Kombination biometrischer Merkmale

Michael Buss , Christine Zähringer
7. Februar 2004

1. Einleitung

Wir haben bislang Methoden kennengelernt, welche in der Lage sind, eine Person durch ein spezielles biometrisches Merkmal, beispielsweise den Fingerabdruck, zu verifizieren oder zu identifizieren.
Ungünstige Aufnahmebedingungen wie beispielsweise laute Umgebung bei der Sprechererkennung oder ungünstiger Lichteinfall, Reflexe und variierende Kopfpositionen bei der Gesichtererkennung aber auch mangelnde Flexibilität der Klassifikatoren verschlechtern die Leistung eines Systems, welches nur einen einzigen biometrischen Indikator benutzt.
Durch die Kombination mehrerer Merkmale kann man die Leistung solcher Systeme verbessern, wenn die verschiedenen Experten sich gegenseitig ergänzende Informationen liefern.

2. Arten der Fusion

Man unterscheidet grob zwischen intramodaler und multimodaler Fusion.
Bei intramodaler Fusion wird ein einziges Merkmal anhand mehrerer unterschiedlicher Klassifikatoren extrahiert und die Ergebnisse zur Verbesserung der Fehlerrate verbunden.
Eine multimodale Fusion hingegen erstreckt sich über mehrere Merkmale. Ein sinnvolles Beispiel für eine solche Fusion stellt beispielsweise die Sprecher-erkennung durch Lippentracking und Audioerkennung dar.

Im Folgenden werden verschiedene Möglichkeiten vorgestellt, eine Klassifikations-Fusion durchzuführen und Ergebnisse von Fusionsexperimenten zusammengefaßt.

2.1 Fusionsebenen

Fusion kann auf unterschiedlichen Ebenen erfolgen:

a.) Merkmalsebene
Die durch verschiedene Sensoren extrahierten Merkmalsvektoren werden zu einem einzigen gemeinsamen Merkmalsvektor verbunden, der dann als Einheit klassifiziert wird.

b.) Ebene der Auswertung (soft fusion)
Die Auswertungsergebnisse werden verbunden und somit erhält man ein einziges Ergebnis, welches anzeigt, wie ähnlich der Merkmalsvektor zu dem Templatevektor ist.

c.) Entscheidungsebene (hard fusion)
Die zwei Entscheidungsklassen accept/reject können beispielsweise durch Summe oder Mehrheitsbeschluß fusioniert werden.

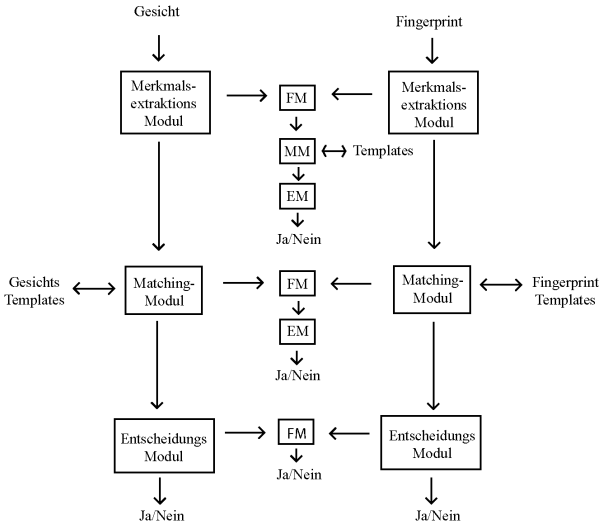

Abbildung 1: Drei Möglichkeiten, eine Fusion durchzuführen (FM= Fusionsmodul, MM=Matching-Modul, EM=Entscheidungs-Modul)

2.2 Algorithmen

Man unterscheidet grob zwischen festen und trainierbaren Regeln. Erstere, wie beispielsweise die Summe eignen sich am besten bei Kombinationen von Experten mit unterschiedlicher Leistung. Trainierbare Regeln hingegen sind flexibler, da die Parameter auf einem Trainingsset trainiert und sie somit besser angepaßt werden.
Für die multimodale Fusion, in denen die Experten unterschiedliche Leistungen aufbringen, sollten zwar theoretisch trainierbare Regeln geeignet sein, doch die schlechte Qualität der kann diesen Vorteil Trainingssets in realen Applikationen schnell wieder aufwiegen.

