Universität Karlsruhe (TH)

Studienarbeit

Konfidenzbasierte multimodale Fusion von Audio und Video zur Personenidentifikation

Studienarbeiter:	**Philipp Große**
Betreuer:	**Prof. Dr. A. Waibel**
Betreuer:	**Dipl.-Inform. H. Holzapfel**
Tag der Abgabe:	**10.06.2008**

Wintersemester 2007/2008

Fakultät für Informatik
Institut für Theoretische Informatik (ITI)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Ein langjährige Traum vieler Wissenschaftler ist es, einen Roboter nach dem Vorbild des Menschen zu erschaffen. Neben dem menschlich motivierten Äußeren, ist bei dieser Art von Roboter auch eine möglichst menschliche Verhaltensweise gewünscht. Der Roboter soll laufen, Gestik zeigen und sogar Gefühle empfinden können, wie man dies vom menschlichen Vorbild gewohnt sind. Ein äußerst wichtiger Aspekt in diesem Zusammenhang ist die Kommunikation und Interaktion mit dem Menschen. Ist die Maschine in der Lage, Menschen zu erkennen und zu identifizieren, so ermöglicht dies dem Roboter, auf natürlichere Art und Weise mit Personen zu kommunizieren.

Inzwischen existieren Systeme, die das Gesicht oder die Stimme eines Menschen wiedererkennen oder eine Identifikation anhand anderer biometrischer Merkmale vornehmen. Schwerpunkt der vorliegenden Studienarbeit ist die Entwicklung eines Systems, das zwei der soeben erwähnten Systeme zu einem neuen zusammenfügt. Das bringt vor allem den Vorteil einer robusteren Erkennung mit sich, da ein solches multimodales System auch dann brauchbare Hypothesen liefern kann, wenn die Aufnahmebedingungen dies für eine einzelne Modalität schwer oder gar unmöglich machen würde. Insgesamt erhofft man sich also eine Verbesserung der Erkennungsrate, gegenüber der Einzelmodalitäten.

Ekenel, Fischer, Jin und Stiefelhagen konnten in ihrer Arbeit [1] zeigen, dass für die Fusion von Audio und Video zur Personenidentifikation, eine adaptive Gewichtung der Modalitäten einen positiven Einfluss auf die Erkennungsrate hat. Ihr adaptiver Ansatz gewichtete die Modalitäten dabei jedoch relativ starr anhand der Erkennungsrate und deren Zusammenhang mit einem einzelnen Konfidenzmaß. Dabei konnte bereits in der Arbeit von Könn, Holzapfel, Ekenel und Waibel [7] gezeigt werden, dass

sich mit Hilfe der logistischen Regression verschiedene Konfidenzmerkmale zu einer guten Konfidenzvorhersage kombinieren lassen. Eines der Ziele der Studienarbeit war es daher auch weitere Konfidenzmerkmale in ihrer Eignung zu untersuchen, und die daraus resultierenden Konfidenzen dann einerseits zum Zweck einer adaptiven Gewichtung nutzbar zu machen, und andererseits weitere Komponenten - wie dem Dialog-System - als Maß der Sicherheit über getroffene Hypothesen anzubieten.

1.2 Überblick

Die vorliegende Arbeit unterteilt sich in mehrere Abschnitte, die unterschiedliche Schwerpunkte setzen. Kapitel 2 bietet einen Überblick über grundlegende Verfahren, Algorithmen und Methoden, die in dieser Arbeit zum Einsatz kamen. Dabei betrachten wir zunächst die Einzelmodalitäten Video und Audio und kommen schließlich auf die Fusion dieser Beiden für eine multimodale Gesamtausgabe zu sprechen. Kapitel 3 stellt den eigentlichen Kern der Arbeit dar und beschreibt die mehrstufige Architektur des Erkenners und den Einsatz der verschiedenen Konfidenzmaße. Kapitel 4 befasst sich mit den gesammelten Daten und beschreibt die durchgeführten Experimente und Evaluationen. Kapitel 5 diskutiert die erzielten Ergebnisse und zeigt Ansätze für zukünftige Arbeiten auf. Das Abschließende Kapitel fasst die gesamte Arbeit schließlich zusammen.

2 Grundlagen

Das folgende Kapitel schafft ein Überblick über die verschiedenen Komponenten, Methoden und Algorithmen, die in der vorliegenden Arbeit zum Einsatz kamen und auf die wir in späteren Kapitel zurückgreifen werden. Das Kapitel untergliedert sich dabei in drei Teile, wobei sich die ersten zwei Teile mit den Einzelmodalitäten auf Audio und Video befassen, während der dritte Teil Grundlagen für die multimodale Fusion darstellt.

2.1 Bildvorverarbeitung und Merkmalsextraktion

Im weiteren Verlauf wollten wir uns mit den Grundlagen der Personenerkennung auf Videodaten befassen. Dazu beschriebt Kapitel 2.1.1 wie bei einem Bild entschieden wird ob eine Person darauf zu erkennen ist, die beiden Kapitel 2.1.2 und 2.1.3 wie ein solches Bild in eine Vektordarstellung überführt wird und Kapitel 2.1.4 beschreibt schließlich, wie eine solche Vektordarstellung zur Personenerkennung genutzt werden kann.

2.1.1 Haar Kaskaden Klassifikation

Die Erkennung und Identifikation einer Personen auf Videodaten setzt zunächst die Lokalisierung eines Gesichtsauschnittes innerhalb eines Einzelbildes voraus, welches in der Regel größer ist als eben jener. Darüber hinaus müssen neben dem Gesichtsausschnitt auch die beiden Augen detektiert werden, da anhand ihres Abstand und ihrer Koordinaten das Bild rotiert und auf eine fixe Größe skaliert wird. Die „Normierung" des Gesichtsbereichs wird durchgeführt, um eine Basis zu schaffen auf deren Grundlage die Gesichter verschiedener Personen sinnvoll verglichen werden kann, ohne dabei den jeweiligen Neigungs- und Blickwinkel des Ge-

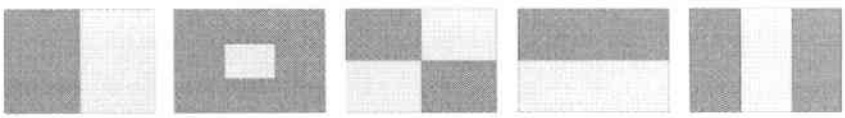

Abbildung 2.1: Einige Beispiele für Haar-like-features.

sichtes berücksichtigen zu müssen. Sowohl für die Gesichtsauschnittsdedektion als auch für die der Augen verwenden wir den von Viola und Jones entwickelten Haar Kaskaden Klassifikator[19], der eine schnelle und effiziente Objekterkennung auf Bildern ermöglicht und somit auch in einem Echtzeitsystem einsetzbar ist. Der Klassifikator arbeitet dabei auf Basis sogenannter *Haar-like-features* und *boosted classifier cascades* auf die wir im Folgenden näher eingehen werden.

Haar-like-features (siehe Abb. 2.1) sind schlichte Rechtecks-Merkmale, die weder Farbinformation noch Veränderungen über Bildsequenzen, sondern lediglich den Intensitätsgrad der Graubildpixel innerhalb bestimmter rechteckiger Bildausschnitte berücksichtigen. Dabei berechnet sich die tatsächliche Ausprägung eines Haar-like-feature aus der Summe und Differenz von bis zu vier solcher Bildregionen. Da Haar-like-features sowohl in ihrer Größe, als auch in ihrer Position innerhalb des Suchfensters variieren können, ist die Anzahl möglicher Haar-like-features bereits für sehr kleine Suchfenster (in unserem Fall 24x24-Pixel) sehr groß [17].

Obwohl Haar-like-features verglichen zu anderen möglichen Merkmalsrepräsentationen sehr einfacher Natur sind, besitzen sie einen entscheidenden Vorteil: Aufgrund ihres einfachen Aufbaus können sie sehr schnell und effizient berechnet werden. Um den Rechenaufwand zu minimieren, kommen dabei so genannte *Integralbilder* zum Einsatz. Das Integralbild stellt dabei eine alternative Repräsentation des Bildes dar, welche eine sehr effiziente Berechnung der Intensitätssummen ermöglicht. Der Wert jedes Punktes innerhalb des Integralbildes entspricht der Summe der Intensitäten innerhalb des Rechtecks, das zwischen Bildursprung und der Position des Punktes aufgespannt wird. Folgende Gleichung gibt diesen Zusammenhang wieder:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y') \tag{2.1}$$

Wobei $i(x',y')$ dem Intensitätswert des Orginalbildes an Punkt (x',y') und $ii(x,y)$ der Summe am Punkt (x,y) entspricht. Diese Formel lässt sich in nur einem Durchlauf über das Orginalbild berechnen in dem man die Intensitätssumme der aktuellen Zeile zwischenspeichert und mit der Intensitätssumme des zuletzt berechneten Rechtecks verrechnet, wobei sich die aktuelle Zeile durch $s(x,y) = s(x,y-1) + i(x,y)$ berechnen lässt und der Wert an der zu berechnenden Position entsprechend durch $ii(x,y) = ii(x-1,y) + s(x,y)$. Das Integralbild bietet so einen sehr schnellen Zugriff auf die Intensitätssummen für Rechtecke die vom Bildursprung aus, also der linken oberen Ecke, aufgespannt werden. Da Haar-like-features jedoch an jeder beliebigen Position innerhalb des Bildes liegen können, bedarf es für den Fall, dass das Rechteck nicht vom Bildursprung ausgeht, noch eines weiteren Rechenschritts. Der Wert für ein derartiges Rechteck lässt sich - wie in Abb. 2.2 gezeigt - wird einfach durch die Differenz aus der Werten des zu berechnenden Rechtecks (D) und den oberhalb (B) und links (C) von ihm liegenden Rechtecken bestimmen, wobei wir das Rechteck linksoben (A) nocheinmal hinzu addieren müssen, um die doppelte Subtraktion der überlappenden Rechtecke (B und C) auszugleichen.

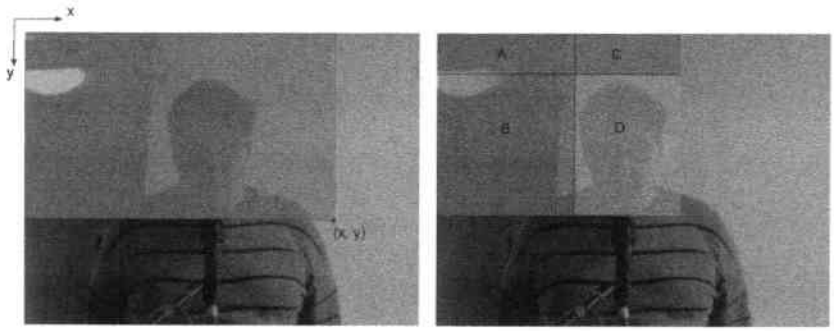

Abbildung 2.2: Das Integralbild an der Position ii(x,y) drückt die Summe der Intensitäten vom Bildursprung bis zu aktuellen Position aus.

Wie bereits erwähnt, ist die Anzahl möglicher Haar-like-features sehr groß, so dass der Klassifikator für die jeweilige Objekterkennung (in unserem Fall der Gesichtsausschnitt bzw. das linke und rechte Auge) eine entsprechende Auswahl an Feature treffen muss, die sich für die Aufgabe besonders eignen. Diese Auswahl geschieht beim Haar Kaskaden

Klassifikator nach Viola und Jones durch einen abgewandelten AdaBoost Klassifikator, der mehrere sogenannte schwache Klassifikatoren zu einem Starken bündelt und diese als Kaskade hintereinander hängt. Die eigentlich Auswahl der zu verwendenden Haar-like-features geschieht nun durch das schrittweise Anhängen des jeweils diskriminierensten (entsprechend der jeweiligen Klassifikationsaufgabe) an die Kaskade. Diese soeben beschriebenen Kaskaden werden in diesem Zusammenhang auch *boosted classifier cascades* genannt und entsprechen in gewisser Weise degenerierten Entscheidungsbäumen, wobei in jedem Schritt ein Klassifikator (in unserem Fall mit Hilfe der Haar-like-feature) entscheidet, ob das Bild zurückgewiesen wird, oder aber an den nächsten Klassifikator weiter gereicht wird, bis das Bild im Zweifelsfall durch die gesamte Kaskade durch gereicht wurde und damit als positiv klassifiziert gilt.

