Uni Hamburg – FB Informatik
Proseminar: Künstliche Intelligenz Architekturen
bei: Dr. Cristina Vertan

Architekturen für multimodale Anwendungen (SmartKom)

Ein Referat von Tatjana Lorenzen, So-Mei Lai, Wanja Johannes Slawski

Gliederung:

Einführung und Integration von Gestik und Mimik (*Wanja Johannes Slawski*) Seite 3
Einführung Seite 3
Das SmartKom-Projekt Seite 3
SmartKom als multimodales Dialogsystem Seite 3
Momentan mögliche Applikationsfelder Seite 3
SmartKom-Mobile Seite 4
SmartKom-Home/Office Seite 4
SmartKom-Public – Ein multimodaler Kommunikationskiosk Seite 5
Integration von Sprache, Gestik und Mimik Seite 6
Modalitätenfusion im Kontext von SmartKom Seite 6
Das multimodales Dialoggedächtnis Seite 7
Der Kontrollbildschirm von SmartKom Seite 7

SmartKom: Gestik (*Tatjana Lorenzen*) Seite 8
Einführung Seite 8
WIZARD-OF-OZ Seite 8
Die erste Phase des Projektes Seite 9
Die Fragen bei der Klassifikation von Gesten: Seite 9
Überblick über Codierungskonzept Seite 9
Definition von Gesten Seite 10
Probleme und zukünftige Arbeit Seite 11

SmartKom: Mimik (*So-Mei Lai*) Seite 11
Einführung Seite 11
Der Gesichtsausdruck Seite 11
Erkennen von Gesichtsausdrücken Seite 12
Beispiel: Erkennung von Ironie und Sarkasmus Seite 14
Anwendung Seite 14

SmartKom: Einführung und Integration von Gestik und Mimik (Wanja Johannes Slawski)

Einführung:

Das folgende Referat behandelt das Thema „Architekturen für multimodale Anwendungen“ am Beispielprojekt SmartKom.

Das SmartKom Projekt

Der Hauptakteur im SmartKom Projekt ist das „deutsche Forschungszentrum für künstliche Intelligenz“ kurz DFKI. Gefördert wird das Projekt von der Bundesministerium für Bildung und Forschung kurz bmb+f als direktes Nachfolgeprojekt von Verbmobil. Außerdem sind mehrere Firmen am Projekt beteiligt wie z.B. Sony, Siemens, Compaq, DaimlerCrysler u.a. die das Projekt mitgestalten und teilweise benötigte Hardware dem Projekt angepasst entwickeln wie z.B. der Virtual Touchscreen von Siemens oder der iPAQ Pocket PC von Compaq.

SmartKom als multimodales Dialogsystem

Der Begriff „Multimodales Dialogsystem“ beinhaltet die Integration von natürlicher Sprache, Gestik und Mimik sowohl für den Input als auch für den Output des Systems. Die Grundlage für die Verarbeitung von natürlicher Sprache stellt das Vorgängersystem Verbmobil dar.
Die Ausgabe des Systems geschieht über den Interface-Agent oder auch virtual-communication-assistant in diesem Projekt genannt “Smartakus” (Abb.1). Smartakus interagiert mit dem Benutzer, indem er spricht, Gesichtszüge (Mimiken) zeigt und gestikuliert.
Unterstützt wird der Interface-Agent über spezifische, also der Situation angepasste bzw. der Anfrage entsprechende, visuelle Darstellungen.

Abb. 1 - © by smartkom. org

Momentan mögliche Applikationsfelder

Momentan sind drei verschiedene Applikationsfelder angedacht und bereits als Prototypen realisiert worden, dazu gehören SmartKom-Mobile, SmartKom-Home/Office und SmartKom-Public. Alle basierend auf dem selben Systemkernel. Die drei Bereiche können auch Geräte übergreifend genutzt werden. So kann z.B. eine am SmartKom-Home/Public angefangene Applikation am SmartKom -mobile fortgeführt werden. Eine Übersicht zeigt die folgende Grafik (Abb.2):

