

Workshop Benchmark

Objekterkennung - Gesichtsdetektion - Gestik - Mimik

IPA Stuttgart, 7. März 2002

Forschungsschwerpunkte von SmartKom und Morpha

Autoren

SmartKom

Nicole Beringer

Dr. Anselm Blocher

Carmen Frank

Dr. Elmar Nöth

Hans Röttger

Rui P. Shi

Dr. Silke Steininger

Morpha

Dr. Carsten Bruckhoff

Markus Ehrmann

Jörg Illmann

Dr. Erwin Prassler

Michael Rinne

1 Einleitung

Im Rahmen des Forschungsschwerpunktes „Mensch-Technik-Interaktion in der Wissensgesellschaft“ werden neben weiteren die Verbundprojekte SMARTKOM und MORPHA gefördert. In beiden Projekten werden, neben weiteren Schwerpunkten, die Forschungsthemen Objekterkennung, Gesichtsdetektion, Gestik und Mimik behandelt.

Dieser Bericht ist das Ergebnis eines Workshops, der am 7. März 2002 am Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA) in Stuttgart stattgefunden hat. In den folgenden Abschnitten werden die Schwerpunkte der Forschung im Bereich Objekterkennung, Gesichtsdetektion, Gestik und Mimik dargestellt. Es werden Unterschiede und Gemeinsamkeiten vorgestellt sowie erste Möglichkeiten der Zusammenarbeit aufgezeigt.

2 Objekterkennung

Das Objekterkennungssystem des Roboters CORA der Ruhr-Universität Bochum verfolgt zwei Ziele: Gelernte Objekte müssen wiedererkannt und von nicht gelernten unterschieden werden; von beliebigen Objekten muss sowohl Ort als auch Orientierung bestimmt werden, so dass diese Objekte gegriffen werden können. Zur visuellen Erfassung der Objekte im Montagebereich dient ein Stereokamerasystem mit einem horizontalen und einem vertikalen Freiheitsgrad.

Zentrale Aspekte der Objekterkennung sind die charakteristischen Farben und die Aufsicht der gelernten Objekte. Hierzu wird aus einem Stereo-Bildpaar zunächst ein Disparitätsbild berechnet und aus diesem die Aufsicht des Objektes. Die Analyse des Farbhistogramms ermöglicht die Bestimmung der charakteristischen Farben eines Objektes. Der Vergleich zwischen gelernter Aufsicht und aktuell wahrgenommener Aufsicht in Verbindung mit einem Vergleich der charakteristischen Farben ermöglicht eine zuverlässige Wiedererkennung gelernter Objekte.

Im Gegensatz dazu dient die Objekterkennung im Projekt SmartKom im Wesentlichen zur Detektion von Schriftstücken. Diese Schriftstücken sollten anschließend bezüglich Bildern oder Text analysiert werden, so dass die Möglichkeit besteht, diese Schriftstücke per Fax oder E-Mail zu versenden. Hierzu ist es ausreichend die Szenen lediglich mit einer Kamera zu beobachten.

Hieraus folgt die gänzlich andere Zielsetzung beider Projekte und die daraus resultierenden verschiedenen Anforderungen an Verfahren und Ergebnisse.

3 Gestenerkennung

3.1 Gestenanalyse in SmartKom

Im Projekt SmartKom ist ein echt multimodales System realisiert, das über die Modalitäten Sprache, Gestik und Mimik als interaktive Benutzereingabe den Benutzer versteht. Bei der Gestik besteht im Vergleich zu den bisherigen Ansätzen zur Gestenerkennung

bzw. -analyse ein wesentlicher Unterschied, der sich im Lauf des Projekts herauskristallisierte: man kann natürliche Gesten anwenden, die man vorher nicht lernen muss. Fast alle gegenwärtigen Arbeiten im Bereich Gestenerkennung bzw. -analyse streben nach einer robusten, effizienten und wenig prozeßintensiven Erkennung, die einem vordefinierten Lexikon unterliegt. Das vordefinierte Lexikon kann z.B. als Kommando dienen, um die Kommunikation zwischen Menschen und Maschine zu ermöglichen und zu erleichtern. Dennoch ist ein solches Lexikon nur quasi-natürlich, denn ein naiver Benutzer hat in den meisten Fällen keinen Zugang zu diesem halbkünstlichen Lexikon, wenn er vor einem echten multimodalen System steht. Natürliche komplexe Gesten der Menschen sind wegen ihres personen- bzw. kulturbedingten Unterschieds kaum erforscht.

