dem Script *cpfeatdiff.pl* in das neu erstellte Trainingsverzeichnis */train* mit folgendem Dateinamen kopiert:

[Emotion][Person]_[Bild].f0

[13] entspricht einer Überlappung von 75%

[14] erzeugt die HTK-Form

[15] spezifiziert den Modelltyp; t=1/2/3 = 1D/P2D/P3D

Ein Beispiel wäre *anger0002_2_001.f0*, wobei es sich hier um die Merkmale der Mimik Ärger des 1.Bildes der Aufnahme 2 von Person 2 handelt. So bleiben die Merkmale auch später noch zuordenbar.

Die zusammengefügten äquivalenten 3D-Modelle werden mit dem Programm *CpFeats.pl* mit der Bezeichnung

[Emotion][Person].f0

in das Trainingsverzeichnis */train/all* kopiert. An dem Fehlen der Bildnummer erkennt man die zusammengefassten P3D-Modelle, die die Merkmale aller 15 Bilder enthalten.

6.4.3 Training

Das Training der Modelle erfolgt wie in den vorhergehenden Abschnitten beschrieben. Dabei wurden nur Anpassungen der Skipten durchgeführt. Der genaue Ablauf findet sich in der Datei *run_all.pl*.

6.5 1D-Vektormodell

Ein andere, einfachere Möglichkeit der Merkmalsextraktion nutzt einen Ansatz aus der Bewegungssegmentierung und -erkennung in „Meeting Room" Szenarios (siehe [FW04]). Da das Auftreten von Mimiken immer mit Bewegung verbunden ist, versucht dieser Ansatz die Merkmale in globale Bewegungsvektoren zu extrahieren und so die Bewegungsinformation zu speichern. Wie auch schon in anderen Ansätzen gezeigt wurde, ist die Differenzbildung eine effiziente Methode, um die zugrunde liegende Bewegungsinformation zu extrahieren. Aus diesen Differenzbildern wird dann der Merkmalsvektor berechnet. Hierbei handelt es sich ebenfalls um ein dynamisches Verfahren.

6.5.1 Merkmalsvektor

Der *Massenschwerpunkt* eines Gesichts $\vec{m} = [m_x(t), m_y(t)]^T$ drückt den Bewegungsschwerpunkt in x- und y-Richtung und berechnet sich zu

$$m_x(t) = \frac{\sum\limits_{(x,y)\in R_i} x|I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \; m_y(t) = \frac{\sum\limits_{(x,y)\in R_i} y|I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|}, \tag{6.12}$$

welches die ersten beiden Einträge des Vektors darstellen. Um die Dynamik der Bewegung (entspricht der Beschleunigung) zu erfassen, werden auch die *Änderungen* des Massenschwerpunktes in x- sowie in y-Richtung

$$\Delta m_x(t) = m_x(t) - m_x(t-1) \text{ und } \Delta m_y(t) = m_y(t) - m_y(t-1) \tag{6.13}$$

berechnet. Zusätzlich wird noch die mittlere *Standardabweichung* eines Pixels (x, y) vom Zentrum der Bewegung ermittelt, um diese zu beschreiben.

$$\sigma_x(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|(x - m_x(t))}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \tag{6.14}$$

$$\sigma_y(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|(y - m_y(t))}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \tag{6.15}$$

Mit diesem Merkmal ist es möglich große Bewegungsänderungen von kleineren zu unterscheiden. Im Hinblick auf die Mimik kann dies zur Unterscheidung der Bewegung des Kopfes und Änderung der Gesichtszüge (also der für die Emotion entscheidenden Merkmale) dienen. Dieses Merkmal wird nach [FW04] auch als „wideness of motion" bezeichnet.

Das letzte wichtige Kennzeichen zur Beschreibung einer Bewegung ist die *Bewegungsintensität*, welche einfach als der absolute Mittelwert der Bewegungsverteilung angesehen werden kann

$$i(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} 1}, \tag{6.16}$$

wobei ein großer Wert von $i(t)$ eine starke Bewegung repräsentiert während ein kleiner Wert ein fast gleich bleibendes Bild beschreibt. Letzteres wäre beispielsweise am Anfang und am Ende der Emotion der Fall, während sich dazwischen (also beim Ausdruck) ein größerer Wert einstellt.

