

SIPBILD – Mimik- und Gestikerkennung in der Mensch-Maschine Schnittstelle

Matthias Wimmer, Bernd Radig und Christoph Mayer
Fakultät für Informatik, Technische Universität München,
Boltzmannstr. 3, 85748 Garching, Germany
wimmer@forsip.de <http://www.radig.cs.tum.edu>

Abstract: Für eine natürliche Mensch-Maschine Interaktion spielt die Interpretation visueller Informationen eine zentrale Rolle. Fehlende Kontrolle der Umgebungsbedingungen wie Helligkeit und Hintergrundfarbe stellt hohe Anforderungen an die Bilderkennungssoftware. SIPBILD schafft es, mit modellbasierter Bildinterpretation die menschliche Mimik und Gestik zu erkennen. Um diese Technik in natürlichen Umgebungen einzusetzen, ist es allerdings notwendig, die bisherigen Techniken entscheidend zu verbessern. Insbesondere stellen wir eine Vorgehensweise vor, die robustes Model-Fitting ohne spezielles Fachwissen in der Bildverarbeitung erreicht und der Einsatz dieser Technik somit keinen Experten mehr verlangt.

1 Einleitung

Die menschliche Kommunikation ist komplex. Menschen tauschen Informationen nicht nur durch Sprache sondern auch durch Stimmfärbung, Körperhaltung, Gestik und Mimik aus. „Man kann nicht nicht kommunizieren“, sagt der Altmeister der Kommunikation, Paul Watzlawick. Mimiken sind intuitive und interkulturell verständliche Mittel zur Verständigung, die auch spontan und unbewusst ausgeführt werden. Somit ist es hilfreich, auch dem Rechner diese Ausdrucksform beizubringen. In den 1970er Jahren entdeckten Ekman und Friesen [Ek72] sechs universelle Mimiken, die alle Menschen unabhängig von Geschlecht, Hautfarbe und Kulturkreis gleich ausdrücken und gleich verstehen: Lachen, Wut, Abscheu, Trauer, Angst und Überraschung, siehe Abb. 1. Unser Ziel ist es, diese sechs universellen Mimiken in Videobildern zu erkennen.

Der Interpretation von Gestik und Mimik kommt beim Einsatz zukünftiger Mensch-Maschine Schnittstellen eine Schlüsselrolle zu. Das Teilprojekt SIPBILD analysiert visuelle

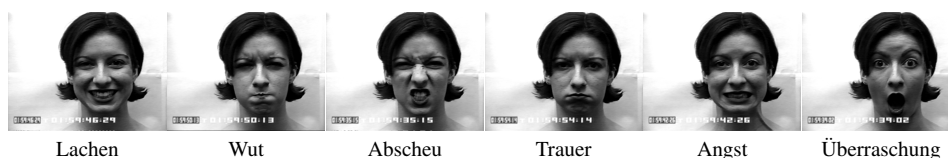


Abbildung 1: Die sechs universellen Mimiken (Cohn-Kanade Datenbank [KCT00]).



Abbildung 2: Vorgefertigtes Modellwissen erlaubt unseren Algorithmen, menschliche Gesichter in Bildern zu erkennen und deren Eigenschaften zu analysieren.

Information und setzt sie für natürliche und intuitive Interaktion zwischen Mensch und Rechner gewinnbringend ein. Im Rahmen dieses Projekts wurden Demonstratoren entwickelt, die geometrische Modelle benutzen, um eine Interpretation komplexer Gestik und Mimik zu ermöglichen, siehe Abb. 2.

2 Problemstellung und Use-Cases

Viele Menschen empfinden technische Geräte als schwierig, aufwändig und umständlich in der Handhabung, denn bis jetzt muss der Mensch das die Benutzung des Gerätes verstehen. Für ein angenehmere Handhabung müssen jedoch Hightechgeräte den Anwender verstehen und auf Körper-sprache oder Stimmführung reagieren. Doch bis jetzt ist dies noch Zukunftsmusik.