Für die im folgenden erläuterten festen Regeln untersucht man die Wahrscheinlichkeiten $P(\omega_1|x_i)$ und $P(\omega_2|x_i)$ des Experten $x_{i,}$ wobei ω_1 und ω_2 die Entscheidungklassen „Identität korrekt“ bzw. „Eindringling“ darstellen.

Zu den festen Regeln gehören unter anderem:

- Mittelwertbildung
  Die Wahrscheinlichkeiten, die jeder der Klassifikatoren für eine der Klassen liefert, werden

addiert und diejenige Klasse, welche das größere Ergebnis liefert, ist entscheidend.

- Mehrheit bei Abstimmung
  Hier wird die wahrscheinlichere Klasse eines Klassifikators als 1 und die unwahrscheinlichere als 0 gewertet. Analog zur Summe werden die Ergebnisse addiert und die Klasse mit dem größeren Ergebnis gewinnt.
- Verfahren der Ordnungsstatistik (Minimum, Median, Maximum)
  Pro Klasse wird von allen Klassifikatoren das Maximum/ Median/Minimum genommen und die Klasse mit dem größeren Wert ist entscheidend.
  Diese Lösung liegt zwischen fixed und trained rules. OS-based rules sind nämlich flexibler als fixed rules, da sie nicht nur die Ergebnisse addieren, sondern zwischen den Ergebnissen auch auswählen.

Zu den trainierbaren Regeln gehören unter anderem:

- Gewichtete Summe
  Die gewichtete ähnlich wie die Summe, mit dem Unterschied, daß die Wahrscheinlichkeiten der einzelnen Klassifikatoren mit normalisierten Faktoren gewichtet werden.
- Behaviour Knowledge Space (BKS)
  Hier wird ein Vektor aus den Entscheidungen jedes einzelnen Klassifikators gebildet. Jeder dieser Vektoren stellt einen Punkt in einem c-dimensionalen (c=Anzahl der Klassen) Raum, dem BKS dar. Für jeden dieser Punkte wird die Klasse geschätzt, welche die höchste Anzahl an Muster aufweist.
  BKS bildet also die Entscheidungsvektoren auf Klassen ab.

2.3 Ergebnisse

Die Experimente mit intramodaler und multimodaler Fusion, einige davon beschrieben in [1], [2] und [3] haben zwar keine goldene Regeln ergeben, welche einem System stets zu einer guten Leistung verhilft, aber es haben sich Tendenzen herauskristallisiert, die hier kurz erläutert werden.

Es wäre naheliegend zu glauben, daß man durch die Selektion möglichst vieler sehr guten Experten und mit viel Training ein immer besseres Ergebnis erhält. Dem ist jedoch nicht so. Vielmehr sollte man die Selektion der Experten anstatt nach deren Einzelleistung aufgrund der Überschneidung ihrer Fehlermengen entscheiden.
Die Fusion von Experten mit unterschiedlichen Fehlern hat sich nämlich als ergiebiger herausgestellt [2]. Je mehr gemeinsame Fehler die Experten hatten, desto uneffektiver war deren Kombination. In diesem Zusammenhang spielt auch die Vorverarbeitung eine wichtige Rolle, da sie für die Diversität der Klassifikatorausgaben verantwortlich ist.
Weiterhin haben Experimente in [3] ergeben, daß Kombinationen, deren Klassifikatoren unterschiedliche Leistung aufwiesen, effektiver waren. Die Kombination der zwei besten Klassifikatoren ergibt also nicht zwangsläufig das beste Ergebnis.
Tests haben außerdem ergeben, daß eine steigende Anzahl von Klassifikatoren nicht zu einer stetigen Leistungssteigerung führt (s. auch Abb. 2). Die Kombination von acht Klassifikatoren fiel in der Leistung sogar so weit zurück, daß die Fusion von nur zwei Klassifikatoren weitaus besser war.

Abbildung 2 gibt einen kleinen Überblick über die ermittelten durchschnittlichen Fehlerraten der Fusionsalgorithmen. Die optimalen Kombinationen der Klassifikatoren wurden auf einem separaten Trainingsset ermittelt.
Hier sieht man deutlich, daß die Kombination zu vieler Klassifikatoren die Fehlerrate verschlechtert.
Auch zu erwähnen ist, daß feste Regeln, vor allem die Mittelwertbildung und die Mehrheit bei Abstimmung, zum Teil besser als trainierbare Regeln abgeschnitten haben.