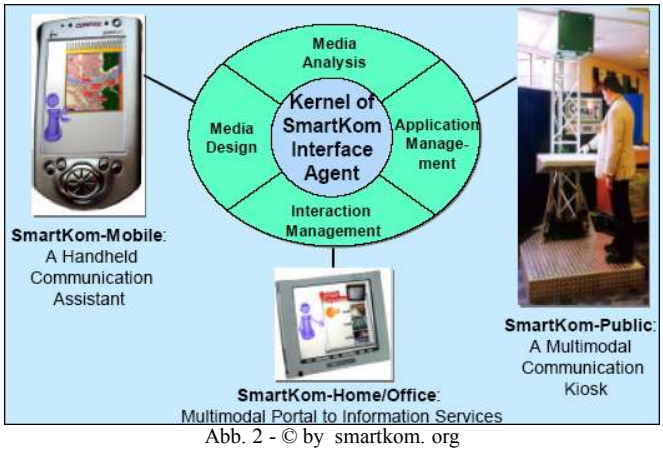

Abb. 2 - © by smartkom. org

SmartKom-Mobile


Abb. 3 - © by smartkom. org

Die Hardware-Basis für SmartKom-Mobile stellt Compaq mit dem „iPAQ Pocket PC“, einem tragbaren mini PC (Abb.3). Der Pocket PC bietet eine Grundlage für momentan zwei Anwendungsbereiche.
Er kann mit dem Navigationsystem eines Autos gekoppelt werden für eine individuelle Routenplanung über GPS und andere Systeme. Einen Prototypen dafür gibt es als Mercedes von DaimlerCrysler.
Die Routenplanung ist aber mit dem Verlassen des Automobils nicht zu Ende, so kann man den Pocket-PC auch mit sich tragen, wobei die im Auto begonnene Anwendung fortgeführt werden kann. Auf diese Weise bietet SmartKom-Mobile auch die Möglichkeit eines individuellen Stadtrundganges. Durch eine Anbindung ans Internet kann der Pocket PC auch Informationen über Sehenswürdigkeiten einer Stadt anzeigen.
Gesteuert wird es sowohl über Sprache, als auch über pen-based -pointing. Die Ausgabe erfolgt über eine vereinfachte Smartakus-version (ohne Rumpf, also nur Gesicht und Hände), beinhaltet also Sprache, Gestik und Mimik und visuelle Darstellungen wie z.B. Stadtpläne, Landkarten, Fotos von Objekten und Informationstexten.

SmartKom-Home/Office


Abb. 4 - © by smartkom. org

SmartKom-Home/Office wurde über den „Fujitsu Stylistic 3500X portable webpad“ (Abb.4) realisiert. Es kann zum Beispiel genutzt werden, gekoppelt an diverse Elektrogeräte wie z.B. TV, Video, DVD etc., um diese zu kontrollieren und zu programmieren. Durch die Kopplung ans Internet sind „Tasks“ möglich wie Beispielsweise, Abfrage des Fernsehprogrammes und anschließende Programmierung des Videorecorders zum ausgesuchten Film. Desweiteren sind auch allgemeine Web-dienste, E-Mail Verwaltung und Telefonieren über das tragbare webpad realisiert worden.
Die Bedienung kann über zwei Modi geschehen: lean-forward oder lean-back. In der lean-forward Methode kann der Benutzer auf Sprache und Gestik mittels eines „Zeige-Stiftes“ zurückgreifen. Die Ausgabe des Systems erfolgt dann über Smartakus, also Sprache, Gestik, Mimik und ebenfalls ünterstüzende visuelle Darstellungen. In der lean-back Methode erfolgt die Eingabe sowie die Ausgabe rein sprachlich.

SmartKom-Public – Ein multimodaler Kommunikationskiosk

Eine Übersicht zeigt folgendes Schemata (Abb.5) Danach folgt eine kurze Erklärung einiger ausgwählten Einzelheiten :