Der Erfolg erster kommerzieller Computerspiele, die bereits rudimentäre, kamerabasierte Bewegungsdetektion zur Interaktion integrieren, zeigt das große Interesse und die Akzeptanz gestenbasierter Mensch-Maschine Kommunikation. Deshalb entwickelten wir mehrere Prototypen von Spielen, die mit unseren Algorithmen zur Detektion der Körperbewegung gesteuert werden. Insbesondere die Analyse menschlicher Emotion während des Spiels ist bisher ein unerforschter Aspekt der Kommunikation zwischen Spieler und Spiel, den wir erproben.

Wir fokussieren uns dabei auf zwei bekannte Spiele: einen Ego-Shooter und den Klassiker Pac-Man. Ego-Shooter sind Computerspiele, bei denen der Spieler die dreidimensionale Spielwelt durch die Augen der Spielfigur sieht. Bei unseren Beispielspielen dirigiert der Spieler die Spielfigur durch seine eigene Körperbewegung inkl. Kopfbewegungen durch die virtuelle Welt und fühlt sich dadurch mehr in das Spielgeschehen hineinversetzt.

Ein weiteres Beispiel für den Einsatz von Mimikererkennung stellt computerunterstütztes Lernen dar, wobei der Computer die Tutorfunktion übernimmt. Er erklärt den Übungsinhalt und fragt das erworbene Wissen anschliessend ab. Versteht der Rechner dabei das menschliche Verhalten und die Emotionen, erhöht sich die Qualität des Unterrichts enorm, da er flexibel auf Langeweile oder Anspannung reagieren kann.

3 Lösungskonzept

Modellbasierte Bildinterpretation hat sich als robuster Ansatz für die Objekterkennung in Alltagssituationen erwiesen. Die geometrischen Modelle enthalten hierbei Wissen zur Beschreibung des Objekts im Bild und reduzieren die immense Datenflut in Bildern auf eine überschaubare Anzahl von Modellparametern. Wir benutzen ein auf Statistiken basierendes, deformierbares Gesichtsmodell nach dem Ansatz von Cootes et al. [CT92]. Modellparameter beschreiben das Aussehen des individuellen Gesichts und dessen momentanen Gesichtsausdruck.

Beim Model-Fitting werden diejenigen Modellparameter bestimmt, die das aktuelle Kamerabild am besten beschreiben. Üblicherweise sind daran zwei Komponenten beteiligt: (1) Die Bewertungsfunktion (engl. objective function) ermittelt wie gut das Modell mit dem aktuellen Kamerabild übereinstimmt und (2) der Fitting-Algorithmus sucht das globale Minimum dieser Funktion. Als inhärenter Bestandteil für diese Aufgabe erweist sich die Bewertungsfunktion (engl. objective function), zu deren Verbesserung wir wissenschaftliche Beiträge liefern.

Diese Funktionen werden oft manuell erstellt, indem der Programmierer passende Bildmerkmale und deren mathematische Verknüpfung selektiert. Danach inspiziert er die Ergebnisse, die diese Funktion für eine kleine Anzahl von Beispielbildern liefert. Ist er mit dem Ergebnis nicht zufrieden, ändert er die Funktion oder erstellt sie komplett neu, siehe Abb. 3 (links). Diese manuelle Vorgehensweise benötigt viel Domänenwissen, die resultierende Funktion ist sehr ungenau und der gesamte Arbeitsaufwand ist im vornherein aufgrund der *Erstellen-Prüfen*-Schleife nicht abschätzbar.

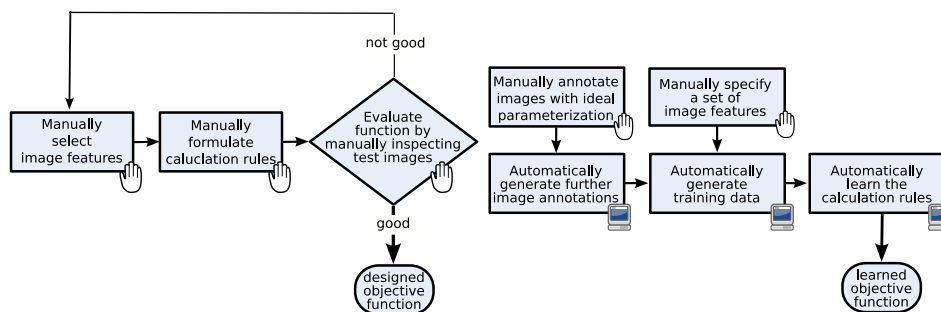


Abbildung 3: Erstellung der Bewertungsfunktion: manuell (links) und maschinell gelernt (rechts).