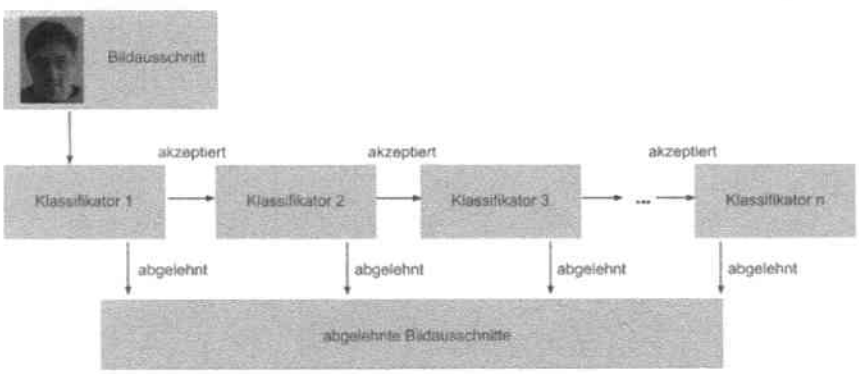

Abbildung 2.3: Boosted classifier cascades vgl. [17]

2.1.2 Diskrete Kosinustransformation

Die diskrete Kosinustransformation (DCT) ist eine lineare, orthogonale Transformation, die ein zeitdiskretes Signal vom Orts- in den Frequenzbereich transformiert und wieder zurück. Seit 1974 ist sie die am weitesten verbreitete Transformation zur Redundanzreduktion von Bildsignalen [10]. Die Gründe dafür sind vielfältig. Einerseits lässt sich die DTC sowohl in Software wie auch Hardware implementieren. Andererseits trans-

formiert die DCT Bilddaten effektiv in eine Form, die gut komprimiert werden kann. Im Gegensatz zur Diskreten Fouriertranformation rechnet man bei der Kosinustransformation nicht mit komplexen, sondern lediglich mit reellen Koeffizienten, was einen weiteren Vorteil darstellt.

Für die Gesichtsidentifikation interessieren wir uns ausschließlich für die Transformation vom Bildbereich in den Frequenzbereich, die im Folgenden mit FDCT[1] abgekürzt wird. Um Korrelation in horizontaler und vertikaler Bildrichtung zu erfassen, wird die zweidimensionale Variante der FDCT benutzt. Zu diesem Zweck wird das Bild in Blöcke von 8x8 Pixel zerlegt. Die folgende Gleichung beschreibt die zweidimensionale FDCT für einen 8x8-Block eines Bildes:

$$F_{x,y} = \frac{2C(x)C(y)}{8} \sum_{i=0}^{7} \sum_{j=0}^{7} f_{i,j} \cdot cos\left(\frac{(2i+1)x \cdot \pi}{16}\right) \cdot cos\left(\frac{(2j+1)y \cdot \pi}{16}\right)$$

wobei $f_{i,j}$ den Wert des Punktes (i, j) im Eingabebild darstellt, $F_{x,y}$ die DCT-Koeffizieten an der Stelle (x, y) und $C(x)$ und $C(y)$ die Konstanten.

$$C(z) = \begin{cases} \frac{1}{\sqrt{2}} & , z=0 \\ 1 & , z \neq 0 \end{cases}$$

Die FDCT repräsentiert jeden Block eines Bildausschnittes durch gewichtete Summen von 2D-Kosinusfunktionen, auch Basisfunktionen genannt. Das Muster links oben hat die niedrigste Frequenz und ist nur ein Einheitsblock. Von links nach rechts nimmt die "Frequenz" zwischen hell und dunkel in horizontaler Richtung zu. Von oben nach unten nimmt hingegen

[1] forward discrete cosinus transformation

die "Frequenz" zwischen hell und dunkel in vertikaler Richtung zu. Folglich nehmen sowohl die horizontalen als auch die vertikalen "Frequenzen" in diagonaler Richtung gleichzeitig zu. Folgende Grafik (2.4) verdeutlicht diesen Zusammenhang:

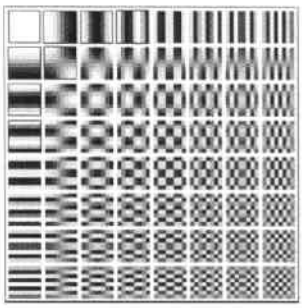

Abbildung 2.4: Basis der Diskreten Kosinustrasformation für 8x8 Pixel. [10]

2.1.3 Zig-Zag-Scan

Durch die DCT erhält man für jeden 8x8 Pixelblock 64 Koeffizienten. Der sogenannte Zig-Zag-Scan sortiert diese nun dahingehend um, dass sie in einem eindimensionalen Vektor dargestellt werden können. Er beginnt dabei in der linken oberen Ecke und bewegt sich im zickzack-Kurs zur rechten unteren Ecke. Folglich stehen die relevanten niedrigfrequenten Koeffizienten anschließend am Anfang des Vektors.

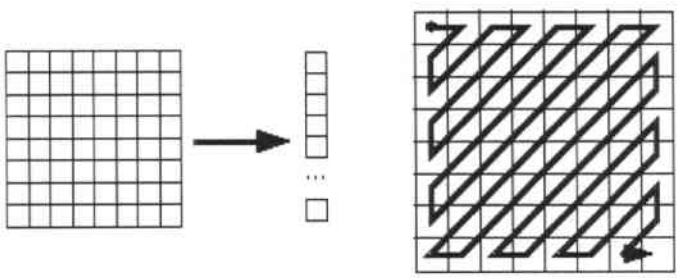

Abbildung 2.5: Der Zig-Zag-Scan ordent die Koeffizienten in einen eindimensionalen Vektor an. Entnommen [11]

2.1.4 k-Nearest-Neighbour

k-nearest-Neighbour-Algorithmen sind die elementarsten Vertreter von instanzbasierten Lernverfahren und setzen voraus, dass eine Instanz als Punkt in einem euklidischen Vektorraum repräsentiert werden kann. Anstatt eine allgemeine, explizite Beschreibung der Zielfunktion zu lernen, werden bei instanzbasierten Lernverfahren einfach die kompletten Trainingsdaten gespeichert [8]. Die Generalisierung über die Daten hinaus erfolgt erst bei der Klassifikation einer Instanz, wobei für jede neue Instanz ihre Beziehung zu den bisherigen Beispielen untersucht wird. Ein großer Vorteil dieser Klasse von Lernalgorithmen ist, dass sie auf diese Weise auch sehr komplexe Zielfunktionen approximieren können. Im Fall des k-nearest-Neightbour-Algorithmus wird - wie der Name schon andeutet - die Beziehung zwischen der zu klassifizierenden Instanz und den "trainierten" Daten durch die Distanz zwischen eben jenen gebildet. Der nächste Nachbar wird dabei i.d.R. gemäß des Euklidischen Abstands

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

ermittelt, wobei x, y Instanzen im n-dimensionalen Verktorraum darstellen. Gilt $k = 1$, erhält die zu kassifizierende Instanz die Klasse ihres nächsten Nachbarn. Für höherwertige k's werden die k nächsten Nachbarn betrachtet, wobei jeder Nachbar entsprechend seiner Distanz zur neuen Instanz gewichtet wird. Dementsprechend haben sehr nahliegende Trainingsbeispiele einen höheren Einfluss auf die letztendliche Klasse der neuen Instanz als die anderen. Die Klassifizierung erfolgt gemäß folgender Gleichung:

$$c(y) = argmax_{v \in V} \sum_{i=1}^{k} w_i \cdot \delta(v, c(x_i))$$

wobei x_i die nächsten Nachbarn, y die neue Instanz, V die Menge der Klassenlabels, $c(a)$ die Klasse von Instanz a, $\delta(a,b) = 1$ sofern $a = b$, ansonsten 0 und

$$w_i = \frac{1}{d(x_i, y)^2}$$

sind.

2.2 Audiovorverarbeitung und Merkmalsextraktion

Was folgt sind die Grundlagen der Personenerkennung auf Audiodaten, die in der Doktorarbeit von Qin Jin[5] zum Einsatz kamen und deren VoiceID Grundlage der vorliegenden Arbeit ist.

2.2.1 Kompensation von Nebengeräuschen

Eine Sprachaufnahme - insbesondere wenn sie mit einem „distent-speech" Mikrophon aufgenommen wurde - wird in der Regel durch Nebengeräusche und Hall verfälscht. Betrachtet man die Raumakustik demnach als eine invariante Verschiebung lässt sich das empfangene Audiosignal $y[t]$ wie folgt darstellen:

$$y[t] = x[t] \cdot h[t] + n[t] \tag{2.2}$$

wobei $x[t]$ die unverfälschte Sprachausgabe, $h[t]$ den Hall und $n[t]$ die Nebengeräusche zum Zeitpunkt t angeben. Eine Möglichkeit Wellenverschiebungen auszugleichen ist der Einsatz der sogenannten *Cepstrum Mean Subtraction* (CMS). Diese setzt jedoch voraus, dass der zu kompensierende Hall kürzer als das *DFT analysis window* ist, welches in der Regel eine Länge von 16ms-32ms hat, wohingegen es unter normalen Aufnahmebedingungen 50ms oder gar mehr nachhallt. Daher teilen wir den Hall $h(t)$ in zwei Abschnitte $h_1(t)$ und $h_2(t)$.

$$h[t] = h_1[t] + \delta(t-T)h_2[t]$$

$$h_1[t] = \begin{cases} h[t] & ,t<T \\ 0 & ,sonst \end{cases}$$

$$h_2[t] = \begin{cases} h[t+T] & ,t\geq 0 \\ 0 & ,sonst \end{cases}$$

womit sich die Formel (2.2) wie folgt ändert:

$$y[t] = x[t] \cdot h_1[t] + x[t-T] \cdot h_2[t] + n[t]$$

$h_1[t]$ beschreibt nun den Teil des Halls der noch innerhalb des DFT analysis windows befindet und damit mit Hilfe der CMS kompensiert werden

kann, wohingegen $x(t-T)\cdot h_2[t]$ zusammen mit den Nebengeräuschen $n[t]$ durch einfache *spectrum subtraction* ausgeglichen wird. Unter der Annahme, die Nebengeräusche $x(t-T)\cdot h_2[t]$ lassen sich durch $y[t-T]$ schätzen, kann diese durch

$$\hat{X}[t,\omega] = max(Y[t,\omega] - a\cdot g(\omega)Y[t-T,\omega], b\cdot Y[t,\omega])$$

berechnet werden, wobei es sich bei a, b und $g(\omega)$ um Korrekturfaktoren handelt, denen empirisch die Werte $a = 1.0b = 0.1$ und $g(\omega) = |1-0.9e^{j\omega}|$ zugeordnet wurden.

2.2.2 Merkmalsverschiebung

Das vorgestellte *feature warping* basiert auf der Arbeit von J. Pelecanos und S. Sridharan [13]. Dabei wird die Verteilung des *cepstral feature stream* innerhalb eines *sliding windows* auf eine Standardverteilung verschoben (siehe Abbildung 2.6), um so die Audiodaten stabiler gegenüber unterschiedlichen Aufnahmebedingungen zu machen und damit die Erkennung unterschiedlicher Sprecher zu verbessern bzw. überhaupt erst zu ermöglichen. Diese Verschiebung lässt sich auch als nichtlineare Tran-

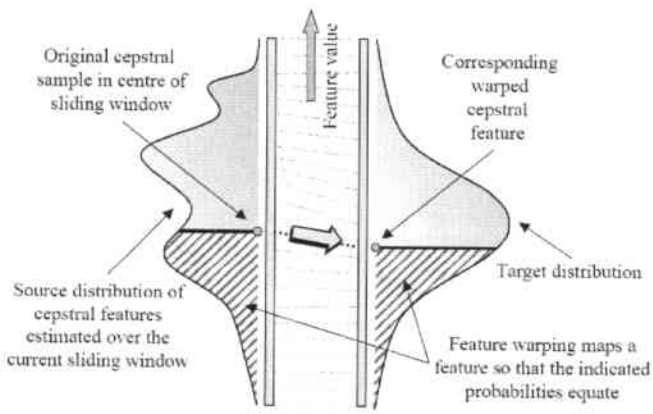

Abbildung 2.6: Merkmalsverschiebung entsprechend einer gegebenen Zielverteilung. Entnommen [13]

formation T auffassen, wobei der ursprüngliche *cepstral feature stream* X

hin zu $\hat{X}$ transformiert wird.

$$\hat{X} = T(X)$$

Eine solche Tranformation ermöglich die CDF[2] [21], welche für jede Dimension des MFCC-Vektors[3] separat jeweils das zentrale Merkmal innerhalb eines kurzen *sliding windows* dahingehend verschiebt, dass nach Ende der Transformation eine gegebene Zielverteilung erreicht wird.