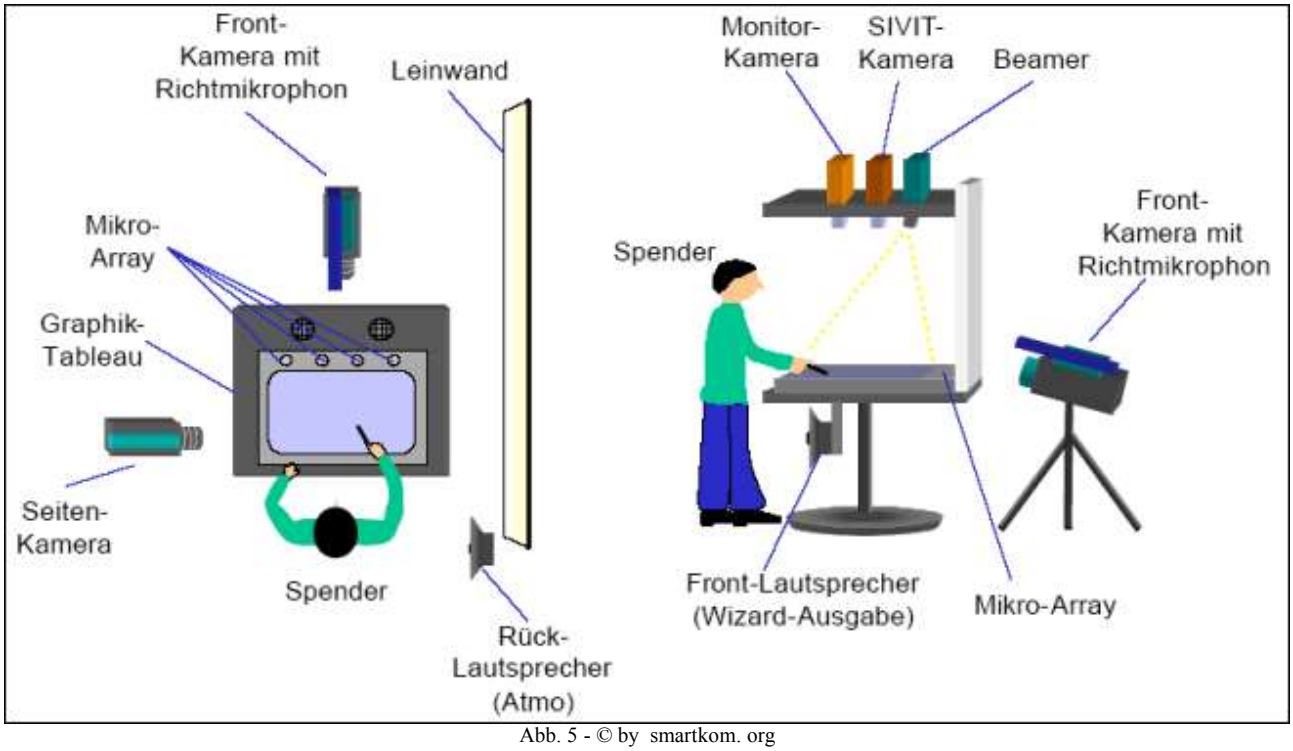

Abb. 5 - © by smartkom. org

Übersicht:

Frontkamera mit Richtmikrophon	Die Frontkamera mit Richtmikrophon ist auf das Gesicht des Benutzers hier Spender ausgerichtet, um die Mimiken des Spenders über die Kamera und mittels des Richtmikrophons die Spracheingabe aufzuzeichnen.
Graphiktableau	Das Grafiktableau dient zur visuellen Darstellung und als Interaktionsfläche.
Beamer	Der Beamer ist auf das Graphiktableau gerichtet und projeziert die visuellen Darstellungen.
Monitorkamera	Die Monitorkamera fotographiert Objekte die auf die auf das Graphiktableau positioniert werden können.
Seitenkamera	Die Seitenkamera zeichnet die Gesten des Spenders auf.
SIVIT-Kamera	Die SIVIT-Kamera dient der Aufzeichnung der Koordinaten von deutenden Gesten.
Lautsprecher	Die Lautsprecher geben die akustische Ausgabe des Systems wieder.

Der SmartKom-Kumminationskiosk ist für öffentliche Einrichtungen gedacht. So soll mittels des Kommunikationskiosk z.B. die Kartenreservierungen in Kinos, oder das Buchen und Reservieren von Tickets in Flughäfen und Bahnhöfen ermöglicht werden. Mit SmartKom-publicist ebenfalls Telekomunikation möglich, was die Idee der Aufstellung von Smartkom-Kiosken anstelle von Telefonzellen aufbringt. Außerdem ist ein Zugang zu persönlichen Webdienste realisiert worden, welche mit dem Sicherheitsaspekt von biometrischen Abfragen gekoppelt wurde.

Integration von Sprache, Gestik und Mimik

allgemeine Modalitätenverarbeitung

Die Leitvorstellung bei der Umsetztung von jeglichen Anwendungen, Anfragen und Modalitäten ist ein einheitlicher Verarbeitungsansatz mit folgenden Grundsätzen.
Es soll keine Speziallösungen für definierte Anwendungen geben, sondern eine Modalitätenverarbeitung mittels übergreifender, wissensbasierter parametisierten Ansätzen. So soll eine Basis für verschiedenste Applikationsfelder geschaffen werden.
Zudem soll für alle Ein- und Ausgaben ein einheitlicher, allgemein im System verwendeter Repräsentationsformalismus verwendet werden, um die Kommunikation aller Module untereinander zu ermöglichen.Desweiteren basiert das System auf einer Multi-Blackboard-Systemarchitektur, welche eine Weiterentwicklung der Integrationsmiddleware von Verbmobil darstellt. Wie die einzelnen Benutzerintentionen zu einer Hypothese kombinert werden, wird im folgenden Abschnitt anhand Sprache und Gestik näher erläutert.