Aus diesem Grund ist nur in der Anfangsphase des Projekts SmartKom ein Gestenlexikon definiert worden, in dem Gesten wie Zeigen und Einkreisen stehen. Diese Gesten werden durch den Virtual Touch Screen (SiVit) der Firma Siemens aufgenommen und anschließend durch das Modul Gestenanalyse hinsichtlich der projizierten Benutzeroberfläche analysiert. Die Konzentration in der jetzigen Forschungsphase liegt in der Analyse der möglichen natürlichen Gesten. Natürliche Gesten enthalten viele Informationen über den Benutzer. Was der Benutzer denkt, wie er sich verhält und zu welchem Zeitpunkt er auf das System reagiert, kann allein durch die Sprache oder die Mimik in vielen Fällen nicht identifiziert werden, wenn der Benutzer mit einem multimodalen System kommuniziert. Das dynamische Verhalten der Gesten spielt dabei eine wichtige Rolle und wird etwa durch die Geschwindigkeit, die Beschleunigung, die kinetische Energie und die Varianzamplitude als Merkmale dargestellt, die durch ein Hidden-Markov-Modell (HMM) analysiert werden.

Um detaillierte Informationen über die Gesten zu erhalten, die in einer Mensch-Maschine-Kommunikation mit einem System wie SmartKom vorkommen, werden am Institut für Phonetik und Sprachliche Kommunikation in München Wizard-of-Oz (WOZ) Aufnahmen gemacht. Die Versuchspersonen interagieren mit einem simulierten multimodalen Dialogsystem und können dabei beliebige Gesten einsetzen (d.h. sie wissen nur, daß das System Gesten erkennt, aber nicht welche - auf diese Weise soll eine große Bandbreite möglicher Gesten aufgenommen werden). Die Gesten werden auf einem Video mit einer Ansicht der Geste von oben und einer Seitenansicht des Benutzers bezüglich Anfangs- und Endzeitpunkt markiert und in Kategorien eingeteilt, die sich auf die Intention der Geste beziehen. Unter anderem ergab diese WOZ-Untersuchung, daß der Benutzer meist einem multimodalen System wie SmartKom gegenüber während der Interaktion nur Zeigegesten verwendet. Der obige Ansatz kann zur Analyse des Benutzerzustands eingesetzt werden, um zu entscheiden, ob der Benutzer gerade Entschlossenheit oder Unentschlossenheit zeigt. Das Ergebnis kann zur richtigen Reaktion des Systems auf die Benutzeranfrage beitragen, wenn es mit anderen Modalitäten integriert wird. Im Falle der Unentschlossenheit des Benutzers kann zum Beispiel eine dynamische Hilfe dem Benutzer zur Verfügung gestellt werden. Der bisherige Verlauf der Gestenanalyse ist der folgende:

- Gesten werden durch SiVit erkannt und anschließend an das Modul Gestenanalyse gesendet.

3 Gestenerkennung

- Die aktuelle Disposition vom Display wird im XML bzw. M3L Format an das Modul Gestenanalyse gesendet.
- Die Gestenhypothesen werden an weitere Module wie Medienfusion geleitet, die die Ausgaben aller Modalitäten fusionieren und verarbeiten. Am Schluß wird eine benutzergerechte Systemreaktion produziert.

3.2 Gestenerkennung in Morpha

Die Beschäftigung mit Gesten hat an der Universität Karlsruhe zwei Ziele: Gesten sollen sowohl zur Kommandierung eines Roboters wie auch zur Kommentierung situationsbezogener Charakteristika dienen.

Angewandt werden die Verfahren zur Gestenerkennung zum einen zur Interaktion mit einem Serviceroboter (siehe Abb. 1) und zum anderen in einer speziell aufgebauten Vorführungsumgebung, die zur Aufnahme von Benutzerdemonstrationen im Rahmen des *Programmierens durch Vormachen* dient (siehe Abb. 2). In beiden Szenarien sind neun statische und fünf dynamische Gestentypen verabredet worden (Embleme und Zeigegesten).