2.2.3 Gaussian mixture model

Mixturemodelle gehören zur Klasse der unüberwachten Clustermethoden, was bedeutet, dass sie Daten in mehrere Cluster gruppieren, ohne dass diese Cluster im Vorfeld benannt werden könnten. Eine Mixture beschreibt dabei k Cluster durch einen Satz von k Wahrscheinlichkeitverteilungen, die die Eigenschaften der Daten innerhalb der Cluster widerspiegeln[20]. Jedes Cluster wird dabei durch eine eigene Wahrscheinlichkeitsverteilung $b_i(\vec{x})$ dargestellt und jeweils mit einen Faktor p_i gewichtet

$$p(\vec{x}, \lambda) = \sum_{i=1}^{k} p_i b_i(\vec{x})$$

wobei $\vec{x}$ ein D-dimensionaler Zufallsvektor ist und die Wahrscheinlichkeitsverteilungen bei einem Gaussian Mixturemodell jeweils durch eine D-dimesionale Gausglocke der Form

$$b_i(\vec{x}) = \frac{1}{\sqrt{(2\pi)^D \sum_i}} \cdot e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \sum_i^{-1} (\vec{x}-\vec{\mu}_i)}$$

dargestellt wird, wobei μ_i den Durchschnittsvektor und $\sum_i$ die Covarianzmatrix angeben. Darüber hinaus gilt für die Gewichtungsfaktoren p_i, dass sie sich auf eins aufaddieren lassen und folglich selbst „echte" Wahrscheinlichkeiten darstellen.[15] Demnach wird jedes Mixturemodell durch den Durchschnittsvektor μ_i, die Covarianzmatix $\sum_i$ und den Mixturegewichten p_i parametrisiert, welche häufig durch $\lambda = \{\mu_i, \sum_i, p_i\} i = 1, .., k$

[2] cumulative distribution function

[3] MFCC steht für Mel Frequency Cepstral Coefficients, welche eine aufs menschliche Gehör ausgerichtete logaritmische Darstellung eines Audiosigals ist

zusammengefasst werden. Da diese Parameter jedoch zu Beginn unbekannt sind, müssen sie mit Hilfe des so genannten EM-Algorithmus[4] geschätzt werden.

2.2.4 Sprechermodellierung

Für die Sprechermodellierung wird jeder Sprecher i durch ein eigenes GMM-Modell λ_i repräsentiert. Für die Identifiktion eines Sprecher folgt man dann folgender Regel[2]:

$$s = argmax_i \{L(Y|\lambda_i)\} \qquad Y = (y_1, y_2, ..., y_n)$$

wobei s der identifizierte Sprecher und L die *likelihood* (also Wahrscheinlichkeit) dafür ist, dass die Merkmalsreihe Y durch das Sprechermodell λ_i hervorgerufen wurde. Man entscheidet sich also für denjenigen Sprecher, dessen Modell mit der größten Wahrscheinlichkeit die Merkmalsreihe (Aufnahme) hervorgerufen hat.

2.3 Grundlagen der Fusion

2.3.1 Gewichtung der Modularitäten

Möchte man verschiedene Modalitäten fusionieren, kann man dies auf sehr unterschiedliche Wege erreichen. Und zwar auf Ebene

1. der zugrundeliegenden Merkmale
2. der Entscheidungsfindung der einzelnen Modalitäten
3. der Ergebnisse der einzelnen Modalitäten

Da wir bereits fertige Klassifikatoren für beide Modalitäten vorliegen hatten, haben wir uns für die dritte Möglichkeit entschieden, nämlich der Fusion der Ergebnisse der Modalitäten. Bei einer derartigen Fusion gibt es im wesentlichen drei Haupteinflussparamter die zu berücksichtigen sind. Und zwar die Art und Weise wie die *Scores* der einzelnen Modalitäten

[4] Expectation-Maximization-Algorithmus

normalisiert werden, wie die Modalitäten miteinander gewichtet werden und schließlich wie die Einzelergebnisse der Modalitäten zu einem neuen Ergebnis verrechnet werden. Kemal Ekenel, Fischer und Jin konnten in ihrer Arbeit[11] zeigen, dass für die Fusion von Audio und Video zur Personenidentifikation, eine sinnvolle Gewichtung der einzelnen Modalitäten essenziell ist, da in der Regel nicht davon auszugehen ist, dass jede der beteiligten Modatitäten im gleichen Maße Einfluss auf die korrekte Vorhersage des endgültigen Klassifikators haben wird. Das einfachste Vorgehen, um das Problem der Gewichtung zu lösen, wäre es sich für eine *statische Gewichtung* der Modalitäten zu entscheiden, wobei z.B. anhand der Erkennungsleistung der einzelnen Modalitäten eine feste Gewichtung bestimmt wird. Offensichtlich hat dies den Nachteil, dass eine solch allgemein festgelegte Gewichtung für einen gegebenen Einzelfall alles andere als passend sein muss und sogar kontraproduktiv für das Ergebnis der Klassifikation sein kann. Es ist daher sinnvoll statt einer *statischen Gewichtung* eine *adaptive Gewichtung* zu verwenden, also die Gewichtung der Modalitäten jeweils abhängig von den vorliegenden Daten zu gestalten. Wir verwenden zu diesem Zweck *Konfidenzen*, auf die wir in Kapitel 3.2 näher eingehen werden.

2.3.2 Logistische Regression

Unter logistischer Regression versteht man ein Verfahren zur multivarianten Analyse des Zusammenhangs zwischen binär abhängigen Variablen und mindestens einer unabhängigen Variablen[12]. Typische Beispiele für binäre abhängige Variablen sind Variablen, die das Eintreten eines Ereignisses erfassen. Diese Variablen haben nur zwei mögliche, sich gegenseitig ausschließende Ausprägungen, wie z.B. Ereignis findet statt $(Y = 1)$ und Ereignis findet nicht statt $(Y = 0)$. Nun interessiert der Einfluss der jeweiligen unabhängigen Variablen auf die Eintrittswahrscheinlichkeit. Die logistische Regression löst diese Aufgabe durch eine geeignete Transformation der abhängigen Variablen. Sie geht von der Idee der Odds aus, d.h. dem Verhältnis von $P(Y = 1)$ zur Gegenwahrscheinlichkeit $1 - P(Y = 1)$ bzw. $P(Y = 0)$ aus.

$$Odds(Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{P(Y = 1)}{P(Y = 0)}$$

Die Odds können zwar Werte >1 annehmen, doch ist ihr Wertebereich nach unten beschränkt, da sie sich asymptotisch Null annähern. Ein unbeschränkter Wertebereich wird durch die Transformation der Odds in die sog. Logits

$$Logit(Y_{1/0}) = ln\frac{P(Y=1)}{P(Y=0)}$$

erreicht, die Werte zwischen minus und plus unendlich annehmen können. In der logistischen Regression wird dann die Regressionsgleichung

$$Logit(Y_{1/0}|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

geschätzt. Es werden also die Regressionkoeffizienten β_i (auch Logit-Koeffizienten) für jede unabhängige Variable X_i bestimmt, nach denen die Logits für ein gegebenes X berechnet werden können.

Durch zwei Transformationsschritte lassen sich die Einflüsse der logistischen Regression auch als Einflüsse auf die Eintrittswahrscheinlichkeit P(Y=1) ausdrücken:

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1-e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

Insbesondere Aufgrund dieser Interpretation haben wir uns für den Einsatz der logistischen Regression entschieden, denn sie ermöglicht uns mit Hilfe verschiedener in Kapitel 3.2 vorgestellter Konfidenzmerkmale als X_i die Eintrittswahrscheinlichkeit $P(Y=1)$ zu schätzen, welche wiederum ein ideales Konfidenzmaß darstellt.

3 Klassifikation und Konfidenzberechnung

Dieses Kapitel erklärt den Ablauf und die einzelnen Schritte, die innerhalb der einzelnen Modalitäten (Audio und Video) für eine Erkennung durchgeführt werden müssen, als auch wie diese schließlich zu einer multimodalen Fusion zusammengefügt werden. Dazu verschafft der erste Teil zunächst einen allgemeinen Überblick über das komplette System und die Systemarchitektur. Nachfolgend wird im zweiten Teil zunächst näher auf die Grundlagen der Hypothesenfusion eingegangen, die sowohl innerhalb der FaceID und VoiceID als auch zur abschließenden multimodalen ID verwendet werden, bevor wir schließlich in Teil drei, vier und fünf auf jede dieser Komponenten noch einmal gesondert eingehen.

3.1 Überblick

Auf dem Weg zur Erstellung einer Annahme über die sich vor dem Roboter befindliche Person werden verschiedene Prozesse durchlaufen. Abbildung 3.1 verschafft auf Ergebnisebene einen schematischen Überblick über das Gesamtsystem und das Zusammenspiel der verschiedenen Komponenten. Das System ist in sechs separate Teilsysteme (dargestellt durch gestrichelte Linien) unterteilt: Die Einzelbild-Ebene (1), die Bildsequenz-Ebene (2), die Einzelturn-Ebene (6), die Concatturn-Ebene (5), die Audiosequenz-Ebene (4) und schließlich in der Mitte die multimodale Ebene (3). Abgesehen von der Einzelturn-Ebene werden in jedem der Teilsysteme Hypothesen in Form von n-besten Listen erzeugt, sowie eine dazugehörige Konfidenz. Der zusätzliche Übergang von der Einzelturn-Ebene hin zur Concatturn-Ebene wurde eingeführt, um der VoiceID ausreichend Audiodaten für deren Hypothesenbildung zur Verfügung zu stellen, da das verwendete Dialogsystem dazu tendierte sehr kurze (< 1 sek) Dialogwech-

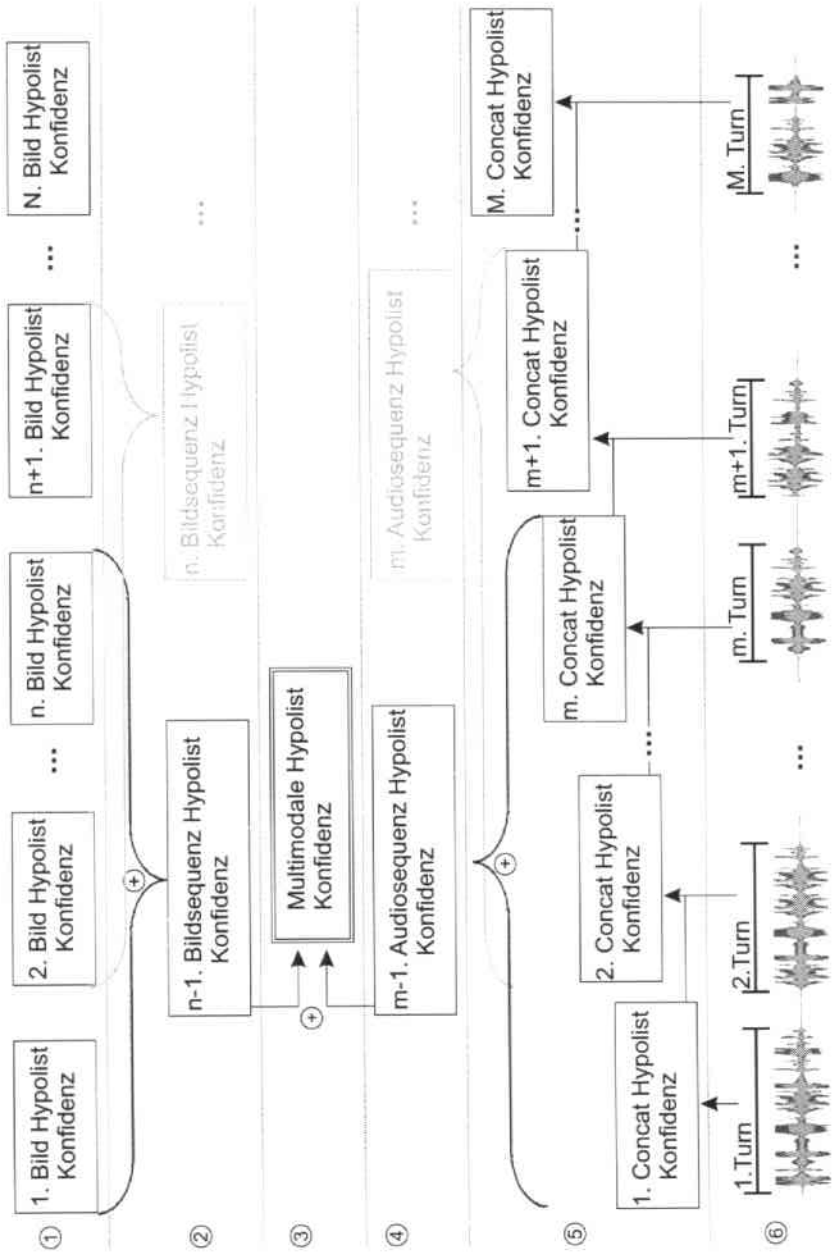

Abbildung 3.1: Aufbau des Klassifikators

sel (in diesem Zusammenhang auch *Turns* genannt) zu produzieren. Ein *Concatturn* ist also nichts anderes als eine Konkatenierung der segmentierten Audiodaten, die während des Dialogs aufgenommen wurden.