Modalitätenfusion im Kontext von SmartKom

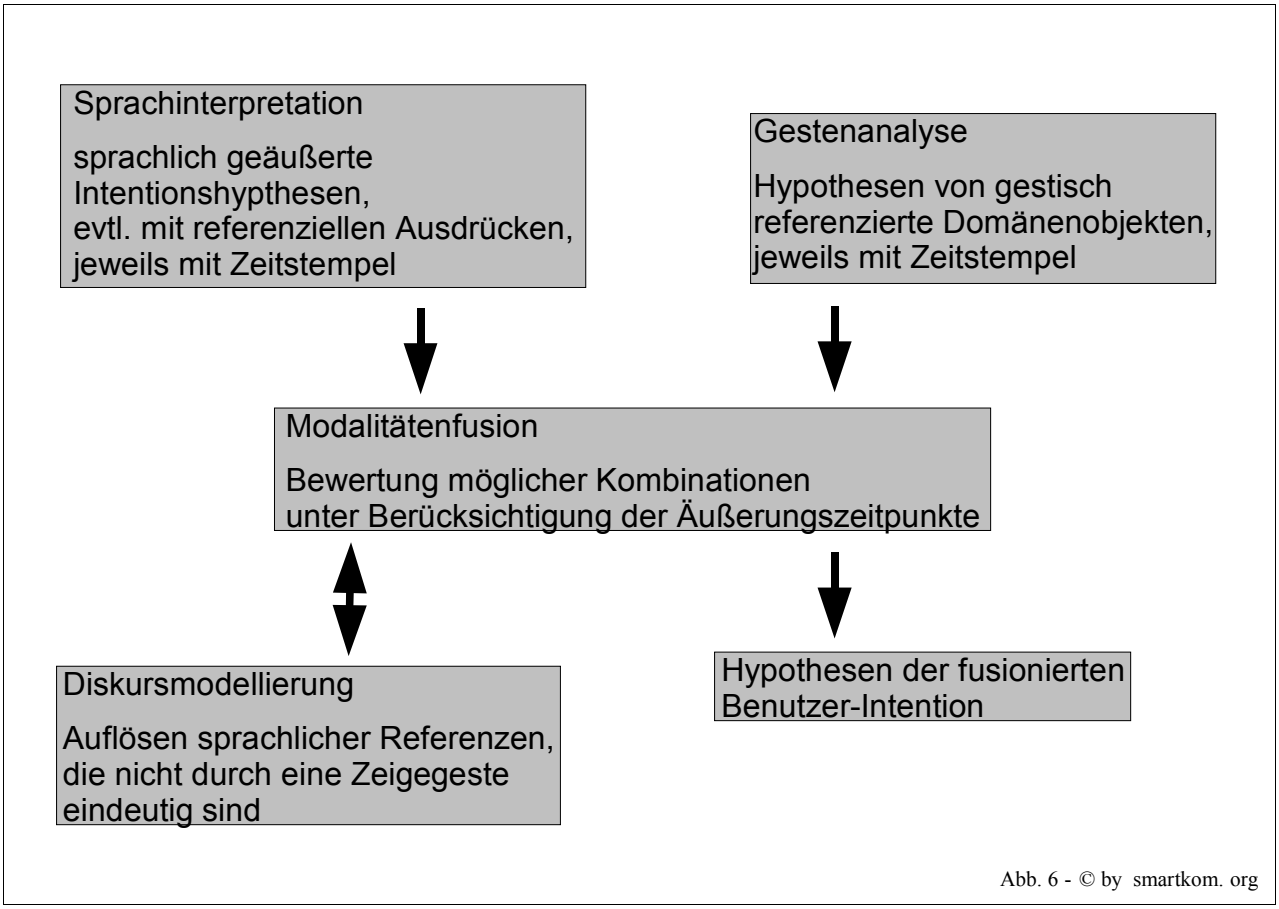

Abb. 6 - © by smartkom. org

Das Schaubild in Abb.6 zeigt einen Ausschnitt der Modalitätenfusion im Kontext von SmartKom. Es werden vorerst die Sprache und die Gesten getrennt analysiert und interpretiert. Die Hypothesen werden danraufhin zusammen mit dem Äußerungszeitpunkt an das Modalitätenfusionsmodul übergeben . Dort werden die eingeganen Daten unter Berücksichtigung der Äußerungszeitpunkte und der Kommunikation mit dem Diskursmodellierungsmodul kombiniert und bewertet. Damit wird eine Hypothese über die Benutzerintention erstellt und an das Aktionsplanungsmodul (hier nicht dargestellt) übergegeben.

Das multimodales Dialoggedächtnis

Das multimodale Dialoggedächtnis arbeitet wie im Schaubild Abb. 7 zu sehen auf drei Ebenen. Die unterste stellt die Modalitätsebene dar. Hier werden linguistische und gestische Objekte gespeichert. Linguistische Objekte sind z.B. „heute abend“ oder „Fernsehprogramm“ und gestische z.B. „[->]“, was eine Referenzierung zu einem Objekt auf dem Graphiktableau hat wie eine z.B. eine „Liste des Fernsehprogrammes“.
Auf der Diskursebene können dann Referenzen auf mehrere Modalitätsobjekte in einem Diskursobjekt gespeichert werden, wie z.B. das gestische Objekt„Liste des Fernsehprogrammes“ und das linguistische „Fernsehprogramm“. Dies ermöglicht zum Beispiel den sprachlichen Bezug auf ein gestisches Objekt ohne eine Geste benutzen zu müssen. Die Anwendungsebene enthält mehrere Diskursobjekte die in sogenannten Dömenenobjekten zusammengefaßt referenziert werden. Eine Domänenobjekt enthält Sinneinheiten wie eine Anfrage an das System z.B. „Zeig mir das Fernsehprogramm von heute abend!“ oder auch Ausgaben „Hier [->]sehen Sie das Programm von heute abend.“

