Sprechen: Ein Hindernis in der modernen Mensch-Maschine-Kommunikation?

Carmen Frank, Elmar Nöth

Lehrstuhl für Mustererkennung, FAU Erlangen-Nürnberg, Email: frank,noeth@immd5.informatik.uni-erlangen.de

Zusammenfassung

In einem multimodalen Dialogsystem beeinträchtigt das Sprechen des Benutzer die Mimikerkennung. Die hier vorgestellte Fusion von Einzelerkennungsergebnissen erlaubt es trotzdem emotionale Gesichtsausdrücke klassifizieren zu können.

Einleitung

In der natürlichen Mensch-Mensch-Kommunikation werden Informationen in sehr viel mehr Kanälen übertragen, als in Standard-Dialogsystemen heutzutage verarbeitet werden. Zu diesen Informationen gehören Emotionen, die über den Gesichtsausdruck geäußert werden und den Dialogverlauf sehr stark beeinflussen können:

- ein verärgerter Dialogpartner ist von einem Spracherkenner schwer zu verstehen → eine angepasste Dialogführung beruhigt den Anwender
- ein unsicherer/hilfloser Anwender ist dankbar für erweiterte Hilfestellungen; ein geübter Anwender reagiert darauf eher genervt

Für ein multimodales Dialogsystem ist die Erkennung von Benutzerzustandsklassen wie Ärger, Zufriedenheit oder Ratlosigkeit interessant [Ste01]. Die Erkennung dieser Benutzerzustände wurde im Projekt SmartKom[1] untersucht. Aufgrund der kulturellen *Display Rules*, verbergen Menschen ihre Emotionen vor Fremden [Buc84]. Dies zeigt auch die Auswertung der WoZ-Aufnahmen in Tabelle 1. Nur ein sehr geringer Anteil zeigt Emotionen und die, aus Dialogsicht interessanten negativen Emotionen, sind besonders selten. Aus diesem Grund verwenden die vorgestellten Experimente den Datensatz aus [Mar98] mit gespielten Emotionen.

Benutzerzust.	ges. Dauer	mittl. Dauer	Anteil
Freude	19.6 min	13.4 sec	5.0 %
Ärger	9.0 min	8.3 sec	2.3 %
Ratlos	19.9 min	11.8 sec	5.0 %
Überlegen	54.4 min	25.9 sec	13.7 %
Überrasch.	1.6 min	2.7 sec	0.4 %
Neutral	292.2 min	199.3 sec	73.6 %

Tabelle 1: Die Tabelle zeigt, dass positive Emotionen in den SmartKom-Daten häufiger vorkommen als negative, aber dennoch einen sehr geringen Anteil der Gesamtdaten ausmachen.

[1]Diese Forschung wird unterstützt vom Bundesministerium für Bildung und Forschung (*BMBF*) als Projekt SmartKom unter Grant 01 IL 905 K7. Die Verantwortlichkeit für den Inhalt dieser Studien liegt bei den Autoren.

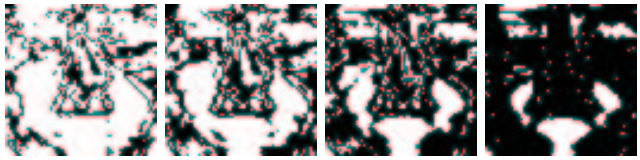

Abbildung 1: Die hellen Bereiche markieren emotional aussagekräftige Gesichtsbereiche mit verschiedenen Detailtiefen.

Ein Problem bei der Gesichtsausdruckserkennung für ein multimodales Dialogsystem ist die gegenseitige Beeinflussung der verschiedenen Modalitäten. Die Signifikanz der verschiedenen Gesichtsbereiche ist in Abbildung 1 dargestellt. Das Sprechen verändert beispielsweise die Erscheinung der für die Mimikerkennung wichtigen Mundregion und damit den Gesichtsausdruck. Für eine sichere Gesichtsausdruckserkennung muss dies beachtet werden. Das Erkennungsverfahren, das in dieser Arbeit vorgestellt werden soll, ist ein angepasstes Eigenraumverfahren.

Einführung in Eigenräume

Eigenraumverfahren sind bekannt für das Einsatzgebiet der Personenerkennung [Mog94, Yam00].
Um einen Eigenraum aus Trainingsbildern zu erzeugen, wird eine partielle Karhunen-Loéve Transformation verwendet, auch Hauptachsentransformation (PCA). Dabei handelt es sich um eine Dimensionsreduktion, die die Streuung aller projizierten Muster maximiert, wobei N Musterbilder einer Person $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ aus einem n-dimensionalen Merkmalsraum verwendet werden. Sei $\vec{\mu}$ der Mittelwert aller Merkmalsvektoren. Dann ist die Streuungsmatrix definiert als

$$S_T = \sum_{k=1}^{N} (\vec{x}_k - \vec{\mu})(\vec{x}_k - \vec{\mu})^T \qquad (1)$$

Bei einer PCA wird die optimale Projektion W_{opt} in einen nieder-dimensionalen Unterraum so gewählt, dass die Determinante der Streuungsmatrix der projizierten Muster maximiert wird,

$$W_{opt} = arg\max_{W} |W^T S_T W| = [w_1, w_2, \ldots, w_m] \qquad (2)$$

wobei $\{w_i | i = 1, 2, \ldots m\}$ die Menge der n-dimensionalen Eigenvektoren von S_T in absteigender Größe der Eigenwerte darstellen. Diese Eigenvektoren haben die selbe Dimension, wie die Eingabevektoren und werden deshalb auch *Eigenfaces/Eigengesichter* genannt. Die Dimension der projizierten Muster wird mit m bezeichnet und ist im Fall der Merkmalsreduktion $m < n$.

reference facial expression	results			
	smile	*neutral*	*anger*	*shout*
smile	**77**	6	7	10
neutral	1	46	**47**	6
anger	2	25	**71**	2
shout	22	7	5	**66**

Tabelle 2: Bei Verwendung der Nasen-Region zur Emotionsklassifikation ergibt sich diese Verwechslungsmatrix (Werte in %) bei einer Erkennungsrate von 65%.