Mit diesen Merkmalen, die die Aktivität der wichtigsten Änderungen einer Bewegung beschreiben, kann die Zahl der ansonsten hochdimensionalen Muster drastisch reduziert werden. Es entsteht nun anstatt des in Kapitel 6.4.2.2 beschriebenen DCT-Koeffizienten ein siebendimensionaler Vektor, der die Hauptcharakteristika der beobachtbaren Bewegungsänderungen erfasst.

$$\vec{x}_t = [m_x, m_y, \Delta m_x, \Delta m_y, \sigma_x, \sigma_y, i]^T \tag{6.17}$$

6.5.2 Ablauf der Extraktion

Die Realisierung der Merkmalsextraktion erfolgt dabei durch ein in Matlab implementiertes Programm. Dieses lädt die Bildsequenzen, führt die Differenzbildung durch, berechnet die Merkmalsvektoren und speichert diese für die Bildsequenz in einem speziellen Format ab (siehe Grafik 6.5.

Die bereits vorverarbeiteten Bilder müssen dabei in einem bestimmten Verzeichnis namens „/examples" der Form [Emotion/Bildverzeichnisse] liegen. Der Aufruf der Extraktion erfolgt durch das Programm *domimikextraction.m*. Nachfolgendes Training der Daten geschieht im Verzeichnis „/actionrec". Dort müssen zuerst die Merkmalsdateien durch Aufruf des Shell-Scripts *catchfeatures.sh* in ein für die HTK-Tools lesbares Format umgewandelt

Abbildung 6.5: Extraktion des siebendimensionalen Merkmalsvektors

und die entsprechenden Trainings- und Testverzeichnisse kopiert werden. Dann beginnt das Standard-Training der eindimensionalen HMM'e.

6.5.3 Training

Das Training läuft grundsätzlich wie bei den äquivalenten P3DHMM'en in Abschnitt 6.4 nur für den eindimensionalen Fall ab. Die Scripten laufen ähnlich wie bei dem 3D-Training unter Perl ab.

Durch Starten des Scripts *runme* wird für verschiedene Anzahlen von Zuständen des HMM's das Programm *complete_run* aufgerufen. Dieses erstellt einen Prototypen abhängig von den Zuständen, erzeugt und trainiert die Modelle gemäß der Modellliste *model.list* mit den zugehörigen Bilder und führt im Anschluß sofort die Erkennung mit den Daten aus dem „/test"-Verzeichnis durch. Die Ergebnisse werden dabei in die Datei *res.txt* geschrieben.

Im weiteren Verlauf wird von dem Hauptprogramm *runme* noch für jeden Zustand das Script *mix_up* gestartet, dass die Anzahl der Mixtures erhöht und damit wieder obige Trainings- und Erkennungsprozedur ausführt. Dabei werden insgesamt 6 Mixtures berechnet. Die Ergebnisse sind in der Datei *res.txt* aufgelistet.

Anzumerken ist an dieser Stelle, dass die erstellten Merkmalsdateien auch als Eingang einer Support Vektor Machine angewendet wurden. Die Ergebnisse mit den SVM'en waren allerdings nicht aussagekräftig.

Der wesentliche Vorteil dieses Verfahrens liegt in der Geschwindigkeit für die Extraktion der Merkmale beziehungsweise der Erstellung der Modelle. Die Merkmalserstellung dauert dabei rund 10 Minuten für 200 Modelle. Das Training der Modelle mit Erstellung der Mixtures und Variation der Zustände sowie der gleichzeitige Test nehmen dabei nur 10 Minuten in Anspruch. Der Nachteil ist natürlich die Reduktion auf nur sieben Merkmale, wobei ein Großteil der Information verloren geht.