3.2 Konfidenz und Konfidenzmerkmale

Unter *Konfidenz* verstehen wir ein Maß für die Zuverlässigkeit einer Klassifikationsaussage. Sie wird dabei im Kontrast zum *Score* einer Hypothese auf seperaten Merkmalen bestimmt und als Wahrscheinlichkeit angegeben.

Je nach Klassifikationsaufgabe unterscheiden sich demnach auch die Merkmale die sich als Konfidenzmerkmale eignen. So gibt es einige Konfidenzmerkmale wie das Grauwertmittel des Bildes (Image) oder die approximierte Distanz zwischen Proband und Kamera (Dist), die sich lediglich für die Klassifikation auf Bildern eignen, da sie als Merkmale nur für diese zur Verfügung stehen. Wieder andere Konfidenzmerkmale, wie die Übereinstimmung (Agre) oder die Stabilität (Stab) der Hypothesen, lassen sich nur für Sequenzhypothesen [6] sinnvoll anwenden. Aber einige Konfidenzmerkmale konnten wir auch durchgängig anwenden, da sie sich direkt aus der Struktur der n-besten Liste gewinnen lassen, die wir bei beiden Modalitäten zur Verfügung stehen hatten. Zu diesen Konfidenzmerkmalen zählten die Entropy innerhalb der n-besten Liste (Ent), die Differenz zwischen den zwei höchsten Scores der n-besten Liste (Diff0), sowie zwei weitere differenzbasierte Merkmale(Diff1, Diff2), die auf unterschiedliche Art und Weise sämtliche Scores der n-besten Liste verrechnen.

Wie bereits angesprochen sind die Konfidenzmerkmale „Image" und „Dist" bildspezifische Merkmale, denn sie lassen sich - mehr oder minder - direkt aus den Bilddaten ableiten. Das Grauwermittel des Bildes (Image) entspricht direkt dem ersten Koeffizienten des zigzag-gescannten DCT Vektors des Bildes, und stellt ein Maß für die Beleuchtungsverhältnisse der Aufnahme dar, welche wiederum einer der großen Herausforderungen einer kamerabasierten Erkennung darstellt und sich somit als Konfidenzmaß sehr gut eignet [6]. „Dist" hingegen lässt sich impliziet aus der Größe des Gesichtsauschnitt ableiten bzw. approximieren. Es eignet sich als Konfidenzmaß, da sowohl die Haar-like-Features als auch die

Bildvektoren abhängig vom jeweiligen Training nur für einen bestimmten Distanzbereich geeignet sind [6].

Die zwei sequenzspezifischen Konfidenzmerkmale, berechnen sich aus dem Vergleich der ersten Hypothese (die Hypothese innerhalb der n-besten Liste mit höchsten Score) innerhalb der betrachteten Sequenz. Sie eigenen sich als Konfidenzmerkmal natürlich insbesondere deswegen, da sie eine Aussage der Häufigkeit der Übereinstimmung (Agre) bzw. der Entscheidungswechsel (Stab) treffen. Genau gesagt erfasst „Agre" dabei die Anzahl der Einzelbildhypothesen, die mit der aktuellen Sequenzhypothese äquivalent sind, geteilt durch die Anzahl der berücksichtigen Frames (entsprechend der Breite des sliding windows). Stab hingegen erfasst die Anzahl der Einzelbildhypothesenwechsel pro Frame (innerhalb des sliding windows).

Die vier n-besten Listen basierten Merkmale (Ent, Diff0, Diff1 und Diff2), eignen sich als Konfidenzmerkmale, da sie ein direktes Maß über die Struktur der n-besten Listen bieten und damit auch die Verwechslungswahrscheinlichkeit zwischen möglichen Hypothesen widerspiegeln. Sie berechnen sich wie folgt:

$$Ent = -\sum_{i=1}^{N} k_i \cdot log_2(k_i)$$

$$Diff0 = k_1 - k_2$$

$$Diff1 = \sum_{i=1}^{N} \frac{k_i - k_{i+1}}{i}$$

$$Diff2 = \sum_{i=1}^{N} \frac{k_i - k_{i+1}}{e^{i-1}}$$

wobei, k_i den Score der i-ten Hypothese und N die Länge der n-besten Liste angibt.

Da die vorgestellten Konfidenzmerkmale jedoch so noch nicht als Konfidenzmaß im Sinne einer Wahrscheinlichkeit dienen können, setzten wir in diesem Zusammenhang die logischtische Regression ein, um mit Hilfe entsprechender Logit-Koeffizienten eine tatsächliche Konfidenz zu erhalten.

3.3 Hypothesenfusion

In Kapitel 2.3.1 hatten wir bereits angesprochen, dass wir die Konfidenzen für eine adaptive Gewichtung der Modalitäten verwenden. Darüber hinaus verwenden wir die Konfidenzen allerdings auch, um innerhalb der Modalitäten aus einzelnen Hypothesenlisten Sequenzhypothesen zu generieren. Die Konfidenzen werden dann als Gewichte verwendet, um die n-besten Listen zur Hypothese der nächsthöheren Ebene aufzusummieren (in der Graphik 3.1 durch $\oplus$ dargestellt). Mathematisch ausgedrückt berechnet sich die Hypothesenliste der nächst höheren Ebene also durch:

$$H_{new} = \sum_{i=1}^{N} conf(H_i) \cdot H_i \tag{3.1}$$

wobei H jeweils eine n-besten Liste darstellt, die für jede ihrer n Hypothesen einen Score führt und eine Konfidenz $conf(H_i)$ besitzt. N gibt die Breite des *sliding windows* an bzw. Anzahl der berücksichtigten Hypothesenlisten. Für die multimodale Fusion ist N beispielsweise 2 und jede der Modalitäten bringt jeweils eine Hypothesenliste mit entsprechender Konfidenz ein, die dann entsprechend Formel 3.1 zur multimodalen Hypothesenliste zusammengefügt wird.

3.4 Normalisierung der Hypothesenlisten

Bevor die Hypothesenlisten jedoch fusioniert werden können, müssen sie normiert werden. Dies gilt insbesondere für die multimodale Fusion, deren zugrundeliegenden Hypothesenlisten von zwei ganz unterschiedlichen Methoden (k-Nearest-Neighbour und GMMs) erzeugt werden. Erst durch die Normierung wird sicher gestellt, dass die verwendeten Scores innerhalb der n-besten Listen vergleichbar sind und nicht etwa unterschiedliche Wertigkeit besitzen, so dass einzelne Hypothesen Listen einen über- oder unterpropotionalen Anteil am Gesamtergebnis bekommen. Das von uns verwendete Normierungsverfahren [16] folgt folgender Formel

$$\bar{s}_i = \frac{s_i - min}{\sum_{i=1}^{n} (s_i - min)}$$

wobei s_i für den Score an der i-ten Stelle, *min* für den kleinsten Score innerhalb der betrachteten Hypthesen Liste und n für die Länge dieser steht.

Dieses Verfahren verkürzt unsere Hypothesenliste natürlich um eine Stelle, denn für $s_i = min$ wird der Wert auf Null normiert. Da jedoch Hypothesenlisten der Länge zehn verwendet werden, wird hierdurch die Qualität nicht wesentlich beeinflusst.

4 Experimente

In diesem Kapitel wird zunächst der Aufbau der durchgeführten Experimente beschrieben, bevor im folgenden Abschnitt die im Rahmen dessen aufgezeichneten Daten erläutert werden. Im Anschluss daran erfolgt eine Beschreibung des Trainings und der durchgeführten Evaluationen zur Bestimmung der benötigten Parameter für die Konfidenzberechnung.

4.1 Aufbau der Experimente

Zum Zweck der Experimente wurden zwei Kennenlern-Dialog-Szenarien geschaffen [18] [14], wobei der Proband bei beiden lediglich die Anweisung erhielt sich in das Sichtfeld des Robortes zu stellen. Der Flurroboter[4] seinerseits startete den Dialog sobald er den Probanden mit seiner Kamera durch „tracking" [9] detektierte und verfolgen konnte oder eine Spracheingabe aufzeichnete. Ziel des Dialogs war es, möglichst viel Information über den Probanden zu sammeln, darunter einige einfache Nachfragen, die sich mit „Ja" oder „Nein" beantworten ließen, aber auch offene Fragen wie nach Vor- und Nachnamen. Das in dieser Arbeit vorgestellt System wurde bereits als Teil des Dialog-Systems eingesetzt, um diesem eine Hypothese (mit entsprechender Konfidenz) über die Identität des Gesprächspartners zur Verfügung zu stellen. Das Dialogmanagement-System konnte so vermeiden, bereits bekannte Antworten von Probanden erneut abzufragen.

4.2 Beschreibung der Daten

Die so erzeugten Daten bestehen aus einer Sammlung von Einzelbildern (8-15 Frames pro Sekunde) und kurzen Audiosegmenten, die mit Beginn des Dialogs mitgeloggt wurden. Die Bilder wurden dabei als Blocks von

JPEGs der Auflösung 640×480 Pixel mit 24-bit Farbtiefe und die Audiomitschnitte als WAV-Dateien gespeichert. Insgesamt wurden auf diese Weise 38 Probanden in 85 Sessions aufgezeichnet, wobei die Länge der jeweiligen Sessions abhängig vom Dialogablauf stark variieren, so dass keine allgemein gültige Aussage über die in einer Session aufgezeichneten Daten möglich ist. Im Durchschnitt wurden jedoch pro Session 1019 Einzelbilder mit 378 Gesichtsdetektionen (die Gesichtsdetektionsrate liegt damit bei 36,81%) und 16 Einzelturns, die insgeamt eine Länge von etwa 14 Sekunden hatten, aufgezeichnet. Für unser System bedeutet dies, dass eine derartige Durchschnittssession 378 Einzelbild- und 16 Concatturnhypothesen generiert.

4.3 Aufteilung der Daten in Sets

Da wir sowohl für das Training der FaceID und VoiceID als auch zur Ermittlung der diversen Regressionskoeffizienten separate Trainings- und Evaluationsdaten benötigen, ist der Bedarf unseres Systems an unabhängigen Datensets recht groß. Um die Anzahl an benötigen Sets dennoch möglichst klein zu halten, haben wir sie (siehe Abbildung 4.1) derart gestaffelt, dass die verwendeten Evaluationssets zwar einerseits unabhängig von den dazugehörigen Trainingssets sind, sie jedoch dennoch ein weiteres Mal wiederverwendet werden können und zwar fürs Training der nächst höheren Ebene.

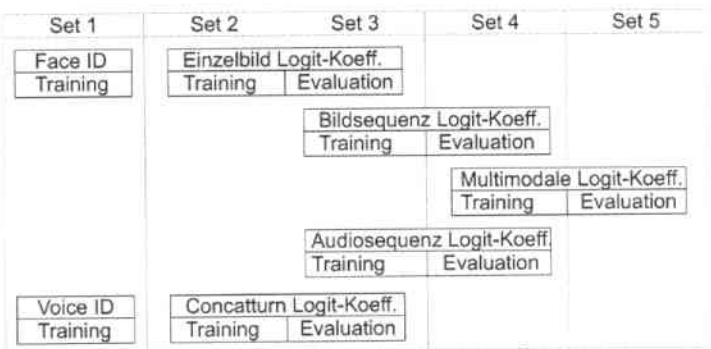

Abbildung 4.1: Verwendung der Sets als Trainings- und Evaluationssets

Insgesamt werden so fünf verschiedene Sets für das Training und die Evaluation der verschiedenen Logit-Koeffizienten benötigt. Die Tabelle 4.1 zeigt eine Übersicht und gibt dabei die Anzahl der Sessions sowie die der verschiedenen Probanden bzw. verwendeten Labels innerhalb der jeweiligen Sets an. Weiter führt die Tabelle für jedes Set die Erkennungsrate der FaceID auf Einzelbildern auf, wobei das entsprechende FaceID Modell auf Set1 trainiert. Das Set „Unknown" enthält, wie der Name schon vermuten lässt, ausschließlich Probanden die nicht im Set 1 repräsentiert und der FaceID folglich unbekannt sind. Da die Konfidenz auch die Wahrscheinlichkeit für Unbekannte berücksichtigen soll, wurden die 20 Sessions des Sets „Unknown(B)" auf die vier verbleibenden Sets verteilt, woraus Set 2B, 3B, 4B und 5B resultieren, welche wiederum Grundlage weiterer Analysen waren. Der Unterschied in den Erkennungsraten (hier exemplarisch anhand der FaceID abzulesen) zwischen den B Sets und den entsprechenden ursprünglichen Sets, ist also lediglich der Hinzunahme von „Unknown" geschuldet.

