Automatische Erfassung und Analyse der menschlichen Mimik

Ulrich Canzler

Lehrstuhl für Technische Informatik
Rheinisch-Westfälische Technische Hochschule (RWTH), 52074 Aachen
Email: canzler@techinfo.rwth-aachen.de

Zusammenfassung. Erkennung und Interpretation von Gebärdensprachen, den rein visuellen Sprachen unter Gehörlosen, stellen unter Echtzeitanforderungen eine besondere Herausforderung an high-level Verfahren der Bildverarbeitung. Gebärden bestehen aus manuellen Komponenten (Handform, Handstellung, Ausführungsstelle...) und nicht-manuellen Komponenten (Kopf, Blickrichtung, Gesichtsausdruck, Mundbild). Letztere spielen eine elementare Bedeutung für Grammatik und Textverstehen der Sprache. Im folgenden wird ein System vorgestellt, welches in der Lage ist, zuverlässig das menschliche Gesicht innerhalb eines Bildes aufzufinden und anschließend die Mimik zu analysieren. Die Mimikmerkmale werden dann mittels des Facial Action Coding Systems (FACS) codiert und anschließend einem HMM-Klassifikator übergeben.

1 Einleitung

Gebärdensprachen stellen vollwertige lebendige Sprachen für die Kommunikation von und mit Gehörlosen dar. Die Vermittlung linguistischer Inhalte erfolgt dabei durch die Kombination von manuellen Parametern (Handform, Handstellung, Ausführungsstelle, Handbewegung) und nicht-manuellen Parametern (Oberkörper, Kopf, Blickrichtung, Gesichtsausdruck, Mundbild).
Innerhalb der Gebärdensprache spielen die nicht-manuellen Parameter eine besondere Rolle, da in ihnen nicht nur – wie bei den Hörenden – Gefühle mitgeteilt werden, sondern zusätzlich Informationen mitverschlüsselt werden, die von zentraler Bedeutung für die Grammatik der Sprache sind[1]. Unter anderem werden durch sie folgende Sprachsignale vermittelt:

- Nicht manuell ausdrückbare Adjektive und Adverbien (nah-fern, intensiv, etc...)
- Verschiedene Satztypen (Verneinung, Frage-, Relativ- und Konditionalsätze)
- Direkte und indirekte Rede
- Mundbilder (Lippenbewegungen, die denen der Artikulation ähneln)

Erst durch die Beachtung der nicht-manuellen Parameter wird also die Klassifikation bestimmter Gebärden möglich und deren grammatikalischer Zusammenhang interpretierbar. Des weiteren können die gewonnenen Merkmale die Klassifikation der

manuellen Parameter unterstützen und somit eine größere Robustheit und Erkennungsrate des Gesamtsystems ermöglichen.

2 Methode und Vorgehensweise

Vorgestellt wird das Konzept eines Systems zur automatisierten Analyse der menschlichen Mimik, welches sich derzeit am Lehrstuhl für Technische Informatik, Aachen in der Entwicklung befindet. Es besteht aus vier Modulen, die im folgenden näher beschrieben werden.

2.1 Gesichts-Detektion

Eine Gesichts-Detektion muß generell dann erfolgen, wenn das System gestartet wird, also noch keine Gesichtsposition bekannt ist, oder aber die vermutete Gesichtsposition nur unzureichende Ergebnisse liefert. Dies kann auftreten, wenn der Benutzer in das Bild ein- bzw. austritt oder zuvor wegen Verschmelzungen von Gesicht und Händen die Gesichtsposition zu ungenau bestimmt wurde.
Zunächst wird der weitgehend statische Anteil des Bildes durch Temporal-Templates maskiert und für die weiteren Bearbeitungsschritte ausgeblendet. Auf den verbleibenden „areas of interest“ werden sowohl formorientierte, durch Kanten und Konturen der menschlichen Gestalt definierte, als auch farborientierte, durch die charakteristische menschliche Hautfarbe implizierte Verfahren zum Auffinden des Gesichtes angewandt.
Ein selektiver Threshold im Farbraum clustert Regionen, die eine bedingte Wahrscheinlichkeit für das Vorkommen hautfarbenen Flächen wiederspiegeln. Dabei wird auf Histogramme einer größeren Trainingsmenge zurückgegriffen.
Diese beruht auf einem Training mit zuvor manuell maskierten Bildern unterschiedlichster Inhalte und Beleuchtungen nach dem Verfahren von Jones und Rehg [2]. Ein Template-Matching mit einem multiskalen „Durchschnitts-Gesichts“ sucht durch Korrelation weitere Übereinstimmungen innerhalb der Regionen mit gesichtsähnlichen

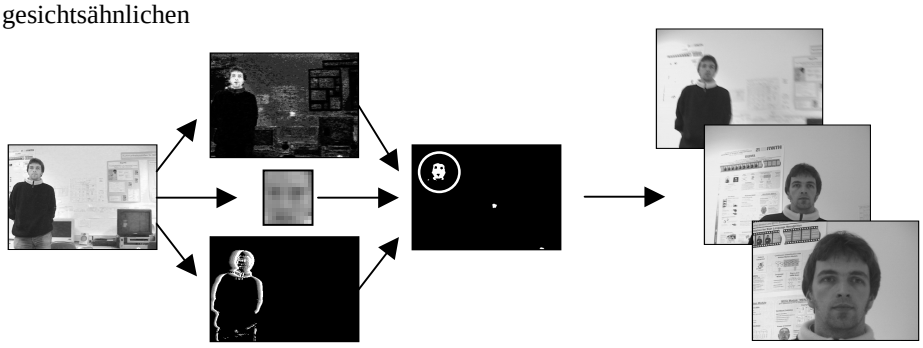

Abb.1: Ablauf der Gesichtsdetektion und Optimierung des Bildausschnittes

Objekten (Abb. 1).