Kapitel 7

Ergebnisse

In diesem Kapitel werden alle durchgeführten Tests aufgelistet und deren Ergebnisse präsentiert. Zum Training der Modelle wurde dabei verschiedenen Listen (wie auch in 6.4.1.1) der aufgenommenen Personen aus der Datenbank erstellt und verwendet. Das Training erfolgt wie in Kapitel 6.1.4 auf Seite 50 beschrieben, die Erkennung mit dem HTK-Tool *HVite*. Aufgrund der relativ langen Rechenzeit der Modelle wurden für einige Versuche die Anzahl der trainierten Personen reduziert, um die grundlegende Tendenz festzustellen. Im Anschluß wurden Trainingsparametervariationen durchgeführt, um Verbesserungen für die Modelle zu erzielen. Um einen Vergleich mit den alten Daten zu erhalten, wurden die Emotionen auf vier Klassen reduziert und nochmals einige Tests durchgeführt. Abschließend erfolgt eine Zusammenfassung und eine Bewertung der Ergebnisse.

7.1 7 Klassen

Dieser Abschnitt beinhaltet alle durchgeführten Trainings- und Testdaten mit ausgewählten Personen der neuen Datenbank für die sechs Basisemotionen und den neutralen Zustand. Dabei wurden die Modelle mit unterschiedlichen Anzahlen von Personen trainiert.

7.1.1 Modelle für zehn Daten pro Emotion

Die berechneten Modelle wurden mit den zehn besten Aufnahmereihen unterschiedlicher Personen pro Emotion trainiert. Die dazu verwendete Trainingsliste *traindir* findet sich im */list* Verzeichnis. Bei der Vorverarbeitung wurde zum Zuschneiden der Bilder nur im ersten Bild Augen, Nase und Mund markiert und anhand dieser Daten die weiteren 14 Bilder zugeschnitten.

Die Überprüfung mit den trainierten Daten ergab folgende Ergebnisse:

90%	anger	disgs	fears	happy	surpr	sadns	neutral
anger	9	0	0	0	0	0	0
disgs	0	10	0	0	0	0	0
fears	2	1	6	1	0	0	0
happy	0	0	0	10	0	0	0
surpr	0	0	0	1	9	0	0
sadns	0	0	0	0	0	10	0
neutr	0	0	0	0	0	1	9

Die Erkennungsrate liegt bei 90%. Im Idealfall müsste die Erkennung der trainierten Modelle bei 100% liegen, allerdings ist es möglich, dass lokale Maxima dicht beieinander liegen oder das auf ein lokales Maxima trainierte Modell nicht auf das optimale Maxima trainiert wurde und dadurch bei der Erkennung Fehlzuweisungen entstehen.

Diese Modelle wurden mit zu den Trainingsdaten unterschiedlicher Aufnahmereihen (insgesamt 17 pro Emotion) getestet. Die besten Testergebnisse weisen dabei folgende Verteilung auf:

31,93%	anger	disgs	fears	happy	surpr	sadns	neutral	Erkannt
anger	5	3	2	2	1	4	0	29,4%
disgs	3	3	1	2	3	5	0	17,6%
fears	1	2	0	7	3	4	0	0,0%
happy	1	1	1	12	2	0	0	70,6%
surpr	0	1	3	7	4	2	0	23,5%
sadns	3	3	1	1	0	6	3	35,4%
neutr	3	0	0	1	0	5	8	47,1%

Aussderdem wurde für diese Modell die Anzahl der Mixtures erhöht, diese ergab allerdings keine Verbesserung.

7.1.2 Modell für drei Sequenzen/Emotion

Aufgrund der langen Berechnungszeit der Modelle im letzten Trainingsschritt wurden für die ersten Versuche und für die Berechnung der Mixtures nur jeweils drei gute Mimiksequenzen pro Emotion verwendet (Liste *traindir_3emot*). Die Erkennungsrate der trainierten Daten ist in diesem Fall 100%, was auf eine fehlerfreie Merkmalsextraktion schließen lässt. Als Testdaten wurden wieder die 17 nicht in den Trainingsdaten enthaltenen Datensätze verwendet. Die Erkennungsraten sind hier aufgelistet:

27,73%	anger	disgs	fears	happy	surpr	sadns	neutral	Erkannt
anger	5	6	2	1	2	1	0	29,4%
disgs	1	5	0	4	3	3	1	29,4%
fears	1	3	1	4	5	2	0	6,2%
happy	0	0	0	7	10	0	0	41,2%
surpr	0	1	3	3	10	0	0	58,8%
sadns	2	7	1	0	1	4	2	23,5%
neutr	0	13	0	0	2	1	1	5,9%

7.1.3 Erhöhung der Mixtures

Für diese Modelle wurden nun die Mixtures erhöht. Die Erhöhung ergab eine deutliche Verbesserung der Erkennungsraten. Für das 5 Mixture haben sich die Erkennungsraten auf 30%verbessert.