Set Übersicht			
	Sessions	Labels	Erkennungsrate der FaceID
Set 1	27	25	-
Set 2	16	15	92,04%
Set 3	16	14	93,72%
Set 4	17	16	84,26%
Set 5	16	14	90,85%
Unknown(B)	20	13	0%
Set 2B	21	15	60,16%
Set 3B	21	14	62,49%
Set 4B	22	16	63,23%
Set 5B	21	14	56,05%

Tabelle 4.1: Set Übersicht

4.4 FaceID

4.4.1 Training

Wie bereits in Kapitel 4.3 angemerkt, wurden die Modelle der FaceID auf dem Set1 trainiert. Dazu wurde zunächst für jedes Label eine zufällige Auswahl[1] an Bildern getroffen auf denen trainiert werden sollte. Dieser Auswahlschritt wurde eingeführt, da das Set1 über 30000 Bilder beinhaltete, und wir das FaceID Modell nicht unnötig aufblähen wollten. Anschließend wurde für jedes Bild, wie in Kapitel 2.1.1 beschrieben, nach Gesichtern und Augen durchsucht und ggf. ein Merkmalsvektor des Gesichts extrahiert. Dieser wurde dann zusammen mit dem Personenlabel als Trainingsinstanz in der Datenbasis hinterlegt (siehe Kap. 2.1.2 f.). Von den 6545 Trainingsbildern konnten auf diese Weise auf insgesamt 4283 Frames Merkmalsvektoren extrahiert werden.

4.4.2 Bestimmung der Einzelbildkonfidenz

Nachdem das Training der FaceID abgeschlossen wurde, bestand der nächste Trainingsschritt darin, Logit-Koeffizienten für die Konfidenz zu trainieren bzw. zunächst geeignete Konfidenzmaße auszuwählen. Dazu wurden zum einen für die Bilder in Set 2B und Set 3B Merkmalsvektoren generiert und anhand ihres Abstandes (siehe Kap. 2.1.4) zu den trainierten Merkmalsvektoren aus Set 1 Hypothesenlisten erstellt und zum anderen zu jeder Hypothesenliste alle relevanten Konfidenzmerkmale (Diff0, Diff1, Diff2, Ent, Image und Dist) gespeichert. Anschließend wurde für eine Vielzahl an Konfidenzmerkmalskombinationen Logit-Koeffizienten auf Set 2B trainiert und auf Set 3B evaluiert. Abbildung 4.2 zeigt einen Ausschnitt aus dem dazugehörigen ROC Graph [3].

Abgesehen von Entropy als Konfidenzmaß, dass mit einer True-Positiv-Rate von $0,71$ weit abgeschlagen wurde, sind alle anderen untersuchten Konfidenzmerkmale (bzw. deren Kombinationen) etwa im gleichen TP-/FP-Bereich[2], mit einer sehr niedrigen False Positiv-Rate (<0,1) und einer sehr guten True Positive-Rate (>0,8). An dieser Stelle sollte jedoch noch einmal gesondert darauf aufmerksam gemacht werden, dass sowohl

[1] nicht mehr als 280 pro Label

[2] TP: True Positiv, FP: False Positiv

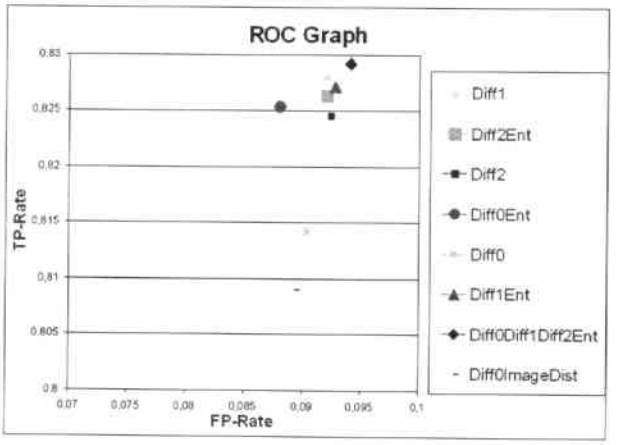

Abbildung 4.2: ROC Graph der verschiedenen Konfidenzmerkmale zur Vorhersage der Einzelbildkonfidenzen (FaceID) evaluiert auf Set3B

Diff0[3] als auch Diff0ImageDist[4] von allen weiteren von uns untersuchten Merkmalen bezüglich der TP-Rate übertroffenen wurden. Da die Prognose der Konfidenz eine möglichst große Zuverlässigkeit haben soll, ist für unsere Wahl eines geeigneten Konfidenzmerkmals (bzw. einer Kombination), bei etwa gleich starker Abweichung eine niedrige FP-Rate einer etwas höheren TP-Rate vorzuziehen, so dass wir uns schließlich für Diff0Ent entschieden haben. Diese Wahl erklärt sich insbesondere, wenn man sich den Vergleich zwischen der Evaluation auf Set 3 (ohne Unknown) und Set 3B (mit Unknown) ansieht, der in Abbildung 4.3 dargestellt ist.

Anhand der Graphik erkennt man, dass die Konfidenzmaße weitestgehend stabil gegenüber Erkennungsratenverschiebungen sind, jedoch mit Verbesserung der Erkennungsrate offensichtlich auch die FP-Rate der Konfidenzmaße etwas verbessert. „Diff0Ent" schneidet ohne Unknown nun zwar nicht mehr als „bestes" Maß ab, was aber umgekehrt bedeutet, dass es das stabilste Maß gegenüber der Hinzunahme von Unknown bzw. der Verschlechterung der Erkennungsrate ist, so dass sich die Entscheidung für dieses Konfidenzmaß hier nochmal bestätigt. Die auf Set 2B ermit-

[3] verwendet in [7]
[4] ähnlich dem in [6]

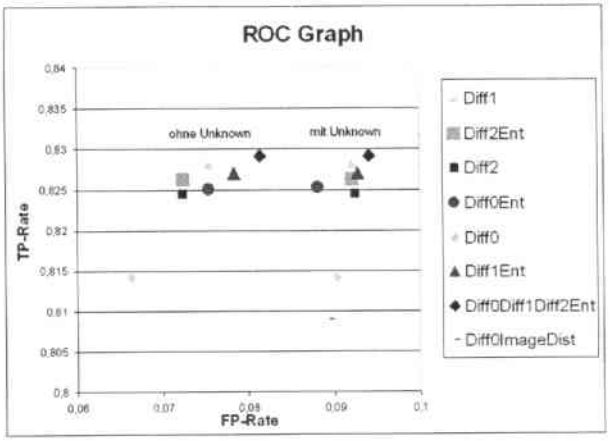

Abbildung 4.3: Einflusses von Unknown Labels auf Vorhersage der Einzelbildkonfidenzen (FaceID) dargestellt in ROC Graph

telten Logit-Koeffizienten für Diff0Ent werden in Tabelle 4.2 aufgeführt.

Logit-Koeffizienten				
Konfidenzmerkmal	i	Koeffizient β_i	⌀ Merkmalswert d_i	$\beta_i \cdot d_i$
	0	-2.2539		
Diff0	1	-18.2106	0,215	-3,9153
Ent	2	2.1548	2,513	5,4150

Tabelle 4.2: Logit-Koeffizienten für Konfidenzen der Einzelbilder

Abschließend zeigt Abbildung 4.4 die Konfidenz auf Basis dieser Logit-Koeffizienten im Verhältnis zur Erkennungsrate. Dazu wurden die Hypothesenlisten anhand ihrer Konfidenzen in Bins eingeteilt, wobei die Bins eine Konfidenzspanne von 10% (die äußeren beiden 5%) haben, und die durschnittliche Erkennungsrate innerhalb der Bins ermittelt.

Insgesamt wird die durchschnittliche Erkennungsrate innerhalb der Bins durch die Konfidenz etwas unterschätzt. Das bedeutet, dass die Konfidenz etwas restriktiver ist als sie sein müsste, jedoch umgekehrt sicherstellt, dass die Konfidenz die Erkennungsrate nicht überschätzt, was ganz in unserem Sinne ist.

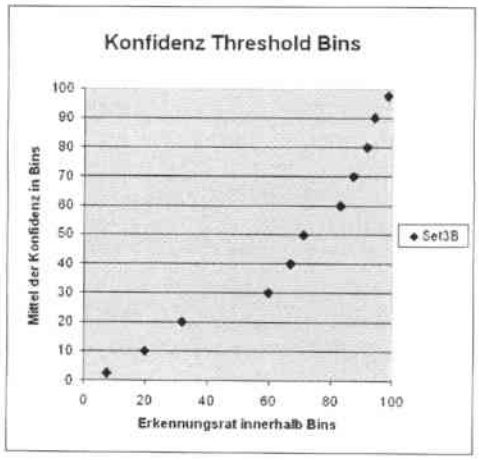

Abbildung 4.4: Konfidenz der Einzelbildhypothesen im Verhältnis zur Erkennungsrate

4.4.3 Bestimmung der Bildsequenzkonfidenz

Die Bildsequenzhypothese ist eine Fusion der Einzelbildhypothesen. Ihre Länge gibt dabei die maximale Anzahl an berücksichtigen Einzelbildhypothesen an. Die erste Bildsequenzhypothese wird jedoch, um im Live-System möglichst schnell Ergebnisse liefern zu können, bereits aus den ersten zwei Einzelbildhypothesen gebildet, um sie dann schrittweise um eine weitere Einzelbildhypothese zu erweitern bis sie schließlich die maximale Sequenzlänge erreicht. Ab diesem Zeitpunkt wird die Sequenz mit einer festen Breite wie ein Sliding Window über die Einzelbildhypothesen geschoben. Auf Grundlage der Einzelbildhypothesen und den dazu gehörigen Konfidenzen werden dann gemäß Formel 3.1 Bildsequenzhypothesen berechnet. Um anschließend auch für diese Hypothesen Konfidenzen bereitstellen zu können, müssten erneut geeignete Konfidenzmaße ausgewählt und entsprechenden Logit-Koeffizienten trainiert werden. Zu diesem Zweck wurden für Set 3B und Set 4B Bildsequenzhypothesen mit unterschiedlicher Sequenzlänge berechnet und mit relevanten Konfidenzmerkmalen (Agre, Stab, Diff0, Diff1, Diff2, Ent) gespeichert. Abbildung 4.5 zeigt den ROC Graph für die Sequenzlänge 15, wobei die

Logit-Koeffizienten auf Set 3B trainiert und auf Set 4B evaluiert wurden.

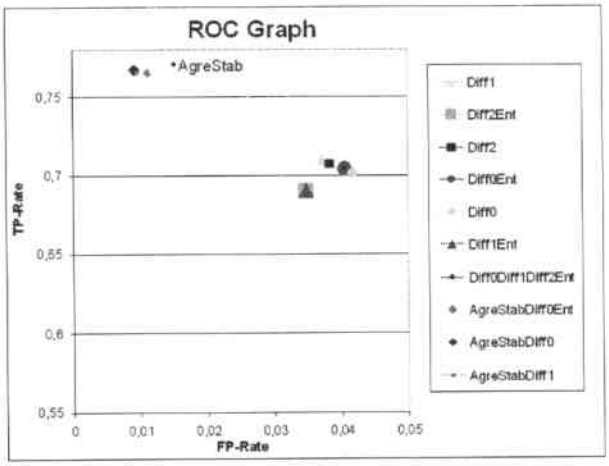

Abbildung 4.5: ROC Graph der verschiedenen Konfidenzmerkmale zur Vorhersage der Bildsequenzkonfidenzen (FaceID) evaluiert auf Set4B

Anhand des ROC Graph erkennt man, dass die Merkmale mit Berücksichtigung von Agre und Stab deutlich besser abscheiden als jene, die allein auf der Struktur der Hypothesenlisten basieren. Bevor wir uns jedoch für eine endgültige Konfidenzmerkmalskombination entschieden haben, wurden diese über verschiedenen Sequenzlängen hinweg betrachtet, da mit Erhöhung der Sequenzlänge auch die Erkennungsrate steigt (siehe Abbildung 4.8), dies jedoch nicht unbedingt für die Zuverlässigkeit der verwendeten Konfidenzmerkmale gelten muss.

Abbildung 4.6 zeigt den Leistungsverlauf von vier möglichen Konfidenzmerkmalskombinationen bei den Sequenzlängen 4, 15, 50, 100 und 200. Hierbei lässt sich erkennen, dass die Merkmale Agre und Stab zwar recht gute Konfidenzmaße darstellen, jedoch bei längeren Sequenzlängen stärke TP-Rate Einbußen hinnehmen als, wenn sie durch weitere Merkmale wie Diff0 und Diff1 verstärkt werden. Wir haben uns so schließlich für die Kombination „AgreStabDiff0Ent" entschieden, da sie von den Untersuchten Konfidenzmerkmalen die beste Stabilität gegenüber Sequenzlängenveränderung zeigte.