Der Roboter soll sich ohne zusätzliche, invasive Sensoren oder Markierungen befehlen lassen. Deshalb werden hier für die Gestenerkennung ausschliesslich bildbasierte Verfahren eingesetzt. Als grundlegendes Merkmal dient die Hautfarbe, mit Hilfe derer die Hände und der Kopf des Benutzers identifiziert werden können. Die Verfolgung einer der beiden Hände dient dann zur Klassifikation von Bewegungsbahnen und zur Klassifikation von Handstellungen auf Basis der Handkontur. Während die Bewegungsbahnen mit Hilfe von Hidden-Markov-Modellen unterschieden werden, dient im zweiten Fall die Beschreibung der Handsilhouette mit Fourier-Deskriptoren als Grundlage der Klassifikation.

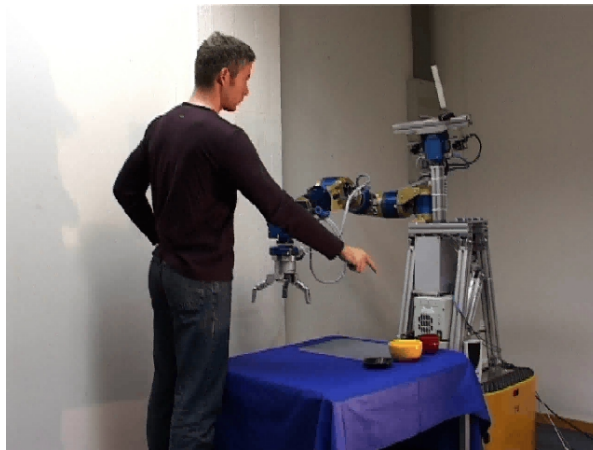


Abbildung 1: Interaktion mit einem Roboter

An dieser Stelle ergeben sich Unterschiede zu den Forschungspartnern im Hinblick auf die Situation (da der Roboterkopf beweglich ist, kann nicht von stabilem Hintergrund

3 Gestenerkennung

ausgegangen werden - die Klassifikation verwendet aus Geschwindigkeitsgründen nur 2D-Information) wie auch im Hinblick auf die Verfahren (Konturkodierung, Bewegungsklassifikation).

Im zweiten Fall trägt der Benutzer einen Datenhandschuh (Abb. 2), der zur Erkennung manueller Eingriffe in die Szene dient. Hier findet ein hierarchischer Klassifikator Verwendung, der auf Basis von neuronalen Netzen und der Betrachtung von Bewegungsparametern der Hand und der einzelnen Fingergelenke eine Entscheidung für eine Griff- oder eine Gestenklasse vornimmt.

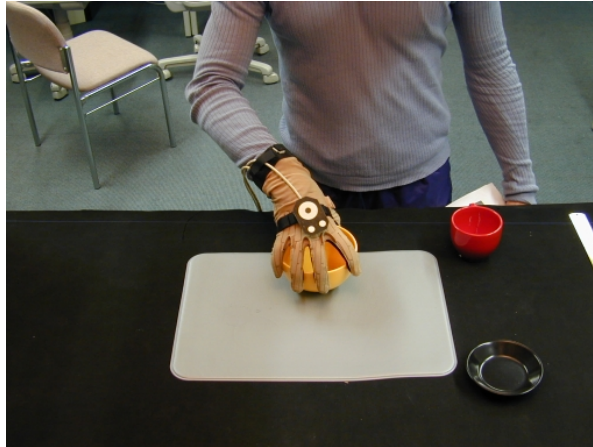


Abbildung 2: Programmieren durch Vormachen

Die Abgrenzung von anderen Forschungspartnern liegt in der Verwendung des Datenhandschuhs als Sensor sowie im darauf zugeschnittenen Klassifikator.

3.3 Detektion und Klassifikation statischer Gesten

Statische Handgesten werden am FAW zur Kommandierung von Robotern genutzt. Statische Gesten stellen eine einfache und intuitive Möglichkeit zur Kommunikation mit dem System dar, die dem Benutzer als Kommunikationsmechanismus geläufig ist. Die Erfassung der Daten erfolgt visuell, so daß keine komplexen technischen Maßnahmen zur Datenaufnahme notwendig sind. Dadurch ist weder eine Veränderung des Umfeldes noch eine Einschränkung des Benutzerverhaltens notwendig. Während Anforderungen an die Umgebungsbedingungen minimiert werden sollen, wird vom Benutzer erwartet, daß er den Satz verwendbarer Kommandierungsgesten und die Wirkung der Gesten kennt.