Parallel wird die Entfernung zum Objekt durch Verrechnen der Linsenstellung innerhalb der Kamera abgeschätzt und dadurch ein Zusammenhang mit der absoluten Fläche der segmentierten Objekte hergeleitet. Wird durch die zuvor beschriebenen Verfahren eine Region als Gesichtsregion klassifiziert, so ist die erste Phase abgeschlossen, der Bildausschnitt wird für die nachfolgenden Verarbeitungsstufen optimiert und die zuvor benötigten Parameter, wie beispielsweise die Farbraumhistogramme des Hautfarbmodells werden adaptiv an die lokalen Lichtverhältnisse für spätere Aufrufe der Methode angepaßt.

2.2 Tracken charakteristischer Punkte

Ist das Gesicht mit einem bestimmten Sicherheitsmaß detektiert worden, so ist es anschließend nicht mehr notwendig, weitere zeitaufwendige Vollbild-Segmentierungen durchzuführen. Es reicht nun aus, einige aussagekräftige Punkte innerhalb des Gesichtfeldes zu verfolgen.
Dazu bieten sich charakteristische Punkte der Körper- und Kopfkontur an, da aus diesen gleichzeitig über Symmetriebeziehungen auf bestimmte Gesichtzonen, wie Augen, Nase und Mund, zurückgeschlossen werden kann. Beispielsweise kann es sich hierbei um Schultern, Halsansatz, seitliche und obere Kopfbegrenzungen handeln. Um diese Punkte zu definieren, wird ein Susan-Edge-Kantendetektors eingesetzt, der sich mit einem Modellprototypen abgleicht. Die Knoten dieses Prototypen beziehen sich auf ein durchschnittliches Gesicht und können begrenzt auf das akquirierte Bild gedehnt werden (Abb.2). Dieses Verfahren ähnelt dem elastic bunch graph matching, welches an der Ruhruniversität Bochum entwickelt worden ist. Dabei wird das Gesicht als Graph modelliert, dessen Kanten die räumlichen Entfernungen zueinander und die Knoten die mittels Gabor-Wavelets gewonnenen Energien charakteristischer Punkte repräsentieren[3][4].
Diese Punkte dienen als Stützstellen innerhalb einer Bildsequenz. Durch Prädiktion - basierend auf der Methodik des „Optical Flows“ und der Korrelation korrespondierender Punkte in Bildfolgen– lassen sich diese Stützstellen über die Zeit verfolgen und als erneute Ausgangspunkte zur folgenden Merkmalsextraktion verwenden.

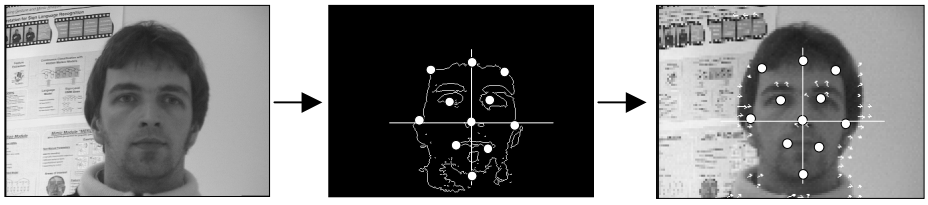

Abb.2: Bestimmen und Tracken charakteristischer Punkte mittels des optical flow

2.3 Extraktion von Mermalen

Zur Lokalisierung einzelner Gesichtsmerkmale innerhalb festgelegter Regionen wie Augen- und Mundregion, müssen zunächst letztere in Bezug zu den Stützstellen aufgefunden und festgelegt werden.
Hierzu kommt ein parametrisches Modell zur Anwendung, welches empirisch gewonnen wird. Es beruht auf geometrischen Eigenschaften, beispielsweise der ellipsenförmigen Augen und Mund-Regionen oder der trapezförmigen Nasen-/Nasenfalten-Region.
Innerhalb der so gewonnenen Regionen werden anschließend mittels eines Template-Matchings und anderer Filter-basierter Ansätze einzelne Merkmale aus diesen Gebieten extrahiert, wie z.B. die Blickrichtung, Öffnungswinkel der Augen oder Stirnfaltenbildung. Zusätzlich können Merkmale wie Größe, Rundheit, Exzentrität und weitere geometrische Größen zur weiteren Verwendung herangezogen werden. Vereinfacht wird die Informationsauswertung, da eine absolute Größeninformation vorhanden ist.
Die erhaltenen Ergebnisse werden nun dem nachfolgenden Analyse-Modul übergeben und gleichzeitig auf Güte untersucht. Wird hierbei ein bestimmtes Maß unterschritten, muß das Gesicht neu detektiert werden.