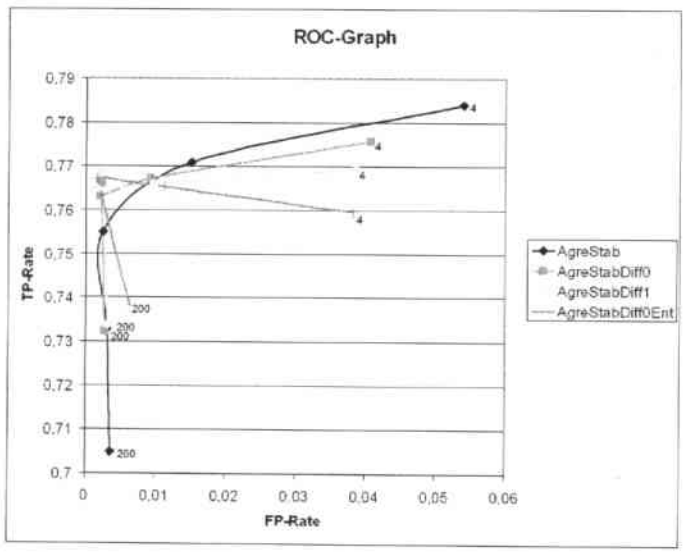

Abbildung 4.6: Einfluss der Sequenzlänge auf Vorhersage der Bildsequenzkonfidenzen (FaceID) dargestellt in ROC Graph

Die auf Set 3B ermittelten Logit-Koeffizienten für AgreStabDiff0Ent mit einer Sequenzlänge von 200 lauten werden in Tabelle 4.3 aufgeführt.

Logit-Koeffizienten				
Konfidenzmerkmal	i	Koeffizient β_i	⌀ Merkmalswert d_i	$\beta_i \cdot d_i$
	0	-9.4296		
Agre	1	-3.1466	0,718	-2,2593
Stab	2	-6.0574	0,74	-4,4825
Diff0	3	2.9128	0,343	0,9991
Ent	4	6.0418	2,219	13,4068

Tabelle 4.3: Logit-Koeffizienten für Konfidenzen der Bildsequenzen

Abbildung 4.7 zeigt - wie schon im vorherigen Kapitel - das Verhältnis zwischen Konfidenz und Erkennungsrate. Auch hier wird die tatsächliche Erkennungsrate durch die Konfidenz unterschätzt, die Tendenz ist jedoch nach wie vor gut erkennbar.

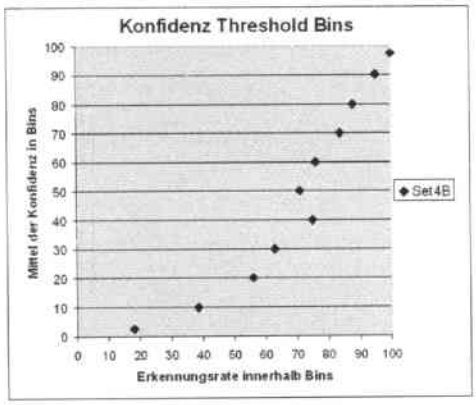

Abbildung 4.7: Konfidenz der Bildsequenzhypothesen (FaceID) im Verhältnis zur Erkennungsrate

Abschließend zeigt Grafik 4.8 für alle vier Sets die Erkennungsraten der Einzelbildhypothesen verglichen zu den Bildsequenzhypothesen verschiedener Länge. Hier zeigt sich besonders deutlich, welchen Einfluss die Länge der Sequenz auf die Erkennungsrate hat. Darüber hinaus erkennt man anhand des Verlaufs bei Set 4, dass Einzelbildfehler der FaceID durch eine stärkere Steigung der Erkennungsrate der Bildsequenzen kompensiert werden können.

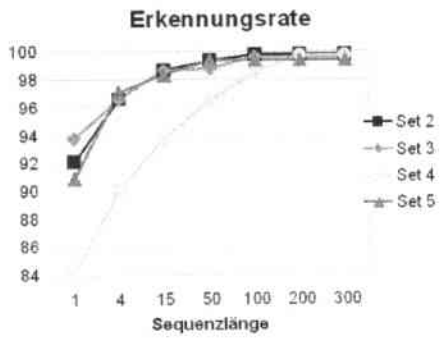

Abbildung 4.8: Erkennungsrate der Bildsequenzhypothesen (FaceID) auf den vier verschiedenen Sets bei unterschiedlicher Sequenzlänge

4.5 VoiceID

Im Unterschied zur FaceID wurde die VoiceID zunächst auf einem separatem Datensatz evaluiert. Die für die folgenden Konfidenzberechnungen und Evaluationen verwendeten Daten wurden also nicht wie in 4.1 beschrieben im Rahmen des Flurroboter-Dialogs aufgezeichnet, sondern stattdessen handelt es sich hierbei um unabhängig von einem Dialog aufgezeichnete freigesprochene Sprache. Auch dieser Datensatz wurde, aus Gründen die bereits in Kapitel 4.3 ausgeführt wurden, in 5 separate Sets unterteilt. Die D Sets unterschieden sich von den C Sets nur dahingehend, dass auf sie die 13 Sessions des Sets „Unknown(D)" verteilt wurden, was die Erkennungsrate natürlich entsprechend beeinflusst.

Set Übersicht - für zweiten Datensatz			
	Sessions	Labels	Erkennungsrate der VoiceID
Set 1C	8	8	-
Set 2C	7	5	72,31%
Set 3C	8	5	94,05%
Set 4C	8	5	95,10%
Set 5C	12	4	91,54%
Unknown(D)	13	3	0%
Set 2D	10	5	54,65%
Set 3D	11	5	75,24%
Set 4D	10	5	79,43%
Set 5D	17	4	67,54%
Sets 2D-5D	48	8	63,61%

Tabelle 4.4: Set Übersicht - für zweiten Datensatz

4.5.1 Training

Die VoiceID wurde auf Grundlage von Set 1C trainiert. Dazu wurde für jedes der acht zu trainierenden Labels zehn Sekunden Audiomitschnitt zugrunde gelegt und die in Kapitel 2.2 beschriebenen Schritte durchgeführt, um schließlich für jedes Label ein entsprechendes GMM trainieren zu können.

4.5.2 Bestimmung der Concatturnkonfidenz

Wie schon bei der FaceID bestand der nächste Schritt darin, passende Logit-Koeffizienten für die Konfidenzberechnung zu trainieren bzw. zunächst passende Konfidenzmerkmale auszuwählen. Zu diesem Zweck wurden für die Sets 2D und 3D VoiceID Hypothesenlisten für alle Concatturns generiert und mit ihnen alle zur Verfügung stehenden Konfidenzmerkmale (Diff0, Diff1, Diff2, Ent) gespeichert. Anschließend wurde - analog zu dem Vorgehen bei der FaceID - für einige Kombinationen von Konfidenzmerkmalen Logit-Koeffizienten auf Set2D trainiert und auf Set3D evaluiert. Abbildung 4.9 zeigt einen Ausschnitt aus dem dazugehörigen ROC Graph.

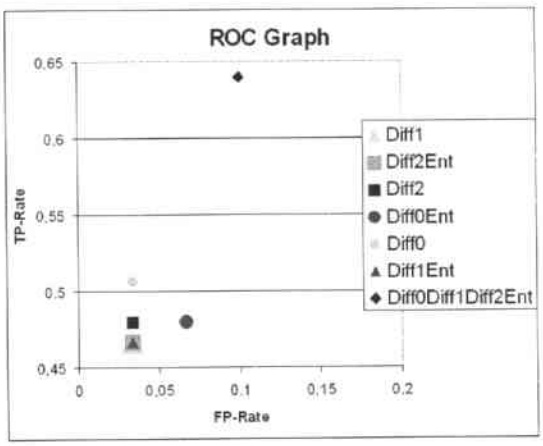

Abbildung 4.9: ROC Graph der verschiedenen Konfidenzmerkmale zur Vorhersage der Concatturnkonfidenzen (VoiceID) evaluiert auf Set3D

Auch hier lässt sich, wie bereits bei der FaceID beobachten, dass die FP-Rate aller untersuchten Konfidenzmerkmale recht gut ist (< 0,1 bzw. sogar < 0,05), die TP-Rate hingegen, verglichen zu den Ergebnissen bei der FaceID, etwas schwächer sind. Die Gründe hierfür sind in der Struktur bzw. Generierung der n-besten Listen zu suchen, die bei der VoiceID auf Grundlage der GMMs erzeugt werden und somit gänzlich anderen Ursprungs sind, als die der FaceID. Dies wird insbesondere deutlich, wenn

man die beiden ⊘ Merkmalswerte d_i für „Diff0" in den Tabellen 4.2 und 4.5 vergleicht.

Nach dieser ersten Evaluation waren die beiden in Frage kommenden Konfidenzmerkmale „Diff0" und „Diff0Diff1Diff2Ent". Bei einer Evaluation auf Set 4D zeigte sich jedoch für letztere eine erhebliche höhere Abweichung bezüglich der FP-Rate und da darüberhinaus „Diff1" eine statistische Abhängigkeit zu „Diff2" aufweist, haben wir uns schließlich für „Diff0" als Konfidenzmaß entschieden. Die auf Set 2D ermittelten Logit-Koeffizienten für Diff0 werden in Tabelle 4.5 aufgeführt.

Logit-Koeffizienten				
Konfidenzmerkmal	i	Koeffizient β_i	⊘ Merkmalswert d_i	$\beta_i \cdot d_i$
	0	1.0373		
Diff0	1	-11.3955	0,088	-1,0028

Tabelle 4.5: Logit-Koeffizienten für Konfidenzen der Concatturns

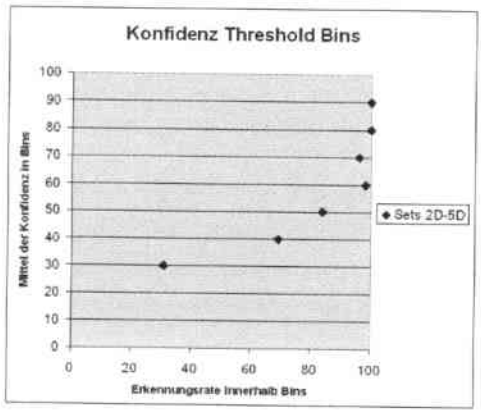

Abbildung 4.10: Konfidenz der Concatturnhypothesen im Verhältnis zur Erkennungsrate

Abbildung 4.10 zeigt das Verhältnis zwischen Konfidenz und Erkennungsrate, für die Konfidenzen der Concatturnhypothesen. Da die Anzahl der Einzel- bzw. Concatturns innerhalb der Sets deutlich kleiner ist, als die der Einzelbilder, wurde die Graphik auf Basis des gesamten Datensatzes

bzw. der Übermenge der Sets 2D, 3D, 4D und 5D berechnet, anstatt nur auf Basis des Sets 3D.

Auch in diesem Fall unterschätzt das von uns ausgewählte Konfidenzmaß die tatsächliche Erkennungsrate. Dies gilt insbesondere für den oberen Konfidenzbereich, was mitunter auch an der geringen Datendichte dieser Bins liegen mag. Deutlich auffälliger ist jedoch, dass die Konfidenz nicht unter 25% fällt. Das ist ein Indiz dafür, dass „Diff0", trotz der guten FP-Rate, als alleiniges Konfidenzmaß nicht ausreicht, um „Unknown" bzw. Fehlklassifikationen anhand der Konfidenz sicher vorherzusagen.

Im Anschluss zu den Evaluationen auf dem separaten Datensatz waren weitere Evaluationen auf den in Kapitel 4.2 beschriebenen Daten geplant. Dabei erreichte die VoiceID jedoch nur eine Erkennungsrate von 32,77% (ohne „Unknown") was sich entsprechend negativ auf die Ergebnisse der Audiosequenzhypothesen und schließlich die multimodale Fusion ausübte. Daher waren keine aussagekräftigen Evaluationsergebnisse entstanden und in folge dessen konnte auch für die multimodale Fusion keine aussagekräftigen Ergebnisse erzielt werden. Wir vermuten die Ursache für die schlechte Erkennungsrate in einem zu geringen Anteil an natürlichsprachlichen Äußerungen im Rahmen der aufgezeichneten Dialoge. Vermutlich bräuchte man wesentlich mehr Trainingsmaterial, um mit Hilfe der VoiceID die Sprecher allein anhand ihrer „Ja" und „Nein" Äußerungen oder ihrer Buchstabierung differenzieren zu können als wir zur Verfügung hatten.