Am FAW werden Mechanismen untersucht, die die visuelle Erkennung von Handgesten anhand von Einzelbildern und Bildsequenzen erlauben. Dabei werden drei wesentliche Ziele verfolgt: Erkennung von statischen Gesten aus einem definierten Satz von Gesten, Bestimmung von Zeigegesten und angezeigten Objekten oder Bereichen sowie die kontinuierliche Lokalisierung der Hand des Benutzers in der Umgebung. Als Eingangsdaten zur Klassifikation statischer Gesten werden Farbbilder und Stereobilder verwendet. Anhand von Hautfarbensegmentierung und Entfernungssegmentierung werden der Be-

nutzer und die Hand des Benutzers in der Umgebung lokalisiert. Die Gestenerkennung erfolgt dann anhand von formbeschreibenden Merkmalen der Hand.

Die Forschungsschwerpunkte liegen dabei in der robusten kontinuierlichen Detektion von Person und Hand durch Integration unterschiedlicher Segmentierungsverfahren, der Modellanpassung an variable Umgebungsbedingungen und der Klassifikation von Kommandierungsgesten aus einem vorab festgelegten Gestenalphabet.

Zur Evaluierung des Ansatzes wurden sechs Gesten ausgewählt, die unter veränderlichen Umgebungsbedingungen zur Kommandierung eingesetzt werden. Variationen betreffen hierbei Szenenhintergrund und Beleuchtung, die ausführende Person sowie Position und Orientierung der Geste in der Umgebung.

3.4 Abgrenzung der Gestenerkennung in Morpha und SmartKom

Obwohl die Projekte Morpha und SmartKom zur Untersuchung der Mensch-Technik-Interaktion umfangreiche Verfahrensentwicklung und Evaluierung im Bereich der Gestenerkennung betreiben, unterscheiden sich die Forschungsschwerpunkte beider Projekte doch recht deutlich. Im Projekt Morpha werden Handgesten zur Kommandierung und Belehrung von Maschinen eingesetzt. Die verwendbaren Gesten sind vorab definiert und in einem beschränkten Gestenalphabet festgelegt. Daraus ergibt sich die Anforderung an den Benutzer, daß er die ausführbaren Gesten und deren Bedeutung sowie den Kontext der Anwendung kennen muss. Diese Handgesten sollten anschließend automatisch unter Verwendung von Videokameras detektiert und erkannt werden, wobei die Umgebungsbedingungen (Szenenhintergrund, Beleuchtung, vorführende Person) variieren können. Die Menge der zu erkennenden Gesten beschränkt sich auf Kommandierungsgesten. Ziel ist nicht die Erkennung von Emotionen oder Intentionen anhand von Handstellungen oder Bewegungen. Eingabedaten sind Farbbilder, Bildsequenzen und Stereobilder.

Im Gegensatz dazu detektiert und klassifiziert das Projekt SmartKom neben Kommandierungsgesten auch Unterstützungsgesten und emotionale Gesten. Die Klassifizierung der Gesten erfolgt nicht anhand der Morphologie oder der Form der Gesten sondern anhand der Benutzerintention. Ziel ist es hierbei, die natürlichen Gesten automatisch so zu klassifizieren, wie sie ein trainierter Beobachter in den durchgeführten Wizard-of-Oz Experimenten zugeordnet hätte. Die Umgebung zur Gestenerkennung ist festgelegt und gliedert sich in drei unterschiedliche Szenarien: Home, Public und Mobil. Neben der Geste selbst wird von SmartKom auch der Kontext und die Intention des Benutzers analysiert. Als Eingabedaten dienen Infrarotbilder, Einzelbilder und Bildsequenzen. Gemeinsam ist beiden Projekten, dass jeweils nur die Gesten eines Benutzers erkannt werden und die Datenerfassung anhand von visueller Information.

4 Gesichtserkennung und Mimik

Der Bereich der Gesichts- und Mimikerkennung unterteilt sich in die Klassifikation des Benutzerzustandes, der Gesichtsdetektion und die Personenidentifikation. Die folgenden

Abschnitte stellen die Arbeiten beider Projekte auf diesen Gebieten gegenüber.

4.1 Klassifikation des Benutzerzustands

SmartKom SmartKom ist ein multimodales Dialogsystem, das von einem einzelnen Anwender bedient wird. Die Kamera befindet sich am dem Anwender gegenüber liegenden Ende einer waagerechten Bedienoberfläche. Dadurch ergeben sich Bilder, die das Gesicht genau einer Person enthalten und von schräg unten aufgenommen sind. Das System soll den mimischen Ausdrucks des Anwenders erkennen, unabhängig von der Person. Als Datenmaterial sind deshalb Bilder mit mimischen Ausdrücken von möglichst vielen verschiedenen Personen nötig.