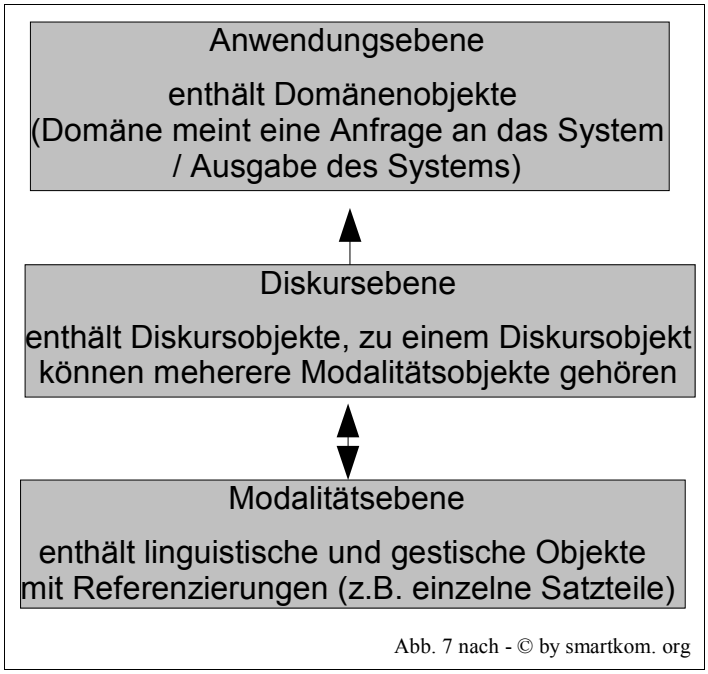

Abb. 7 nach - © by smartkom. org

Der Kontrollbildschirm von SmartKom

Die Abbildung 8. zeigt den Kontrollbildschirm von SmartKom. Er dient als Übersicht und zur graphischen Darstellung aller Vorgänge zur Laufzeit im Ablauf des SmartKom- Programmes. Es zeigt die Verknüpfungen der einzelnen Module miteinander und zur Laufzeit die aktiven Module.
Der rot eingekreiste Bereich zeigt Abschnitt der unter Punkt „Modalitätenfusion im Kontext von SmartKom“ erläutert wurde.

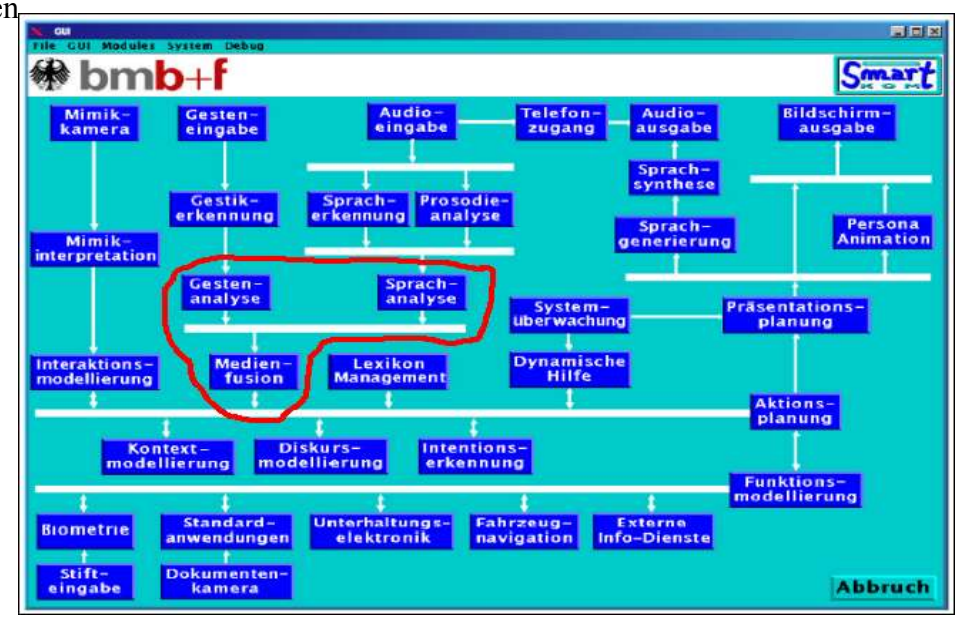

Abb. 8 - © by smartkom. org

SmartKom: Gestik (Tatjana Lorenzen)

Einführung

Neben der Sprache kann der Benutzer auch als Ergänzung die Geste benutzen, um mit SmartKom zu „sprechen".
Bei der Gestik besteht im Vergleich zu den bisherigen Ansätzen zur Gestenerkennung bzw. -analyse ein wesentlicher Unterschied: man kann natürliche Gesten anwenden, die man vorher nicht lernen muss. Fast alle gegenwärtige Arbeiten in Bereich Gestenerkennung bzw. -analyse streben nach einer robusten, effizienten und wenig prozeßintensiven Erkennung, die einem vordefinierten Lexikon unterliegt. Es kann z.B. also Kommando dienen, um die Kommunikation zw. Mensch und Maschiene 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 multimodallen System steht.

Auf diesem Grund ist nur in der Anfangsphase des Projektes SmartKom ein Gestenlexikon definiert worden, in dem Gesten wie Zeigen und Einkreisen stehen. Diese Gesten werden durch den Virtual Touch Screen (SiVit) 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 vilen Fällen nicht identifiziert werden, wenn der Benutzer mit einem multimodallen 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.