30,25%	anger	disgs	fears	happy	surpr	sadns	neutral	Erkannt
anger	5	3	1	0	3	5	0	29,4%
disgs	1	3	0	7	2	4	0	17,6%
fears	0	2	2	5	5	3	0	11,8%
happy	0	0	1	7	8	1	0	41,2%
surpr	0	1	4	2	10	0	0	58,8%
sadns	3	4	0	1	2	6	1	35,3%
neutr	0	2	0	6	0	6	3	17,6%

7.1.4 Anpassung der Varianz

Mit diesen Testsequenzen habe ich auch Modelle mit angepasster Varianz trainiert. Dabei ergab sich für die Senkung der Varianz von $0,0001$ auf $0,00005$ eine Verbesserung auf 32,77%. Auch die Erhöhung auf $0,0005$ brachte eine Verbesserung auf sogar 33,61%.

Die Ergebnisse sind in den beiden dargestellten Tabellen veranschaulicht.

Varianz $0,00005$:

32,77%	anger	disgs	fears	happy	surpr	sadns	neutral	Erkannt
anger	4	6	1	2	2	2	0	23,5%
disgs	0	8	1	3	2	3	0	47,1%
fears	0	6	3	1	6	1	0	17,6%
happy	0	0	2	2	10	3	0	11,8%
surpr	0	2	4	4	7	0	0	41,2%
sadns	2	6	3	0	1	1	4	5,9%
neutr	0	2	0	0	0	1	14	82,4%

Varianz $0,0005$:

33,61%	anger	disgs	fears	happy	surpr	sadns	neutral	Erkannt
anger	5	6	1	1	2	2	0	29,4%
disgs	1	7	1	4	3	0	1	41,2%
fears	1	5	2	2	7	0	0	11,8%
happy	0	0	3	1	13	0	0	5,9%
surpr	0	2	4	1	10	0	0	58,8%
sadns	4	4	4	0	1	1	3	5,9%
neutr	0	2	0	0	0	1	14	82,4%

Hierbei ergeben sich zwar bessere Werte für die Gesamterkennungsrate, allerdings werden dabei auch einzelne Emotionen deutlich schlechter erkannt als bei dem Standard Wert für die Varianz. Beispielsweise wird Freude, welche von den bisherigen Modellen gut erkannt wurde, fast gar nicht mehr richtig zugeordnet. Dafür wir plötzlich Angst besser erkannt. Als Tendenz bleibt das Mittel zwar gleich, aber die Unterschiede für die einzelnen Mimiken größer, was als Ergebnis dieser Untersuchung festzuhalten ist.

7.2 Test der neuen Modelle für 4 Klassen

Um nun einen Vergleich der neu erstellten Modelle mit dem alten System durchführen zu können und eine Verbesserung feststellen zu können, müssen die sieben Klassen auf die ursprünglichen vier reduziert werden. Dabei ergeben sich bei dem Test der 17 Serien/Emotion höhere Erkennungsraten:

52,94%	disgs	happy	surpr	anger	Erkannt
disgs	8	0	3	6	47,1%
happy	1	13	2	1	76,5%
surpr	2	6	9	0	52,9%
anger	6	2	3	6	35,3%

Die hier aufgeführten sind ergaben unter allen reduzierten Modellen die besten Erkennungsraten.

7.3 Evaluierung des Verfahrens für 1D Vektormodell

Für die Untersuchung des 1D Verfahrens wurden als Trainings- und Testdaten die besten Emotionssequenzen der Datenbank verwendet. Die Trainingsdaten umfassten 10 Sequenzen pro Emotion unterschiedlicher Personen, die Testdaten aus insgesamt 119 Sequenzen (17/Emotion). Um einen Vergleich mit dem 3D-Verfahren zu ermöglichen, bestanden diese Daten aus den selben Sequenzen wie für die erstellten 3D-Klassifikationen aus Abschnitt 7.1.1.