4.5.3 Bestimmung der Audiosequenzkonfidenz

Die Audiosequenzhypothese ist - analog dem Vorgehen bei der Bildsequenzhypothese - eine Fusion der Concatturnhypothesen. Um auch für diese Hypothesen Konfidenzen bereitstellen zu können, müssten erneut geeignete Konfidenzmaße ausgewählt und entsprechenden Logit-Koeffizietnen trainiert werden. Aus diesem Grunde wurden für Set 3D und Set 4D Audiosequenzhypothesen berechnet und mit relevanten Konfidenzmerkmalen (Stab, Agre, Diff0, Diff1, Diff2, Ent) gespeichert. Die Sequenzlänge wurde dabei mit 50 weit über der maximalen Anzahl an Concatturnhypothesen pro Session festgelegt, um so die Länge der jeweiligen

Sequenz möglichst groß zu halten. Abbildung 4.11 zeigt den entsprechenden ROC Graph.

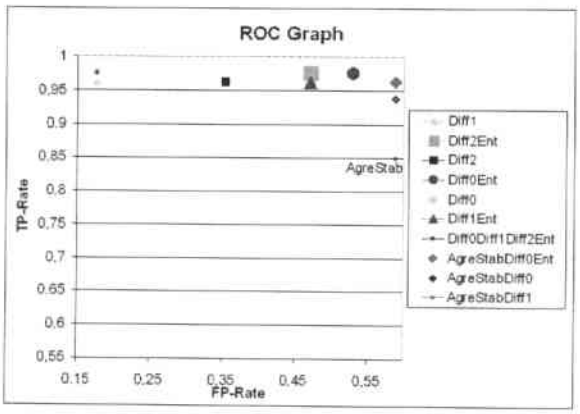

Abbildung 4.11: ROC Graph der verschiedenen Konfidenzmerkmale zur Vorhersage der Audiosequenzkonfidenz (VoiceID) evaluiert auf Set4D

Anhand des ROC Graph erkennt man, dass die Merkmale mit Berücksichtigung von „Agre" und „Stab" im Gegensatz zu den Beobachtungen bei Bildsequenzen deutlich schlechter abschneiden. Die Ursache hierfür ist dem Wesen der VoiceID bzw. der GMMs und ihrer Funktionsweise geschuldet. Während die FaceID Datenpunkte von „Unknowns" im Vektorraum meist fernab trainierter Labels abbildet und bei der Berechnung nächster Nachbarn folglich zu häufig wechselnden Hypothesen neigt, entscheidet sich die VoiceID bei einer unbekannten Verteilung für dasjenige GMM, dass diese am besten abbildet, was bei verschiedenen Aufnahmen des selben unbekannten Sprecher offensichtlich immer wieder zu gleichen Hypothesen führt. „Agre" und „Stab" sind folglich für die Sequenz von VoiceID Hypothesen keine geeigneten Konfidenzmaße. Wesentlich vielversprechender erscheinen hingegen die Konfidenzmaße „Diff0", „Diff1" und „Diff0Diff1Diff2Ent". Wobei sich auch hier in weiteren Evaluationen „Diff0" als das stabilste Maß herausstellte, weswegen wir uns schließlich für dieses entschieden. Die auf Set3D ermittelten Logit-Koeffizienten für „Diff0" werden in Tabelle 4.6 aufgeführt.

Logit-Koeffizienten				
Konfidenzmerkmal	i	Koeffizient β_i	⌀ Merkmalswert d_i	$\beta_i \cdot d_i$
	0	2.4942		
Diff0	1	-31.9234	0,11	-3,511574

Tabelle 4.6: Logit-Koeffizienten für Konfidenzen der Audiosequenzen

Abbildung 4.12 zeigt - wie schon im vorangegangenem Kapitel - das Verhältnis zwischen Konfidenz und Erkennungsrate auf Basis des gesamten Datensatzes bzw. der Übermenge der Sets 2D, 3D, 4D und 5D. Es sind zwar einige Abweichungen in einzelnen Bins zu verzeichnen, was erneut der geringen Datendichte dieser Bins geschuldet sein mag, aber nichtsdestotrotz ist die Tendenz nach wie vor recht gut erkennbar.

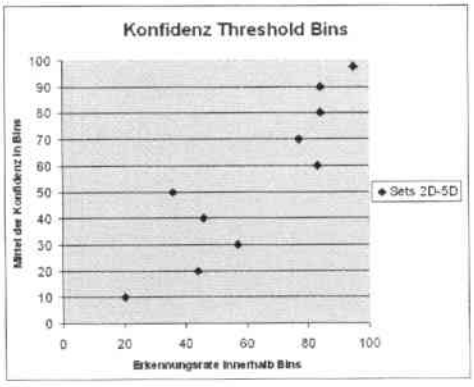

Abbildung 4.12: Konfidenz der Audiosequenzhypothesen (VoiceID) im Verhältnis zur Erkennungsrate

Vergleicht man auf Basis des Datensatzes (Sets 2D-5D) die Erkennungsraten der Audiosequenzehypothesen mit denen der Concatturnhypothesen so zeigt sich nur eine minimale Verbesserung von 63,61% (VoiceID mit Unknown) auf 64,13%. Außerdem geht die Verbesserung der TP-Rate bei den Audiosequenzkonfidenzen mit einer Verdreifachung der FP-Rate gegenüber den Concatturnkonfidenzen einher. Angesichts dessen stellt sich natürlich die Frage, ob der Einsatz von Sequenzen bei der VoiceID überhaupt gerechtfertigt ist. Aus diesem Grund haben wir bei

den weiteren Evaluation beide Varianten berücksichtigt, wobei sich (siehe 4.6.1) zeigte, dass der Einsatz der Audiosequenzen sowohl bezüglich der Erkennungsrate als auch bezüglich der Konfidenzberechnung der multimodalen ID leicht zum Vorteil gereicht.

4.6 Multimodale ID

4.6.1 Bestimmung der Fusionskonfidenz

Nachdem wir nun für beide Modalitäten Hypothesenlisten als auch dazugehörige Konfidenzen berechnen können, bleibt nur noch die Fusion dieser beiden. Um im Live-System möglichst auch dann eine Hypothese liefern zu können, wenn nur eine der beiden Modalitäten Hypothesen liefert, gibt die multimodale ID für diese Sonderfälle eine Kopie der Sequenzhypothesen der zur Verfügung stehenden Modalität wieder. Im Regelfall wird die multimodale ID jedoch durch eine Fusion der Hypothesenlisten der beiden Modalitäten gemäß der Beschreibung in Kapitel 3.3 gebildet. Da wir anschließend auch für diese nun multimodalen Hypothesenlisten Konfidenzen anbieten wollen, müssen erneut geeignete Konfidenzmerkmale ausgewählt werden. Wir haben zu diesem Zweck jeder Session der Sets 4D und 5D passende Videoaufzeichnungen aus den Sets 2B bis 5B zugeordnet und multimodale Hypothesen für sie berechnet, die dann zusammen mit allen relevanten Konfidenzmerkmalen (Stab, Agre, Diff0, Diff1, Diff2, Ent) gespeichert wurden. Anschließend wurden für verschiedene Konfidenzmerkmalskombinationen Logit-Koeffizienten auf Set 4D trainiert und mit Set 5D evaluiert. Da die Ergebnisse der Audiosequenzen keine klare Verbesserung der Erkennungsrate gegenüber der Concatturnhypothesen gezeigt hat, haben wir im folgenden zwei Varianten zur Berechnung der multimodalen ID untersucht. Bei der ersten Variante entfällt die Audiosequenz-Ebene und die multimodale Hypothese wird aus der Fusion von Bildsequenz- und Concatturnhypothesen gewonnen, während bei der zweiten Variante die fusionierte Hypothesenliste, wie in Abbildung 3.1 dargestellt, auf Basis der Bildsequenz- und Audiosequenzhypothesen erstellt wird. Abbildung 4.13 zeigt die entsprechenden ROC Graphen. ROC Graph 1 zeigt die Leistung verschiedene Kon-

fidenzmerkmale der ersten Variante, während ROC Graph 2 selbiges für die zweite Variante zeigt.

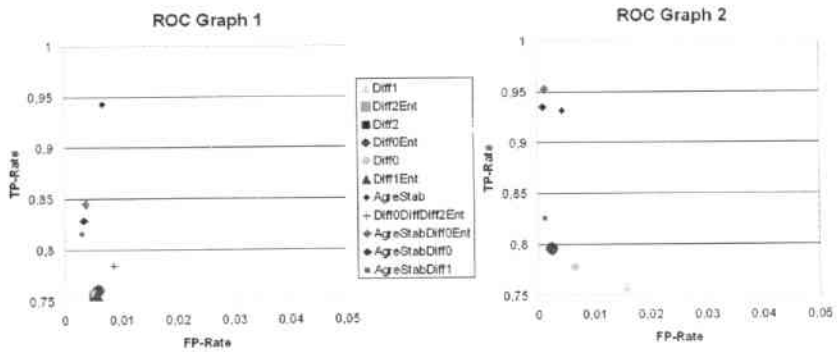

Abbildung 4.13: ROC Graphen der verschiedenen Konfidenzmerkmale für die multimodale ID evaluiert auf Set5D.

Vergleicht man nun diese beiden ROC Graphen wird deutlich, dass die Konfidenzmaße „AgreStabDiff0Ent" und „AgreStabDiff0" bei der zweiten Variante wesentlich besser als bei der ersten Variante abscheiden und sogar beide „AgreStab" Ergebnisse übertreffen. Die Ursache diese Verbesserung ist, dass die Scores der Sequenzhypothesenlisten in der Regel eine größere Varianz als die der Concatturnhypothesen aufweisen. Dadurch sind die Konfidenzmaße „Diff0" und „Diff0Ent" für die Sequenzhypothesen etwas aussagekräftiger. Da die zweite Variante darüber hinaus

Logit-Koeffizienten				
Konfidenzmerkmal	i	Koeffizient β_i	⌀ Merkmalswert d_i	$\beta_i \cdot d_i$
	0	24.1075		
Agre	1	-4.3436	0,893	-3,8788
Stab	2	-4.3436	0,893	-3,8788
Diff0	3	-43.6945	0,376	-16,4291
Ent	4	-4.4368	2,145	-9,516936

Tabelle 4.7: Logit-Koeffizienten für Konfidenzen der multimodalen Hypothese

auch bezüglich der Erkennungsrate nicht schlechter als die erste Variante

abschneidet, haben wir uns für sie in Kombination zu „AgreStabDiff0Ent" als Konfidenzmaß entschieden.Tabelle 4.7 führt entsprechend die auf Set 4D ermittelten Logit-Koeffizienten für „AgreStabDiff0Ent" auf.

Abschließend zeigt Abbildung 4.14 - wie schon in den vorangegangenen - das Verhältnis der Erkennungsrate zur Konfidenz. Dabei mag der Eindruck entstehen, die Konfidenz würde die tatsächliche Erkennungsrate nicht sonderlich gut wiedergeben. Es sei jedoch darauf hingewiesen, dass die beiden gut geschätzten äußeren Bins mehrere Tausend Hypothesen beinhalten, wohin gegen die schlecht geschätzten mittleren Bins teilweise unter hundert fassen, so dass die Extrema bei der Bewertung deutlich stärker ins Gewicht fallen.

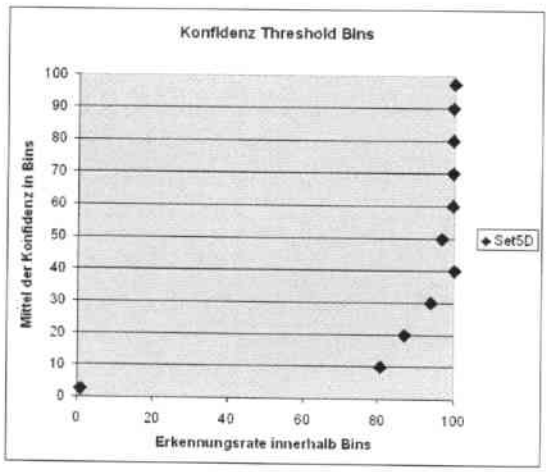

Abbildung 4.14: Konfidenz der multimodalen Hypothesen (multimodale ID) im Verhältnis zur Erkennungsrate

5 Diskussion und Ausblick

5.1 Zusammenfassung und Diskussion der Ergebnisse

Im Folgenden werden die Ergebnisse der Gesamt- und Teilsysteme noch einmal zusammengefasst und diskutiert, wobei die Evaluation auf Set 5C die Grundlage dafür stellt. In Abbildung 5.1 werden zu diesem Zweck die einzelnen Erkennungsraten gegenüber gestellt.