Im Folgenden werden Eigenvektoren mit hohen Eigenwerten als hohe Eigenvektoren bezeichnet.

Für die Klassifikation von Gesichtsausdrücken hat die direkte Anwendung des Standard-Vorgehens der Personenerkennung [Tur91] nicht bewährt. Bessere Ergebnisse werden erzielt, wenn jede Klasse für sich als eigene „Gesichts"-Klasse behandelt wird. Ein zu klassifizierendes Bild wird in jeden Emotions-Eigenraum projiziert und derjenigen Klasse zugeordnet, die den geringsten Rückprojektionsfehler erzeugt.

Verwendet man Gesichtsausschnitte, die um die Nase zentriert sind, so erreicht die vorgestellte Methode eine Erkennungsrate von 65% für die Klassen *neutral, smile, anger, shout* des Datensatzes [Mar98]. Die Verwechslungsmatrix zu diesem Versuch ist in Tabelle 2 dargestellt.

Experimente

Wie oben erwähnt, sollte im Fall eines sprechenden Anwenders auf Merkmale aus der Mundregion verzichtet werden. Um die Erkennungsleistung anderer Gesichtsregionen festzustellen wurden die Versuche auch für den Augen- und Mund-Bereich wiederholt. Der Augenbereich erzielt eine Erkennungsrate von 63% und die Mundregion von 79%. Die hohe Erkennungsleistung des Mundbereichs belegt die Vermutung, dass diese Region aussagekräftig bezüglich des Gesichtsausdrucks ist und auf jeden Fall dann als Merkmal verwendet werden sollte, wenn es nicht durch Sprache verfälscht wird.

In den Fällen, in denen ein Anwender spricht, kann das Klassifikationsergebnis verbessert werden, in dem man die Ergebnisse der anderen Regionen fusioniert. Dies geschieht durch ein automatisch generierte Regelsystem. Zwei der Regeln sind

Regel 3

IF	(one classifier says *shout*) AND (one classifier says *smile*)
THEN	over all result is *smile*

Regel 4

IF	(eyes classifier says *neutral*) AND (nose classifier says *anger*)
THEN	over all result is *neutral*

Dieses Regelsystem wird mit einem Datensatz getestet, der keine Bilder der Klasse *shout* enthält. Die Klassifikation auf dem Augenbereich erreicht darauf 57%, die auf dem Nasenbereich 64%. Durch das Regelsystem erzielt man eine verbesserte Erkennungsrate von 67%, das entspricht einer Verbesserung von 10% bezüglich der Augenregion und 3% bezüglich des Nasenbereichs, wie in Tabelle 3 dargestellt ist. Details zu diesem Vorgehen finden sich in [Fra05].

	Augen-Region	Nasen-Region
Standard	57%	64%
Erweiterung	67%	
Verbesserung	10 %	3%

Tabelle 3: Das vorgestellte Verfahren erzielt für die Augen-Region eine Verbesserung der Klassifikationsrate um 10 Prozentpunkte auf 67%.

Schlussfolgerung

Das Sprechen eines Benutzers in einem multimodalen Dialogsystem ist unvermeidbar, behindert allerdings die Klassifikation der Mimik. Mit Hilfe des Dialogsystems kann allerdings entschieden werden, zu welchen Zeitpunkten dies beachtet werden muss. Während einer Sprechphase des Benutzers kann beispielsweise zusätzlich Prosodie als Indikator für den Emotionalen Zustand herangezogen werden. Oder die Klassenzugehörigkeit des Gesichtsausdrucks wird durch die vorgestellte Fusion der Erkennungsergebnisse von weniger stark beeinflussten Gesichtsregionen bestimmt.

Literatur

[Buc84] Buck, R.: *The communication of emotion*, The Guilford social psychology series, Guilford Press, New York, NY, USA, 1984.

[Fra05] Frank, C.: *The facial expression module*, in Wahlster, W. (Hrsg.): *The SmartKom Project*.

[Mar98] Martinez, A.; Benavente., R.: *The AR Face Database*, Purdue University, West Lafayette, IN 47907-1285, 1998.

[Mog94] Moghaddam, B.; Pentland, A.: *Face Recognition Using View-Based and Modular Eigenspaces*, in *Vismod, TR-301*, 1994.

[Ste01] Steininger, S.; Siepmann, R.; Beiras-Cunqueiro, C.; Glesner, A.: *Labeling von User-States im Mensch-Maschine Dialog*, Technisches Dokument 17, BMBF Projekt SmartKom, DFKI, Saarbrücken, 2001.

[Tur91] Turk, M.; Pentland, A.: *Face Recognition Using Eigenfaces*, in *Proceedings of Computer Vision and Pattern Recognition*, 1991, S. 586–591.

[Yam00] Yambor, W.; Draper, B.; Beveridge, J.: *Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures*, in *Second Workshop on Empirical Evaluation Methods in Computer Vision*, 2000.