7.4 Vergleich

Als Ergebnis ergaben sich die höchste Erkennungsrate zu 38,66%. Dieses Ergebnis trat bei unterschiedlichen Zuständen und zugehörigen Mixtures ein, so bei 3 States für das 4. Mixture, bei 4 States für das 5. Mixture und bei State 6 sowie 7 ebenso für das 4. Mixture.

Im Vergleich mit dem 3D-Verfahren bei bestehender Merkmalsextraktion kommt das 1D-Modell mit den Kopfbewegungen deutlich besser zurecht und erreicht bei gleicher Datenbasis die höheren Erkennungsraten.

Der P3D-Ansatz hat seine Vorteile bereits für statische und gespielte Emotionen bewiesen, allerdings ist für die natürlichen Emotionen die bestehende Merkmalsextraktion noch nicht ausreichend. Für das 1D Modell ergeben sich aufgrund der für Bewegung ausgelegten Merkmale ein wenig bessere Erkennungsraten.

Kapitel 8

Zusamenfassung und Ausblick

Thema dieser Diplomarbeit war die Entwicklung eines Setups für Blickwinkelerfassung, Pointing- und Mimikaufzeichnung sowie die Aufnahme einer Datenbank für sieben Klassen von Gesichtsausdrücken. Im zweiten Teil sollten dann mit Hilfe dieser aufgenommenen Datenbank Untersuchungen zur dynamischen Mimikerkennung mit zwei verschiedenen Ansätzen der Merkmalsextraktion und einer Klassifikation mit Hidden Markov Modellen durchgeführt werden. Abschließend erfolgte ein Vergleich beider Verfahren und die Bewertung des Systems unter Berücksichtigung der zugrundegelegten Mimikdaten.

Im ersten Teil der Arbeit wurde ein Versuchsaufbau zur Aufzeichnung von Blickwinkeln als auch für das Pointing entworfen, die technischen Vorraussetzungen für die Aufnahme geschaffen sowie ein modulares Aufnahmeprogramm zur Steuerung der Aufzeichnung und der technischen Geräte in Perl/Tk entwickelt, welches automatisiert ist und für zukünftige Aufzeichnungen weiterverwendet werden kann. Diesen Aufbau betreffend wurden somit die Grundlagen für die spätere Aufzeichnung geschaffen und das System ist für die geforderten Vorgaben sofort einsatzbereit. Darauf aufbauend und angepasst an die unterschiedlichen Anforderungen für die Aufzeichnung von Emotionen wurde ein neuer Aufbau für die eigentliche Mimikaufzeichnung umgesetzt, der überall und mit weniger Aufwand realisierbar ist. Basierend auf Erkenntnissen der Emotionspsychologie erstellte und realisierte ich ein neues Verfahren, welches die Erzeugung und Aufnahme möglichst natürlicher Mimiken erlaubt, und das Programm entsprechend angepasst. Für die Erzeugung der Emotionen habe ich speziell dafür erstellte Filmausschnitte verwendet. Mit den geschaffenen Vorraussetzungen wurde dann eine Datenbank von 19 verschiedenen Personen und von jeder einzelnen 3 Aufnahmeserien pro Emotion erstellt. Insgesamt ergeben sich somit 399 Aufnahmeserien einer Gesamtdauer von mehr als 30 Minuten vorsegmentiertem Videomaterial von sieben Klassen von Gesichtsausdrücken.

Der zweite Teil der Arbeit befasste sich mit der Merkmalsextraktion aus den aufgezeichneten Emotionssequenzen und der Klassifikation der gewonnen Merkmale mit Hilfe Pseudo 3D Hidden Markov Modellen. Dabei wurde aufbauend auf einem System für vier Klassen dieses auf sieben Klassen erweitert und damit neue Modelle mit Hilfe der natürlicheren Mimikdaten trainiert. Ziel der Untersuchungen war die Verbesserung der Erkennungsleistung des alten Systems aufgrund der größeren Trainingsdatenmenge und eine gleichzeitige Erhöhung der erkannten Klassen. Als Abschluss sollten die Ergebnisse

noch mit einem eindimensionalen Ansatz verglichen und ein Resümee gezogen werden.