Anzahl	2	3	4	5	6	7	8
Klassifikator Mittelwert	14 0.69	134 0.62	1348 0.56	12348 0.56	123468 0.57	1234568 0.60	alle 0.82
Klassifikator Mehrh. b. Abst.	34 2.66	134 0.79	1348 0.45	12348 0.45	123478 0.71	1234678 0.83	alle 2.82
Klassifikator Min	34 2.417	134 2.417	1234 2.417	12348 0.552	123458 0.722	1234578 0.724	alle 50.000
Klassifikator Med	46 0.663	247 0.614	1347 0.532	12467 0.643	123467 0.494	1234678 1.147	alle 2.640
Klassifikator Max	45 0.573	456 0.573	4567 0.573	24567 0.573	245678 0.573	2345678 0.573	alle 0.574
Klassifikator Gew. Summe	46 0.671	246 0.658	1246 0.658	12346 0.650	123468 0.669	1234568 0.691	alle 0.809
Klassifikator BKS	34 0.740	134 0.740	1348 0.680	12348 0.610	123478 0.740	1234678 0.740	alle 1.400

Abbildung 2: Durchschnittliche Fehlerrate der verschiedenen Kombinationen

Wie oben bereits erwähnt, sind trainierbare Regeln nur im idealen Fall, d.h. ein Set zu Training- und Testzwecken, den festen Regeln vorzuziehen ist. In einem solchen Fall bringt eine Fusion wie das gewichtete Mittel der Klassifikatoren eine bessere Leistung dadurch, daß die ermittelten Idealgewichte leistungsschwache Klassifikatoren ausrangieren können. Unter realen Bedingungen jedoch, in denen schlechte Aufnahme-bedingungen wie Rauschen oder Lichtreflexe das Eingangssignal stören, kann sich die Leistung der Klassifikatoren derart verändern, daß die Gewichtung nicht mehr stimmt.
Unter Realwelt-Bedingungen erwies sich die Summe meist als die beste Fusionsmethode.

3. Anwendungsbeispiel: Sprechererkennung mittels Sprache und Lippenbewegung

In dem hier vorgestellten Verfahren [4, 5] wird sowohl die Sprache als auch die synchron mittels Kamera aufgenommenen Lippenbewegungen zur Sprecher-

erkennung herangezogen und zwecks Ergebnisoptimierung fusioniert.

3.1 Merkmalsextraktion

Zur Erfassung der Lippenbewegungen wird zunächst ein Konturmodell erstellt.
Als Referenzpunkte dienen die beiden äußeren Eckpunkte der Lippen, welche die Skalierung, den Winkel und das Zentrum festlegen. Dazwischen werden entlang der äußeren und inneren Lippenkontur in gleichen Abständen weitere Punkte verteilt.

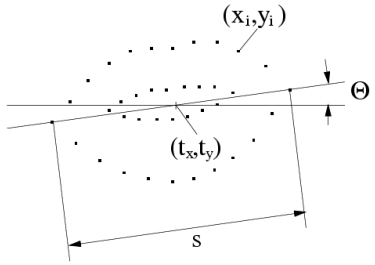

Abbildung 3: Konturmodell mit Zentrum, Skalierung, Winkel und Konturpunkten.

Dieses Modell wird nun von Hand auf Testbilder übertragen und diese dann mittels Hauptkomponentenanalyse untersucht. Die gewonnenen Hauptkomponenten stellen dann die Grundformen dar, aus denen mittels Linearkombination jede Lippenkontur approximiert werden kann.

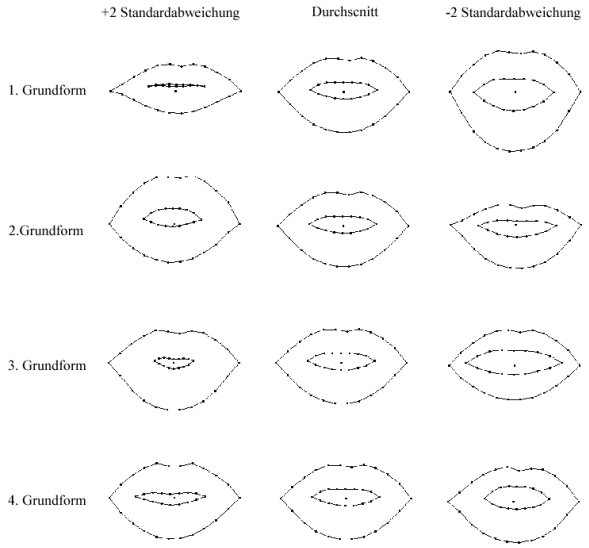

Abbildung 4: Die ersten 4 Grundformen des Konturmodells mit Durchschnitt und +/- 2 Standardabweichung

Aufbauend auf dem Konturmodell wird dann aus denselben Testbildern ein Intensitätsmodell erstellt. Dazu werden jeweils an den Konturpunkten senkrecht zur Lippenkontur die Grauwerte gemessen.