Morpha Es gibt keine vergleichbaren Arbeiten in Morpha.

4.2 Gesichtsdetektion

SmartKom Der Hintergrund der Szene ist unbekannt und dynamisch. Es sind jeweils nur einzelne Personen im Sichtfeld. Die Person ist höchstens eine Armlänge von der Kamera entfernt, daraus ergibt sich eine konstante Kopfgröße von ca. 100x150 Pixel.

Morpha Es werden Gesichter der Personen im Blickfeld des Roboters detektiert. Zu den Anforderungen gehört insbesondere die Detektion der Gesichter von Personengruppen, also mehrere Gesichter in einem Bild. Die typische Anordnung der Kamera ist den Randbedingungen der in Morpha betrachteten Szenarien angepasst. Dies kann eine Perspektive in horizontaler Richtung sein, wobei die Kamera etwa in der durchschnittlichen Höhe einer Person montiert ist, oder eine Perspektive aus leicht erhöhtem Blickwinkel. Die minimale Kopfgröße beträgt verfahrensbedingt etwa 40x60 Pixel. Dieser Wert wird jedoch der Anwendung angepasst und daher in der Regel deutlich größer gewählt. Für Morpha ist eine hohe Detektionsrate wichtig, daher wurde das Verfahren vorrangig hinsichtlich der Verarbeitungszeit optimiert.

4.3 Personenidentifikation

Morpha Im Szenario Produktionsassistent ist es wichtig, autorisierte Personen zu erkennen, um die Benutzung durch Unbefugte zu vermeiden. Im Szenario Haushaltsassistent ist eine Trennung der Personen in autorisierte Benutzer und Personen, die sich nur zeitweise im Umfeld des Roboters aufhalten, notwendig. Hierzu wurde von der ZN Vision Technologies AG eine Komponente zur Identifikation der Personen anhand ihrer Gesichter auf die Bedürfnisse von Morpha angepasst und zur Verfügung gestellt.

SmartKom Es gibt keine vergleichbaren Arbeiten in SmartKom.

4.4 Zusammenfassung

In den Bereichen Klassifikation des Benutzerzustandes und Personenidentifikation gibt es in den Projekten Morpha und SmartKom keine vergleichbaren Arbeiten.

Im Bereich der Gesichtsdetektion ergeben sich Gemeinsamkeiten bezüglich der Aufgabenstellung, doch sind die Randbedingungen und Anforderungen sehr unterschiedlich.

Die genutzten Verfahren müssen auf diese Anforderungen zugeschnitten sein, um effizient arbeiten zu können. So ist ein einfacher Vergleich durch die Anwendung der unterschiedlichen Verfahren auf die gleichen Daten nicht möglich, bzw. erlaubt dies keinen fairen Vergleich.

Die Diskussion ergab, dass ein Benchmark für Gesichtsdetektion nicht einfach aus einer Menge von Bildern bestehen kann, sondern vielmehr einer strukturierten Bild-datenbank und eines Kriterienkatalogs bedarf, um die Anwendbarkeit der Verfahren für unterschiedliche Einsatzszenarien anhand von geeignet gewählten Untermengen der Datenbank und entsprechenden Kriterien beurteilen zu können. Ein direkter Vergleich der Verfahren ist jedoch auch dann nur möglich, wenn die Verfahren für Anwendungen mit ähnlichen Randbedingungen entwickelt wurden. Selbst dann bleibt das Problem der Beurteilung der Übereinstimmung zwischen der detektierten Gesichtsregion und der manuell markierten Gesichtsregion. Dies wird weiter erschwert durch die unterschiedlichen Ausgabeformate der Verfahren: manche liefern die Position der Augen und des Mundes, andere berechnen ein umschreibendes Rechteck und wieder andere eine Maske. Je nach Anwendung ist die jeweilige Ausgabe geeignet. Eine Umrechnung auf ein einheitliches Maß würde jedoch zu Fehlern führen, die man nicht den Verfahren zuordnen darf.

Trotz dieser Probleme wäre eine Art Benchmarksammlung für die Beurteilung von Gesichtsdetektionsverfahren hinsichtlich ihrer Anwendbarkeit auf bestimmte Szenarien wünschenswert.