WIZARD-OF-OZ

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 multimodallen 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 multimodallen System wie SmartKom gegenüber während der Interaktion nur Zeigegesten verwendet. Das kann zur Analyse des Benutzerzustands eingestezt werden, um zu entscheiden, ob der Benutzer gerade Entschlossenheit oder Unentschlossenheit zeigt. Das Ergebnis kann zur richtigen Reaktion des System auf die Benutzeranfrage beitragen, wenn es mit anderen Modalitäten integriert wird. Im Falle der Unentschlossenheit des Benutzers kann zum Beispiel zur Verfügung gestellt werden.

Die erste Phase des Projektes

Die gesammelte Daten wurden für 3 verschiedene Hauptzwecke benutzt:

- für Training von Sprache, Gesten und Erkennung von Emotionen
- für Entwicklung die Benutzer-, Sprache-, Dialog-Modellen usw. und Sprachsynthese Modul
- für allgemeine Bewertung des Verhaltens von Benutzern, während Interaktion mit der Maschine

Bei jedem Wizard-of-Oz Versuch wurden spontane Sprache, Gesichtsaudrücke und Geste auggenommen. Dafür hat man folgendes benutzt:

- eine digital Kamera, die Gesichtsaudrücke erfasste
- Gesten wurden mit der 2. digitalen Kamera aufgenommen. Sie nahm einen Seitenansicht des Benutzers auf
- eine infrarot empfindliche Kamera, die Handgeste (2-dimensional) und die graphische Ebene aufgenommen hat
- die Koordinaten von deutenden Gesten auf dem Display wurden mit SiVit aufgenommen
- und auch Eingaben, die mit dem Stift auf der graphische Ebene gemacht wurden, wurden auch aufgenommen

Aufnahmen, die für Gestencodierung relevant sind, sind die 2-dimensionale schwarz-weiße Bilder von infrarote Kamera, Seitenansicht des Benutzers und Ausgaben des Beamers (Displayansicht). Um Codierung zu vereinfachen sind Ausgaben von Beamer und infrarot empfindliche Kamera überlappt. Auf dieser Weise hat Codierer einen Seitenansicht, Blick von oben und Information auf welchen Teil des Displays sind die Hände gerichtet.

Die Fragen bei der Klassifikation von Gesten:

- welche funktionale Geste kann man identifizieren (z.B. „deutend“, „nein“, „zurück“ usw.)
- welcher Weg der beste Weg ist, um beobachtete Elemente zu kategorisieren?(welche einfache Gesten können kombiniert werden?)
- welcher Weg der beste Weg ist diese Elemente zu beschreiben?
- welche klare morphologische Form haben die Elemente?
- Wie kann man Anfang und Ende von gegebenem Element definieren?
- Ist ein hinweisendes Wort im Audio enthalten, wenn ja welches dann?(z.B. „dieser“, „hier“, „nein“)

Überblick über Codierungskonzept

Label, der zu einer von 3 Kategorien gehört, wird zu jeder identifizierter Geste zugewiesen.
Label wird von eigene „modifizierer“ ergänzt. Die 3 Modifizierer weisen auf Zeit (Anfang, Ende, Takt) hin, während andere 3 weisen auf Inhalt (morphologie, Bezugswort, Bezugsobjekt) hin. Wenn nötig ist wird identifizierte Geste durch einen Kommentar spezifiziert.
Anfang und Ende sind als Zeitpunkte und Takt als Zeitperiode definiert. Bezugswort ist eine Anmerkunng, d.h. das Wort, das der Geste im Audiokanal entspricht wird zetiert.

Begriff „Geste“ bezeichnet eine Folge aus Segmenten, die aus Arm- und Handbewegungen bestehen. Segmente sind in verschiedene Kategorien unterteilt.

Ist der Geste eine interaktion mit dem System („Interaktionsgeste“), eine Unterhaltung mit sich selbst („unterstützende Geste“) oder was anderes („übrige Geste“).
Nur die Geste, die geschieldert (labeld) sind legen so genannten „cubus“ fest. „Cubus“ sind die Felder, die auf dem Display und Raum über dem Display liegen, wo SiVit die Daten aufnimmt.

Definition von Gesten

Es sind 3 Kategorien von Gesten definiert. Interaktionsgeste (I-Geste), unterstützende Geste (U-Geste) und übrige Geste (R-Geste). Das Kriterium für diese Aufteilung ist das Ziel des Benutzers.

Die I-Geste sind konstruktiv, d.h. sie sind die Bedeutung von Interaktion mit dem Computer. Wenn Benutzer macht Hand/Armbewegung um eine Kommande dem Computer zu geben, dann heißt diese Geste Interaktionsgeste. Als Kommande betrachtet man jede Bitte an System. Auch wenn Computer kann die Kommande nicht ausführen, bleibt bittende Geste trotzdem eine Interaktionsgeste. 2. Typ der Interaktion ist die Bestätigung auf eine Frage.
Unterstützende Geste sind auch konstruktiv. Sie kommen in der Phase in welcher Bitte schon abgeschlossen ist vor. Sie bedeuten gestikallische Unterstützung von „Solo-Aktion“ des Benutzers (wie lesen od. Suchen). Unterstützendegeste Geste enden mit dem Ende von Kommanden nicht. Wenn noch Unterstützende Geste eine Kommande kommt dann wird diese Kommande getrennt von U-Geste als Interaktionsgeste geschieldert.
Übrige Geste Diese Kategorie beinhaltet alle Geste die zu Cubus gehören aber sind nicht in einer von beiden vorherigen Kategorien zugeordnet. Einige Geste, die außerhalb von Cubus geschieldert sind, gehören auch zu diese Kategorie. Übrige Geste sind nicht konstruktiv. Das sind keine Bitte und keine Bestätigung. Übrige Geste sind emotionale Geste wie auch unbekannte Geste.