Das bestehende Mimikerkennungssystem birgt durchaus großes Potential im Bereich dynamischer Erkennung von Gesichtsausdrücken. Für spätere Untersuchungen muss lediglich das Verfahren der Merkmalsextraktion so verbessert werden, dass die emotionsunspezifische Bewegung aus den Sequenzen herausgefiltert wird, so dass das Klassifikationssystem die reine extrahierte Dyamik der Mimik verarbeiten kann.

Anhang A

Isotracker

Folgende Optionseinstellungen müssen für die */dev/ttyS0* Schnittstelle unter Linux durch Eingabe folgenden Befehls vorgenommen werden:

> stty -F /dev/ttyS0 -parenb -parodd cs8 -hupcl -cstopb cread clocal -crtscts ignbrk -brkint -ignpar -parmrk -inpck -istrip -inlcr -igncr -icrnl -ixon -ixoff -iuclc -ixany -imaxbel -opost -olcuc -ocrnl -onlcr -onocr -onlret -ofill -ofdel nl0 cr0 tab0 bs0 vt0 ff0 -isig -icanon -iexten -echo -echoe -echok -echonl -noflsh -xcase -tostop -echoprt -echoctl -echoke

und

> stty -F /dev/ttyS0 115200

Im ersten Befehl werden die Parameter ein- und ausgeschaltet und im zweiten Befehl die Verbindungsgeschwindigkeit auf 115200 Baud gesetzt. Nur mit diesen Einstellungen ist ein korrekter Zugriff auf den Isotracker mit Hilfe des Programms *record.pl* über die Schnittstelle möglich.

A.1 Aufbau der Datei Info.txt

Die aktuelle Parameterdatei für alle für diese Diplomarbeit aufgezeichneten Aufnahmen der Datenbank mit allen Zeit- und Videoeinstellungen hat folgenden Inhalt:

```
Gaze-Detection Neutral neutral.mpg /elmo0/misc * sim 0 650 GAZE     smiley_mid.jpg
Gesture Neutral neutral.mpg /elmo0/misc * sim 0 650 POINT     smiley_mid.jpg
Neutral  1.Neutral neutr.gif /elmo0/misc * pic 0 30 neutr     smiley_mid.jpg
Neutral  2.Neutral smiley_mid.jpg /elmo0/misc * pic 0 30 neutr     smiley_mid.jpg
Mimik 1.Surprise surps1_scr.avi /elmo0/misc intro.avi seq 3 100 surpr     surpr.gif
Mimik 2.Surprise surps6of6.avi /elmo0/misc * seq 0 100 surpr     surpr.gif
Mimik 3.Surprise surps5of5.avi /elmo0/misc * seq 0 100 surpr     neutr.gif
Mimik 1.angry anger_road2.avi /elmo0/misc anger_new.avi seq 0 100 anger     angry_s.jpg
Mimik 2.angry anger2_scr.avi /elmo0/misc * seq 0 100 anger     angry_s.jpg
Mimik 3.angry anger_hulk4.avi /elmo0/misc * seq 0 100 anger     neutr.gif
Mimik 1.sad sadness.avi /elmo0/misc sadness_i.avi seq 2 200 sadns     sad.jpeg
Mimik 2.sad sad_eisk.avi /elmo0/misc * seq 2 200 sadns     sad.jpeg
Mimik 3.sad sad_arm.avi /elmo0/misc * seq 2 300 sadns     neutr.gif
Mimik 1.Disgust disgs.avi /elmo0/misc disgust_new.avi seq 0 150 disgs     disgs.jpg
Mimik 2.Disgust Ekel_indy.avi /elmo0/misc * seq 2 150 disgs     disgs.jpg
Mimik 3.Disgust disgs1_13.avi /elmo0/misc * seq 2 150 disgs     neutr.gif
Mimik 1.Fear fear_summer.avi /elmo0/misc fear_new.avi seq 51 150 fears     scared.jpg
Mimik 2.Fear fear.avi /elmo0/misc * seq 32 150 fears     scared.jpg
```

```
Mimik 3.Fear fear_event.avi /elmo0/misc * seq 0 200 fears     neutr.gif
Mimik 1.happy happy_shot.mpeg /elmo0/misc happiness_new.avi seq 2 100 happy     smiley_s.jpg
Mimik 2.happy happy1_scr.avi /elmo0/misc * seq 2 100 happy     smiley_s.jpg
Mimik 3.happy happy2_scr.avi /elmo0/misc * seq 0 150 happy     neutr.gif
Neutral  3.Neutral neutr.gif /elmo0/misc * pic 0 30 neutr     smiley_mid.jpg
Ende

#Art Beschreibung Video-Sequenz HOME-DIR Intro-VIDEO Var Start Dauer DEST.-DIR   Bild
#Args[0] Args[1] Args[2] Args[3] Args[4] Args[5] Args[6] Args[7] Args[8]     Args[9]
```