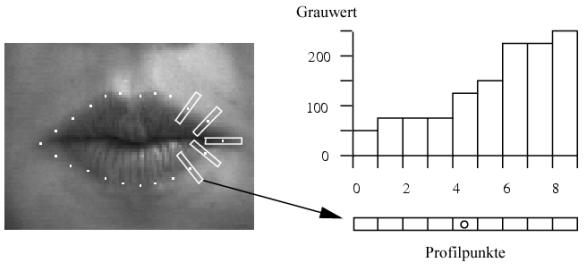

Abbildung 5: Intensitätsmodell

Anschließend werden wieder mittels Hauptkomponentenanalyse die Grundformen der Intensitätsverteilung ermittelt, mit denen die Intensitätsverteilungen durch Linearkombination der Hauptkomponenten approximiert werden können.

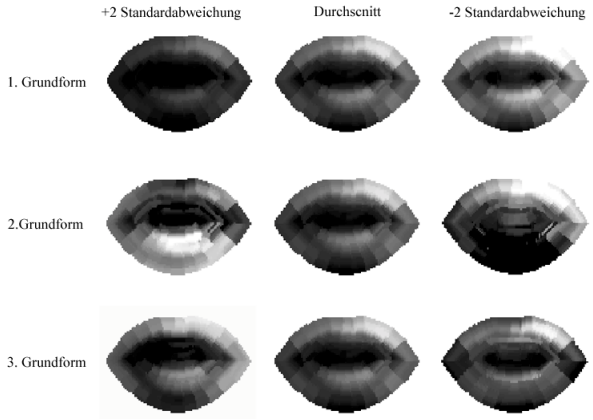

Abbildung 6: Die ersten 3 Grundformen des Intensitätsmodells mit Durchschnitt und +/-2 Standardabweichung

Zur automatischen Lokalisierung der Lippen wird nun das Standard-Intensitätsmodell an zufälliger Stelle im Bild plaziert und versucht, durch Minimierung einer Kostenfunktion, die Lippen zu finden.
In der Kostenfunktion werden lediglich Abweichungen vom Standard-Intensitätsmodell mit höheren Kosten belegt. Das Konturmodell wird nur insofern berücksichtigt, dass völlig unrealistische Konturen ausgeschlossen werden. So wird gewährleistet, daß Abweichungen von der mittleren Kontur nicht zu einer Erhöhung der Kosten führen und dadurch die Lokalisierung bei Personen mit einer Lippenkontur, welche weiter vom Standard entfernt ist, robuster wird.
Für das Lokalisieren in den nachfolgenden Bildern einer aufgenommenen Sequenz wird jeweils die Position im Vorgängerbild als Ausgangsposition verwendet.

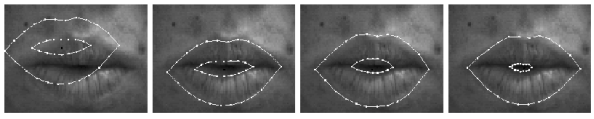

Abbildung 7: Initialisierung und Ergebnis der Lokalisierung nach 5, 10 und 20 Iterationen

Die Merkmalsextraktion der Audioaufnahme erfolgt mittels MFCC (mel frequency cepstrum coefficients) Verfahren, das dem Cepstrum-Verfahren in den Vorträgen zur Spracherkennung entspricht und deshalb hier nicht ausführlicher behandelt werden soll.

3.2 Klassifizierung

Da man für den Test des Systems zwei unterschiedlichen Datenbanken mit verwendete, wurde die Klassifizierung der extrahierten Merkmale an das jeweilige Bild und Tonmaterial der Datenbanken angepaßt. Außerdem wurden unterschiedliche Fusionsmethoden für die einzelnen Datenbanken verwendet.

3.2.1 Tulips1 Datenbank (Fusion auf Merkmalsebene)

Die Tulips1 Datenbank enthält Bild und Tonmaterial von 12 Personen und insgesamt 96 Wörtern (4 Wörter pro Person à 2 Sets), wobei die Hauptschwierigkeit wohl bei der stark variierenden Bildqualität lag und deshalb eher zum Testen des Lippentrackings als zum Klassifizieren verwendet wurde.