5 Ergebnis des Workshops

Das Hauptergebnis des Workshops ist die Tatsache, daß sich die Forschungsschwerpunkte in den Projekten SmartKom und Morpha deutlich unterscheiden, trotz zum Teil gleicher Benennung.

Im Bereich der Objekterkennung befaßt sich das Projekt SmartKom mit der Erkennung eines Blatts Papier durch eine einzelne Kamera. Ziel ist es hierbei, den Text und die Bilder zu erkennen und diese anschließend per Mail oder Fax zu verschicken. Im Gegensatz dazu liegt die Zielsetzung im Projekt Morpha im Bereich der Erkennung von Gegenständen, die ein Roboter greifen soll. Im Haushaltsbereich sind diese typischerweise Gegenstände des täglichen Lebens wie Teller und Tassen und im Montagebereich Werkzeuge oder Werkstücke. Die Erfassung dieser Objekte erfolgt durch ein Stereo-Kamerasystem oder einen Laserscanner. Somit ergeben sich sowohl Unterschiede bei den Gegenständen als auch bei den genutzten Sensoren.

Im Bereich der Gestenerkennung verfolgt das SmartKom Projekt zunächst eine Erfassung aller Gesten, die ein unbedarfter Nutzer im Umgang mit dem zu realisierenden

Assistenten benutzt. Hierzu ist ein Wizard-of-Oz (WOZ) Szenario realisiert worden, bei dem dem Benutzer ein fertiges System präsentiert wird, welches für den Benutzer unsichtbar durch Menschen gesteuert wird. Hierdurch ist es möglich, Gestik und Mimik des Benutzers zu analysieren und in Kategorien einzuteilen. Die so gewonnenen Daten können anschließend zum Training geeigneter Erkennungsmethoden herangezogen werden, und das System kann auf natürliche Gesten reagieren. Im Gegensatz dazu werden im Morpha Projekt spezielle Gesten definiert mit Hilfe derer der Benutzer den Roboter anlernen oder kommandieren kann. Hierbei muss der Benutzer über den Katalog der Gesten informiert sein und deren Bedeutung in unterschiedlichen Szenarien kennen. An dieser Stelle ergibt sich ein erster Ansatz für eine spätere Zusammenarbeit. Die im Projekt SmartKom extrahierten typischen Gesten könnten in einem späteren Projekt anstelle der vordefiniert Gesten in Morpha genutzt werden, so dass der Benutzer auf natürliche Art und Weise den Roboter anlernen oder kommandieren kann. Dies kann leider erst dann erfolgen, wenn die Klassifikation der möglichen Gesten abgeschlossen ist.

Im Bereich der Gesichts- und Mimikerkennung gibt es lediglich bei der Gesichtsdetektion Gemeinsamkeiten. Die Bestimmung des Benutzerzustandes anhand dessen Mimik wird ausschließlich im SmartKom Projekt bearbeitet, wobei die Identifikation des Benutzers lediglich im Morpha Projekt angegangen wird. Im SmartKom Projekt wird der Benutzer aus annähernd konstanter Entfernung und unter fast konstantem Winkel beobachtet, so dass lediglich ein Benutzer im Bild ist und dessen Gesicht eine konstante Größe und konstante Perspektive hat. Im Gegensatz dazu müssen im Morpha Projekt die Gesichter von Personengruppen identifiziert werden. Eine weitere Schwierigkeit ist hierbei, dass die den Roboter anleitende Person aus unterschiedlichen Winkeln gesehen wird und somit die Gesichter aus verschiedenen Perspektiven betrachtet werden.

Auf Grund der unterschiedlichen Anforderungen und Randbedingungen ist hier ein fairer Vergleich der eingesetzten Methoden nicht möglich. Die rege Diskussion dieses Punktes ergab aber, daß der Bereich der Gesichtsdetektion bereits recht weit fortgeschritten ist und daß die Definition eines geeigneten Benchmarks wünschenswert wäre. Diese Benchmark darf aber nicht einfach aus einer Sammlung von Bildern bestehen, sondern muß eine Gruppe von Szenarien beschreiben, in denen verschiedene Methoden ihre Leistungsfähigkeit beweisen müßten. Hierzu müßten für jedes Szenario die Eingangs- sowie Ausgangsdaten definiert werden und die Randbedingungen angegeben werden, unter denen die Verfahren funktionieren müssen. Nur so können verschiedene Verfahren fair miteinander verglichen werden.