Labels

Label bedeutet exakten Typ von Gesten, d.h. er bezieht sich auf funktionalität der Geste im Interaktionsprozess. Es gibt 3 Kategorien von Labels: F-, U- und R-. Folgende F-, U- und R-Labels existieren: F-Kreis (+), F-Kreis (-), F-Punkt (long+),F-Punkt (long-), F-Punkt (short+), F-Punkt (short-), F-Frei (free); U-Kreis(read), U-Kreis(search), U-Kreis(count), U-Kreis (ponder), U-Punkt (read), U-Punkt (ponder); R-emotionell (+cubus), R-emotionell (-cubus),R- unbekannt (+cubus).
Beschreibung des Labels

F-Kreis (+) beschreibt eine ununterbrochene Bewegung mit der Hand. Bewegung spezifiziert einen Objekt auf dem Display durch einkreisen. Display ist berührt. Es muss nicht unbedingt eine Kreis-Bewegung sein. Aber gut gerichtete Bewegung. Blick des Benutzers ist auf gewählten Region gerichtet.
(-) - bedeutet, dass Display ist nicht berührt.
Long – einen gedehnten Takt der Bewegung.
Short – sehr kurzen Takt der Bewegung.

U- Labels haben im allgemeinen gleiche Merkmale. Sie dienen aber nicht zum Auswählen vo Objekten sonder nur zum unterstützen, wie z.B. beim Lesen Text entlang mit der Hand führen.

Probleme und zukünftige Arbeit

Daargestellter Konzept ist nur der erste Schritt im Entwicklungsprozess. Nächste Schritt ist Bewertung und Verbesserung des Konzeptes. Dabei sollen Folgende Punkte beachtet werden:

- Qualitätsmaß soll festgesetzt werden.
- Zeit: bis jetzt sind Zeitspannen nur unklar definiert.
- Takt: es ist jetzt noch nicht klar wie man Takt zuverlässig defenieren kann.
- Zusätzliche Labels: Gesten, die beobachtet werden können sind stark von Displaybenutzung beinflusst. Deswegen soll man zusätzliche Labels einfügen.

Der Konzept ist wahrscheinlich immer noch zu komplex, um Bedürfnise schnellen Systems zu erfüllen. Aber die aufgetauchte Kategorien sind gute Kandidate für selektive Labels und werden auf ihre Qualitätskriterium in der 2. Schritt des Entwicklungsprozess testiert.

SmartKom: Mimik (So-Mei Lai)

Einführung

Dialogsysteme sind für normale Benutzern (keine Fachleute) konstruiert. Diese Benutzer wollen keine lange Beschreibung über die Funktionalitäten durchlesen. Deshalb soll zur Verbesserung des Dialogsystems die Mensch-Maschine Dialoge in Mensch-Mensch Dialoge verändert werden.
Aber wie soll ein Mensch -Mensch Dialog aussehen?
Während eines Gesprächs werden viel mehr Input-Informationen benutzt als nur die Sprache:
- Ohren, um Wörter und Stimmausdrücke zu wahrzunehmen
- Augen, um Körperbewegungen und Gesichtsmuskeln zu erkennen
- Nase, um zu riechen, wo jemand gewesen war
- Haut, um physischen Kontakte zu erkennen

Im folgenden wird nur auf Gesichtsausdrücken konzentriert.

Der Gesichtsausdruck

Gesichtsausdrücke sind nicht nur der emotionale Zustand des Benutzers (Hass, Liebe und Furcht), auch der innere Zustand oder man bezeichnet es Benutzerzustand (Hilflosigkeit oder Ärger) beeinflusst die Weiterentwicklung der Mensch-Maschine Dialog.

Die Idee für das Erkennen von Gesichtsausdrücken ist es, so früh wie möglich Benutzer, die ärgerlich werden, zu erkennen, um die Dialogstrategien vom System abzuändern und mehr Unterstützung zu geben.

Dies verhindert den Benutzern enttäuscht zu werden und, dass sie das System nie wieder benutzen wollen.
Das System beobachtet das Gesicht, um den Benutzerzustand zu erkennen.
Den emotionalen Zustand einer Person zu bestimmen ist die Aufgabe für das Erkennen von Gesichtsausdrücken

Erkennen von Gesichtsausdrücken

Das Modul arbeitet in zwei Schritten:
- Das Gesicht der Benutzer muss erst lokalisiert werden.
- Dann wird der mimischen Ausdruck des Gesichts klassifiziert.