Für den visuellen Teil wurden 10 Merkmale des Konturmodells und 20 Merkmale des Intensitätsmodells verwendet.
Für den Audioteil wurden 12 MFCC Parameter verwendet, welche bei einer Rate von 30Hz (MFCC_30) extrahiert wurden, um mit der Bildrate der Videos übereinzustimmen. Da dies wesentlich weniger ist als die 100Hz (MFCC_100), die normalerweise verwendet werden, wurde dies zum Vergleich zusätzlich noch herangezogen.

Die Klassifikation wurde einmal mit traditionellen Hidden Markov Models (HMM) durchgeführt, wobei für einen Sprecher jeweils pro Wort ein HMM trainiert wurde (multiple HMM/M-HMM).
Da die Datenbank recht wenig Wörter enthält, wurde in einer anderen Versuchsreihe deshalb ein einziges HMM (single HMM/S-HMM) pro Sprecher für alle 4 Wörter trainiert. Eine dritte Versuchsreihe wurde mit Gaussian Mixture Models realisiert.

Für die Fusion wurden die jeweiligen Merkmale zu einem einzigen Merkmalsvektor zusammengefaßt (Fusion auf Merkmalsebene), und als Eingabe für den Klassifikator verwendet.

Die Ergebnisse zeigten, daß bei den wenigen Trainingsdaten die zur Verfügung standen, die visuellen Merkmale ein besseres Ergebnis liefern als die des Audio. Die Ergebnisse der Fusion waren teilweise besser als die der Einzelklassifikation.

Methode	Kontur	Intensität	Kont. +Int.	MFCC_30	MFCC_100	Kont.+Int.+ MFCC_30
M_HMM	72.9%	89.6%	91.7%	70.8%	81.3%	93.8%
S-HMM	83.3%	95.8%	97.8%	70.8%	85.4%	97.9%
GMM	81.3%	97.9%	95.8%	60.4%	75.0%	95.8%

Abbildung 8: Identifikationsrate für multiple HMMs (M-HMM), einzelne HMMs (S-HMM) und GMMs mit unterschiedlichen Merkmalen

3.2.2 M2VTS Datenbank (Fusion auf Entscheidungsebene / Hard fusion)

Die M2VTS Datenbank enthält mit insgesamt 1850 Wörter (37 Personen, 10 Wörter à 5 Sets) und ist damit wesentlich besser für das Trainieren von Klassifikatoren geeignet.

Hier wurden die Audiomerkmale getrennt von den visuellen Merkmalen klassifiziert und danach die Ergebnisse mittels der gewichteten Summe zusammengeführt. Im Unterschied zur Tulips Datenbank wurden für den visuellen Teil 14 Merkmale des Konturmodells und 10 Merkmale des Intensitätsmodells benutzt.
Als Klassifikatoren wurden GMMs und HMMs verwendet.

Die Ergebnisse zeigen, daß die rein akustische Klassifikation bei den vielen zur Verfügung stehenden Trainingsdaten besser abschneidet als die rein visuelle Klassifikation. Zudem verbessert die Fusion das Ergebnis hier doch erheblich.

Typ	Validierung			Test		
	ID	FAR	FRR	ID	FAR	FRR
Akustisch	100.0%	2.5%	0.0%	97.2%	2.3%	2.8%
Visuell	82.3%	4.9%	8.8%	72.2%	3.0%	27.8%
Fusion	**100.0%**	**0.6%**	**0.0%**	**100.0%**	**0.5%**	**2.8%**
Anzahl Tests	36	1332	36	36	1332	36

Abbildung 9 : Ergebnisse von Validierung und Test. ID ist die korrekte Identifikation, FAR die false acception rate und FRR die false reject rate.

Referenzen:

[1] Arun Ross und Anil Jain. Information fusion in biometrics

[2] Jacek Czyz, Josef Kittler und Luc Vandendorpe. Combining Face Verification Experts

[3] Fabio Roli, Josef Kittler, Giorgio Fumera und Daniele Muntoni. An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems

[4] Juergen Luettin. Visual Speech and Speaker Recognition.

[5] Pierre Jourlin, Juergen Luettin, Domenique Genoud, Hubert Wassner. Acoustic-Labial Speaker Verification.