2.4 Gesichtsanalyse

Bei dem vierten Modul des Systems handelt es sich um die Gesichtsanalyse, die dem Gesamt-Klassifikator zu den Merkmalen der manuellen Parametern (von einem weiteren am Lehrstuhl entwickelten Modul) zusätzliche Merkmale in Form sogenannter Action Units (AU‘s) übergibt. Hierbei handelt es sich um einzelne, anatomisch hergeleitete, minimale Bewegungseinheiten des Gesichtes
Zu Grunde gelegt wird das ´Facial Action Coding System´ (FACS) von den Psychologen Paul Ekman und Erich Friesen [5]. Ein wichtiger Punkt dieses Systems war, Beschreibung und Interpretation von Bewegungen im Gesicht voneinander zu trennen, was durch eine Definition in Form einer beschreibenden, deskriptiven Begrifflichkeit erreicht wurde.
Jede AU kann dabei in drei verschiedenen Intensitätsstufen auftreten, gleichzeitiges Auftreten mehrerer AU’s ist gestattet (Abb.3).

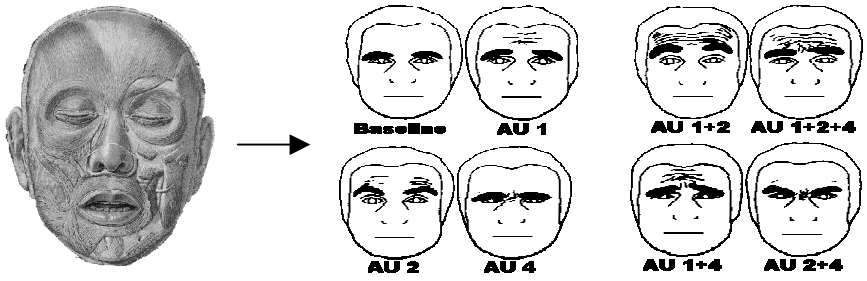

Abb.3: Muskulärer Gesichtsaufbau und exemplarische Action Units (bzw. Kombinationen)

Um das System robust und schnell konzipieren und implementieren zu können, werden nur deutlich sichtbare Bewegungen des Gesicht bewertet, wie sie in der Mimikartikulation der Gebärdensprache auch üblich sind. Nicht sichtbare Unterschiede und Veränderungen, wie z.B. die des Muskeltonus, werden explizit ausgeschlossen.
Die Anzahl der AU's wird eingeschränkt, so dass nur für die Gebärdensprache relevante Gesichtsbewegungen analysiert werden können. Zu jeder protokollierten AU werden vier weitere Eigenschaften für die Ergebnisgewinnung berücksichtigt:

- Art der Gesichtsbewegung
- Intensität der Gesichtsbewegungen
- Lateralität (Gleichzeitigkeit oberer und unterer Gesichtsbewegungen)
- Zeitlicher Verlauf

Die AU's selbst werden aus den extrahierten Merkmale des weiter oben beschriebenen Moduls unter Verwendung eines auf Hidden Markov Modellen basierenden Klassifikators gewonnen. Dieser kann vom Gebärden-Klassifikator abgeleitet werden und basiert auf dem statistischen Vergleich eines Trainingssatzes mit den aktuell extrahierten Merkmalen.
Um dabei eine größere Übereinstimmung bei der Klassifizierung zu erzielen, werden bestimmte Regeln über Fuzzy Sets bzgl. Dominanz, Substitution und Austauschbarkeit von Gebärden zusätzlich eingesetzt, so daß ein Hintergrundwissen bzgl. der Gebärdensprache miteinfließt.

3 Diskussion und Ausblick

Die größten Probleme bei der Erstellung des oben beschriebenen Systems werden definiert durch die Projektrahmenbedingungen der Echtzeitverarbeitung, Personenunabhängigkeit und Robustheit gegenüber gestörten Hintergründen. Da sich das System in der Entwicklung befindet, sind zur Zeit noch keine präzisen Angaben zu Erkennungsrate und Fehlverhalten möglich.

4 Literatur

1. Braem, Penny, Boyes: Einführung in die Gebärdensprache und ihre Erforschung, Intern. Arbeiten zur Gebärdensprache und Kommunikation Gehörloser, Band 11, Signum Verlag Hamburg, 1995
2. Jones M, Rehg J: Statistical color models with application to skin detection, Proceedings Computer Vision and Pattern Recognition, 1999, pp. 274-280
3. Gong S., McKenna S., Psarrou A.: Dynamic Vision,Imperial College Press, London, 2000
4. Würtz R.: Multilayer Dynamic Link Networks for Correspondences and Visual Object Recognition, Verlag Harri, Bochum 1994
5. Ekman P, Friesen W: Facial Action Coding System, Consulting Psychologists Press Inc., California, 1978