A.2 Dokumentation der Skripten

Alle hier beschriebenen Skripten, die zur Vorverarbeitung, zur Merkmalsextraktion und zum Training verwendet wurden, befinden sich in /scripts Verzeichnis. Die im folgenden verwendete Verzeichnisangaben beziehen sich auf das Home Verzeichnis $HOME. In allen Skripten müssen die Pfade in der $HOME-Variable auf das jeweilige Heimat-Verzeichnis angepasst werden. Des weiteren benötigt wird eine Listen-Datei (entweder *traindir* oder *testdir*), angegeben in der Variable $PGM.

Für das Funktionieren der Scripten werden noch folgende Dateien benötigt:

model.list = Datei mit den 7 Emotionen
mimik.slf = Erkennungsnetzwerk erstellt vom Script *mkpersnet.pl*
mimik.dict = Wörterbuch der Emotionen (Überbleibsel aus der Spracherkennung)

Benötigte Umgebungen unter Linux:

Perl + Libraries

- **run_all.pl**:

  Bash-Skript, welches alle für das Training erforderlichen Scripten aufruft, nacheinander startet und die Trainingsmodelle sowie das Master-Model-File im Verzeichnis /scripts/hmm0/ erstellt. Die Erkennung erfolgt mit *run_test.pl*.

- **Cpppm.pl**:

  Kopiert die entsprechenden 16 Bilder aus der Datenbank (definiert in der angegebenen Liste in der Variable $PGM) in das das Verzeichnis /ppm (welches von dem Programm erzeugt wird) und ändert die Nummerierung dieser beginnend in eine aufsteigende Reihenfolge beginnend bei *p_000.ppm*.

- **runxv+.pl**

  Startet für jedes erste Bild (bzw. für alle Bilder) aus der Liste das Tool *xv+* zur Definition der Punkte von Augen, Nase und Mund durch Bewegen des Mauszeigers auf die entsprechenden Positionen im Gesicht und durch nacheinander drücken der Tasten

  5=linkes Auge, 6=rechtes Auge, 7=Nasenspitze, 8=Mittelpunkt des Mundes

Die Abspeicherung der Koordinaten erfolgt in einer *.*ilab* Datei.

- textbfrunPrepfeat.pl

  Ruft das Programm *Prepfeat.pl* auf. Dieses nimmt anhand einer *.*ilab* Datei im jeweiligen Bilderverzeichnis (erstellt von *xv+*) das Ausschneiden und Drehen des Kopfes sowie ein Rescaling auf eine im Programm eingestellte Größe vor (hier 196x172). Es außerdem festgelegt werden, dass entweder die Ilab-Datei des ersten Bildes (p_000.ilab) oder diejenigen für alle Bilder benötigt werden. Im ersten Fall werden die definierten Koordinaten des ersten Bildes auf alle weiteren angewendet, ansonsten für jedes einzelne Bild eine *.*ilab* Datei ausgelesen und zur Berechnung verwendet.

- **cppgm.pl**

  Kopiert alle zugeschnittenen Grauwertbilder aus der angegebenen Liste in ein eigenes */feat* Verzeichnis mit gleicher Struktur wie die Datenbank.

- **differences.pl**

  Erstellt für jedes *p_*.pgm* in */feat* für jede Person die Differenzbilder der vorliegenden Grauwertbilder mit Hilfe der ausführbaren Datei *differencedog*, die sich im */bin* Verzeichnis befindet. Dabei werden eimal die Werte in diff*.dd Dateien und zum anderen die entstehenden Differenzbilder als diff*.pgm abgespeichert.

- **mkfeat.pl**

  Führt die Feature-Extraktion (Berechnung der DCT-Koeffizienten der Differenzbilder) mit dem Programm *dct_trans* durch und schreibt diese in *.*f1* Dateien.