Deshalb schlagen wir ein Vorgehen vor, das auf informationstheoretischen Grundlagen beruht und somit objektive Berechnungen liefert [WPSR07]. Wir haben explizit die gewünschten Eigenschaften von Bewertungsfunktionen spezifiziert und gezeigt, dass manuell erstellte Funktionen nicht ideal sind. Unsere neue Methodik lernt Bewertungsfunktionen aus Beispielbildern, die mit Ergebniswerten einer idealen Bewertungsfunktion annotiert sind. Hierbei spezifiziert der Programmierer manuell die korrekten Modellparameter

für eine große Anzahl von Bildern, für welche die Bewertungsfunktion den besten Bewertungswert, also ihr globales Minimum, liefern soll. Weiterhin wird das annotierte Modell schrittweise verschoben und jeweils ein zugehöriger Wert bestimmt. Die korrekten und die verschobenen Modellparameter inklusive der gewünschten Ergebnisse der Bewertungsfunktion stellen die Trainingsdaten für das Lernen der Bewertungsfunktion dar. Die hierbei automatisch gewonnene Bewertungsfunktion ist hochgradig genau und all-gemeingültig, denn sie approximiert die Eigenschaften einer idealen Bewertungsfunktion und eine Vielzahl von Bildern wird zum Training verwendet.

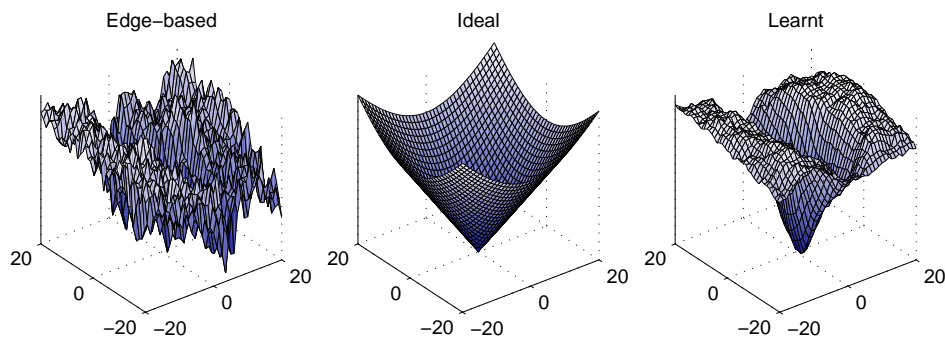


Abbildung 4: Vergleich des Verlaufs der manuell erstellen, einer idealen und der gelernten Bewertungsfunktion, jeweils indem das Modell entlang der x - und y -Achse verschoben wurde.

4 Ergebnisse

Abb. 4 zeigt, wie das Ergebnis der Bewertungsfunktion von den Modellparametern abhängt. Es ist klar ersichtlich, dass die gelernte Bewertungsfunktion näher ans Ideal heranreicht als die manuell erstellte. Die am Rande liegenden Plateaus entstehen dadurch, dass sie außerhalb des Bereichs liegen, für den die Bewertungsfunktion gelernt wurde. In diesen Bereichen liefert die Funktion undefinierte Ergebnisse.

Literatur

- [CT92] Tim F. Cootes and Chris J. Taylor. Active Shape Models – Smart Snakes. In *Proc. of the 3rd British Machine Vision Conference 1992*, Seiten 266 – 275. Springer Verlag, 1992.
- [Ekm72] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symp. on Motivation*, vol. 19, Seiten 207–283, 1972. U of Nebraska Press.
- [KCT00] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Int. Conf. on Automatic Face and Gesture Recognition*, Seiten 46–53, 2000.
- [WPSR07] Matthias Wimmer, Sylvia Pietzsch, Freck Stulp, and Bernd Radig. Learning Robust

Objective Functions with Application to Face Model Fitting. In *Proceedings of the 29th DAGM Symposium*, Heidelberg, Germany, September 2007. to appear.