Lokalisation:
Für die Lokalisation werden alle nicht hautfarbenen Bereiche ausgeblendet.
In den übrigen Bereichen wird mit einem auf Gesichter trainierten Klassifikator die Position mit der größten Übereinstimmung gesucht.
Die Vorteile der Hautfarbensegmentierung sind es, dass der Suchraum reduziert ist, und dass gesichtsähnliche Texturen im Hintergrund eliminiert sind.

Klassifikation des Gesichtsausdruck:
Hier wird ebenfalls von Klassifikatoren durchgeführt.

Verschiedene Gesichtsaudrücke von mehrere Personen werden in Klassen eingeteilt und in ein Datenbank gespeichert. Dieser Datenbank bietet ein Maß dafür, wie gut ein zu klassifizierendes Bild modelliert werden kann.

Beispieltypen von Gesichtsausdrücken

Es gibt folgende Kategorien von Gesichtsausdrücken:
- Freude
- Ärger
- Hilflosigkeit
- nachdenklich
- überrascht
- neutral

Nach folgenden Kriterien kann das System erkennen, zu welchen Kategorien bestimmte Gesichtsausdrücke gehören.

Die Kriterien um den Gesichtsausdruck **Freude** zu erkennen sind z.B.:
- der Benutzer lacht oder lächelt

- Mundwinkel bewegt sich nach oben.
- Augen sind meistens offen.
- Augenbrauen bewegen sich nach oben
- Zähne sind sichtbar
- Stimme ist meist höher und/oder lauter
- Lachen und freundliche Stimme

Die Kriterien um den Gesichtausdruck **Ärger** zu erkennen sind z.B.:

- Lippen zusammen gepresst
- Stirn runzeln
- geschlossene Augen
- Kopf schütteln
- Benutzer spricht langsamer
- schreit oder spricht sehr deutlich
- Pause zwischen Wörter
- tiefer Stimme
- Augenbrauen zusammengekniffen

Die Kriterien um den Gesichtsausdruck **nachdenklich** zu erkennen sind z.B.:

- Stirn runzeln
- An den Lippen beißen
- Langsame Bewegung
- Mund teilweise offen
- auf die Decke schauen
- zögern
- stottern
- leises Gemurmel

Die Kriterien um den Gesichtsausdruck **Hilflosigkeit** zu erkennen sind z.B.:

- Stirn runzeln
- Falte am Stirn sichtbar
- offener Mund
- zögern
- stottern
- Augenbrauen bewegen sich nach oben

Das Problem hier ist die leichte Verwechslung von „nachdenklich“, da sie ähnliche Merkmale haben.

Sie zu unterscheiden gibt es folgende Kriterien:

1. Frage: „ Wie fühlt sich der Benutzer?“
   Antwort: „unter Kontrolle“ → Kategorie „nachdenklich“
   Antwort: „außer Kontrolle“ → Kategorie „hilflos“

2. Konzentration auf Gesichtteile:
   Mundbereich → „nachdenklich“
   Augenbrauen, Stirn → „hilflos“

Allgemeine Probleme:
Das System kann die Intensität der Emotion nicht erkennen.
Die Beurteilung von Emotionen kann schwer in Kategorien eingeteilt werden, da manche Kriterien mehrdeutig sein können.
Z.B kann der Merkmal „Kopf schütteln“ als ärgerlich oder auch hilflos angesehen werden.

Beispiel: Erkennung von Ironie und Sarkasmus

(1) Smartakus: Hier sehen Sie die Übersicht zum heutigen ZDF-Programm.

(2) Benutzer: Echt toll.

(3) Smartakus: Ich zeige Ihnen alternativ das Programm eines anderen Senders.

(2') Benutzer: Echt toll.

(3') Smartakus: Welche Sendungen wollen Sie aus dem ZDF-Programm sehen oder aufzeichnen?

Smartakus gibt ein Programm aus.
Der Benutzer sagt „Echt toll.“, was eine positive Merkmal ist.
Aber seine Mimik wird als Negativ angesehen.
Smartakus interpretiert deshalb die Sprache hier als ironisch oder sarkastisch und sucht weiter nach Programmen bis der Benutzer ein zufriedener Ausdruck gibt.

Anwendung

Ein Anwendungsbeispiel ist wie vorheriges Beispiel ein Fernsehprogramm.

In den Versuchen erscheinen die Benutzer trotz wegen der Kamera nicht verwirrt zu sein.

Doch es gibt keine Ergebnisse über die emotionalen Zustandsinformation bei der Benutzung.
Der Grund dafür ist, dass die Benutzer nicht über diese Funktionalität wussten, und dass Gesichtsausdrücke den Dialog beeinflussen kann.

Beurteilung von Benutzer, die zum ersten mal SmartKom benutzen:

- SmartKom ist einfach zu benutzen.
- Smartkom macht Spaß, weil es was Neues ist.
- Es ist interessant, lustig und unkompliziert.
- Benutzer sind begeistert, dass sie vom System angesprochen werden.

Im Allgemeinen sind die Benutzer mit dem System zufrieden und werden es gerne weiterbenutzen.