- **featdiff.pl**

  Erzeugt durch Aufruf von *dct_trans* aus den in den *.*f1* Dateien stehenden Merkmalen die eindimensionale HTK Form mit speziellen Startzuständen der Spalten für die Berechnung der P2D-Modelle.

-

-

-

Literaturverzeichnis

[Bar99] M. Bartlett, J. Hager, P. Ekman und T. Sejnowski. “Measuring facial expressions by computer image analysis.”, 1999. 40, 42, 43, 44, 45

[Bar02] M. S. Bartlett, G. Littlewort, I. Fasel und J. R. Movellan. “Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.”, 2002. 45

[Coh02] I. Cohen, N. Sebe, F. Cozman, M. Cirelo und T. S. Huang. *Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data*, 2002. 16, 41

[Don99] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman und T. J. Sejnowski. “Classifying Facial Actions.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, Nr. 10, Seiten 974–989, 1999. 16, 40

[Ess95] I. A. Essa und A. Pentland. “Facial Expression Recognition Using a Dynamic Model and Motion Energy.” In *ICCV*, Seiten 360–367, 1995. 43

[Fai98] G. Faigin. *Mimikzeichnen leichtgemacht*. Taschen Deutschland, 1998. 3, 4, 5, 6, 7, 8, 9, 10, 11

[FH01] F. W. Frank Hülsken und G. Rigoll. “Facial Expression Recognition with Pseudo-3D Hidden Markov Models.” In *23. DAGM-Symposium, Tagungsband Springer-Verlag*. Munich, Germany, September 2001. 16, 52, 54, 59, 60

[FW04] M. Z. Frank Wallhoff und G. Rigoll. *Action Segmentation and Recognition in Meeting Room Scenarios*. Munich University of Technology, Institute of Human-Machine-Communication, 2004. 61, 62

[lT02] Y. li Tian, T. Kanade und J. F. Cohn. “Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity.” *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002. 40, 41

[Mic03] P. Michel und R. E. Kaliouby. “Real Time Face Expression Recognition in Video using Support Vector Machines.” *ICMI’03*, 2003. 45

[Mül02] S. Müller, F. Wallhoff, F. Hülsken und G. Rigoll. “Facial Expression Recognition Using Pseudo 3-D Hidden Markov Models.” In *16th Int. Conference on Pattern Recognition (ICPR)*. Quebec, Canada, August 2002. 51, 52

[Oer96] H. Oertel. *Hidden Markov Modelle - Theorie und Anwendungen, Ausarbeitung im Rahmen eines Seminars zur Mustererkennung und Klassifikation*. Technische Universität Berlin, Fachbereich Informtik, 1996. 47, 49

[Ple04] T. Plechinger. *Objektverfolgung mit I-Condensation*. Munich University of Technology, Institute of Human-Machine-Communication, 2004. 51

[Rab89] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE*, 77, Nr. 2, Seiten 257–285, Februar 1989. 47, 49, 51, 53

[Rai04] M. A. Rainer. *Auslöser von Emotionen, Ausarbeitung zur Lehrveranstaltung Motivation und Emotion*. Univ. Prof. Dr. M. Trimmel, 2004. 11

[Rus03] G. Ruske. *Manuskript zur Vorlesung: Automatische Mustererkennung in der Sprachverarbeitung*. Technische Universität München, 2003. 47

[Sam94] F. Samaria. *Face Recognition Using Hidden Markov Models*. Doktorarbeit, Engineering Department, Cambridge University, Trumpington Street, Cambridge CB2 1PZ, UK, Oktober 1994. 48, 50, 53

[St98] E. G. Schukat-talamazzini und H. Niemann. “ISADORA - a Speech Modelling Network Based on Hidden Markov Models.” *Proceedings on an unknown conference*, Januar 1998. 48, 49, 50, 51

[Zeh98] C. A. Zehnder. *Informationssysteme und Datenbanken*. Teubner, Stuttgart, 6 Auflage, 1998. ISBN 3-519-32480-6. 15

[Zha98] Z. Zhang, M. Lyons, M. Schuster und S. Akamatsu. “Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron.” *Proceedings of the Third IEEE International Conference (FG’98)*, 1998. 45