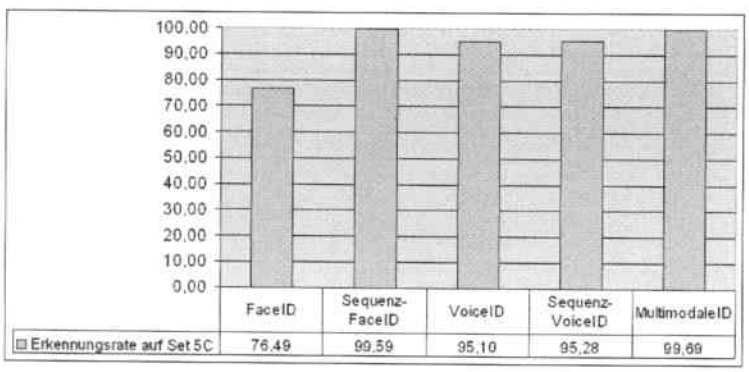

Abbildung 5.1: Erkennungsraten auf Set 5C im Überblick

Dabei bestätigt sich nochmal die Beobachtung von Abbildung 4.8, dass Einzelbildfehler der FaceID durch eine stärkere Steigung der Erkennungsrate der Bildsequenzen kompensiert werden können, so dass die Sequenz-FaceID schließlich sogar eine höhere Erkennungsrate als die, nur auf 8 Personen trainierte, VoiceID bzw. die Sequenz-VoiceID besitzt. Die Verbesserung der Erkennungsrate zwischen VoiceID und Sequenz-VoiceID ist hingegen, wie bereits in Kapitel 4.5.3 angesprochen, nicht so deutlich ausgeprägt. Dies liegt unter anderem daran, dass die auf Basis der

GMMs gewonnenen VoiceID Hypothesen dazu tendieren auf ähnlichen Daten ähnliche Hypothesen zu generieren, auch wenn diese Hypothesen fehlerhaft sein sollten, wohingegen die fehlerhaften Hypothesen der FaceID auf Basis des Nearest-Neighbour-Ansatzes in sich sehr stark variieren, was die Sequenz Hypothese begünstigt. Auch die multimoldale ID zeigt eine Verbesserung der Erkennungsrate, aber auf Grund des sehr hohen Niveaus der Sequenz-FaceID mit einer Erkennungsrate von 99,59% ist die Verbesserung trotz einer relativen Steigerung von 24,3% absolut betrachtet nur 0,1%. Beide Schritte erzielen jedoch darüber hinaus auch eine Verbesserung der Konfidenzen, die in Tabelle 5.1 aufgeführt werden,

FaceID			
	Anzahl	Mittelwert	Standardabweichung
Set 5C Hypothese korrekt	4487	0,583	0,362
Set 5C Hypothese falsch	1379	0,119	0,171
Set Unknown(D)	3501	0,112	0,154
Sequenz-FaceID			
	Anzahl	Mittelwert	Standardabweichung
Set 5C Hypothese korrekt	5830	0,574	0,389
Set 5C Hypothese falsch	24	0,046	0,102
Set Unknown(D)	3488	0,019	0,046
VoiceID			
	Anzahl	Mittelwert	Standardabweichung
Set 5C Hypothese korrekt	233	0,447	0,145
Set 5C Hypothese falsch	12	0,315	0,074
Set Unknown(D)	116	0,315	0,048
Sequenz-VoiceID			
	Anzahl	Mittelwert	Standardabweichung
Set 5C Hypothese korrekt	222	0,668	0,293
Set 5C Hypothese falsch	11	0,188	0,201
Set Unknown(D)	103	0,262	0,246
multimodale ID			
	Anzahl	Mittelwert	Standardabweichung
Set 5C Hypothese korrekt	6068	0,911	0,205
Set 5C Hypothese falsch	19	0,43	0,465
Set Unknown(D)	3591	0,007	0,048

Tabelle 5.1: Übersicht über Konfidenzen für Sets 5C und Unknown(D)

was für sich gesehen bereits eine Rechtfertigung wäre, da eine Verbesserung der Konfidenz auch mit eine Verbesserung der Erkennungsrate auf der nächst höher liegenden Ebene einhergeht bzw. im Fall der multimodalen ID die Zuverlässigkeit des Klassifikators verbessert. Insbesondere die Konfidenzen der multimodalen ID sind mit einem Mittelwert von 0,911 für korrekt klassifizierte Hypothesen und 0,007 für „Unknown" also fehlklassifzierte Hypothesen ausgesprochen gut. Die Konfidenzabschätzungen ermöglichen so neben der reinen Aussage über die Zuverlässigkeit der Hypothesen auch die Detektion von „Unknown" und erweitert den Klassifikator somit nicht unerheblich.

Eine der größten Herausforderungen während der Arbeit war es eine gleichmässige Unterteilung der Daten in Sets zu gewährleisten, die sowohl die Anzahl an beinhalteten Labels, den Umfang der Video- und Audiodaten als auch die Häufigkeit der fehlerhaften Hypothesen und „Unknown" Hypothesen für beide Modalitäten umfasste. Dies war insbesondere deswegen schwierig, da die zugrunde liegenden Sessions nicht nur in ihrem Umfang absolut produzierter Hypothesen, sondern auch in der Häufigkeit von Fehlhypothesen in beiden Modalitäten unabhängig von einander stark variierten. Eine weitere Schwierigkeit während der Evaluationen entstand durch die Abhängigkeiten der verschiedenen Ebenen des Klassifikators, denn einer Änderung oder Anpassung auf unterer Ebene erzwang nicht nur eine erneute Evaluation aller beteiligen Konfidenzmerkmalskombinationen auf dieser, sondern auch auf jeder darüber liegenden Ebene, was Anpassungen des Systems teilweise sehr zeitaufwendig machte und somit das Experimentieren mit unterschiedlichen Parametern oder Verfahren (beispielsweise unterschiedliche Normierungsverfahren) sehr einschränkte.

5.2 Ausblick

Für die Zukunft sind zunächst weitere Experimente und Datensammlungen unter unterschiedlichen Bedingungen geplant, welche die Richtigkeit der vorgestellten Ergebnisse noch einmal bestätigen sollen. Außerdem konnten wir bereits bei unseren bisherigen Aufnahmen beobachten, dass im Live-System gelegentlich fehlerhafte UserIDs bestätigt bzw. bereits fehlerhaft eintrainiert wurden, so dass wir Grund zur Annahme haben, dass derartige Effekte die Erkennungsrate des Systems auf Dauer nachhaltig negative beeinflussen. Wir überlegen daher in Zukunft auf Basis der uns zur Verfügung stehenden Konfidenzen eine Bereinigung der Daten bzw. ein Vergessen des Systems zu integrieren, so dass derart fehlerhaft eintrainierte Daten zusehens „verloren" gehen würden. Da die von uns untersuchten Konfidenzmerkmale bei den gegebenen Daten auf den Hypothesen Listen der FaceID bessere Ergebnisse erzielt haben als auf denen der VoiceID, würden wir in der Zukunft gerne genauer evaluieren, wie es zu diesen Abweichungen kommt, und ob eventuell weitere VoiceID interne Konfidenzmerkmale nötig sind. Des weiteren würden wir gerne untersuchen inwieweit sich unser Ansatz zur Detektion von „Unknown" mit anderen explizit nach „Unknown" klassifizierenden Ansätzen [17] kombinieren lässt.

6 Zusammenfassung

Im Verlauf dieser Studienarbeit wurde ein multimodales Identifikationsmodul für einen mobilen Roboter entwickelt, welches während der Mensch-Roboter-Interaktion Hypothesen über die beteiligte Person generiert. Zur Erstellung dieser Annahmen werden nicht nur einzelne Bilder oder Audiomittschnitte in Betracht gezogen, sonderen ganze Sequenzen dieser betrachtet und die daraus resultierenden Hypothesen fusioniert. Das in dieser Arbeit vorgestellte Verfahren zur Kombination der Einzelklassifikationsergebnisse zur nächst höheren Ebene basiert dabei auf einer individuellen Gewichtung durch Konfidenzfaktoren. Die Konfidenzen werden jeweils mit Hilfe eines logistischen Regressionsmodells bestimmt. Diese Vorgehensweise erweist sich als sinnvoll, da die durchgeführte Evaluation der multimodalen ID signifikant bessere Klassifikationsergebnisse als die der zugrunde liegenden Einzelklassifikatoren liefert.

Darüberhinaus berechnet das entwickelte Modul Konfidenzen für die aufgestellten multimodalen Hypothesen, die erneut mit Hilfe der logistischen Regression bestimmt werden. Die Evaluation zeigt, dass die Konfidenz der multimodalen Hypothesen ein verlässliches Maß für die Korrektheit dieser Annahmen darstellt. Andere Komponenten auf dem Robotersystem, die mit dem multimodalen Identifikationsmodul über den one4all-Server kommunizieren, werden dadurch in die Lage versetzt, die generierten Hypothesen zu verwerfen oder zu übernehmen.

Ferner ist das Modul mit Hilfe der Konfidenz in der Lage „Unknown" als solche auszuweisen und mit Unterstützung des auf dem Roboter vorhandenen Dialogsystems diese als neue Benutzer in die bestehende Datenbank einzufügen.

Literaturverzeichnis

[1] EKENEL, HAZIM KEMAL, MIKA FISCHER, QIN JIN und RAINER STIEFELHAGEN: *Multi-modal Person Identification in a Smart Environment*. CVPR Biometrics Workshop, Minneapolis, USA, June 2007.

[2] EKENEL, HAZIM KEMAL und QIN JIN: *ISL Person Identification Systems in the CLEAR Evaluations*. CLEAR Evaluation Workshop, Southampton, UK, April 2006.

[3] FAWCETT, TOM: *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. Technischer Bericht, Tech Report HPL-2003-4, HP Laboratories., 2003.

[4] HOLZAPFEL, HARTWIG, THOMAS SCHAAF, HAZIM KEMAL EKENEL, CHRISTOPH SCHAA und ALEX WAIBEL: *A Robot learns to know people - First Contacts of a Robot*. KI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science, 4314, 2007.

[5] JIN, QIN: *Robust Speaker Recognition*. Doktorarbeit, Carnegie Mellon University, 2007.

[6] KÖNN, STEPHAN: *Studienarbeit: Gesichteridentifikation auf Bildsequenzen in Mensch-Roboter-Interaktion*. Universität Karlsruhe (TH), 2006.

[7] KÖNN, STEPHAN, HARTWIG HOLZAPFEL, HAZIM KEMAL EKENEL und ALEX WAIBEL: *Integrating Face-ID into an Interactive Person-ID Learning System*. International Conference on Computer Vision Systems (ICVS'07), Bielefeld, Germany, 2007.

[8] MITCHELL, TOM M.: *Machine Learning*. McGraw-Hill, 2. Auflage, 1997.

[9] NICKEL, K. und R. STIEFELHAGEN: *Fast Audio-Visual Multi-Person Tracking for a Humanoid Stereo Camera Head*. IEEE-RAS Intl. Conference on Humanoid Robots, 2007.

[10] N.N.: *http://de.wikipedia.org/wiki/Diskrete_Kosinustransformation*. Stand: März 2008.

[11] N.N.: *http://www.cs.cf.ac.uk/Dave/Multimedia/node238.html*. Stand: März 2008.

[12] N.N.: *http://www.lrz-muenchen.de/~wlm/ilm_l11.htm*. Stand: März 2008.

[13] PELECANOS, JASON und SRIDHA SRIDHARAN: *Feature Warping for Robust Speaker Verification*. Proceedings of Speaker Odyssey Conference, 2001.

[14] PUTZE, FELIX: *Social User Model Acquisition through Network Analysis and Interactive Learning*. Diplomarbeit, Universität Karlsruhe (TH), 2008.

[15] REYNOLDS, DOUGLAS A. und RICHARD C. ROSE: *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*. IEEE Transaction on Speech and Audio Processing, 3(1), Januar 1995.

[16] STALLKAMP, JOHANNES: *Video-based face recognition using local appearance-based models*. Diplomarbeit, Universität Karlsruhe (TH), 2006.

[17] SZASZ-TOTH, LORANT: *Studienarbeit: Open-set Face Recognition*. Universität Karlsruhe (TH), 2007.

[18] ULTES, STEFAN: *Studienarbeit: Lernen von Vor- und Nachnamen im natürlichsprachigen Mensch-Roboter-Dialog*. Universität Karlsruhe (TH), 2008.

[19] VIOLA, PAUL und MICHAEL J. JONES: *Fast Multi-view Face Detection*. Technischer Bericht, Technical Report TR2003-96, Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, Juni 2003.

[20] WITTEN, IAN H. und EIBE FRANK: *Data Mining - Practical Machine Learning Tools and Techniques*. Elsevier, 2. Auflage, 2005.

[21] XIANG, B., U. CHAUDHARI, J. NAVRATIL, G. RAMASWAMY und R. GOPINATH: *Short-time Gaussianization for Robust Speaker Verification*. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.