Architekturen für multimodale Anwendungen (SmartKom)

Vortrag von So-Mei Lai, Tatjana Lorenzen, Wanja J. Slawski

Gliederung:

- Einführung
- Integration von Gestik und Mimik
- Gestik
- Mimik

Teilbereichsgliederung:

- allg. Vorstellung des Smartkom-Projektes (Folien 4 – 11)
  - allgemeines (Folien 4 – 5)
  - “Smartakus” (Folie 6)
  - momentan mögliche Applikationsfelder (Folien 7 – 11)

- Integration von Sprache, Gestik und Mimik (Folien 12 – 14)
  - allg. Modalitätenverarbeitung (Folie 12)
  - Modalitätenfusion im Kontext von Smartkom (Folie 13)
  - multimodales Dialoggedächtnis (Folie 14)
  - Kontrollbildschirm (Folie 15)

Das SmartKom – Projekt

Der Hauptakteur ist das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)

Gefördert wird das Projekt von der Bundesanstalt für Bildung und Forschung (BMBF) als Nachfolgeprojekt des Verbmobils.

Zudem sind mehrere Firmen und Forschungseinrichtungen am Projekt beteiligt (u.a. Siemens, Sony).

SmartKom als multimodales Dialog System

umfasst die Bereiche

- Gestik
- Mimik
- Sprache (Grundlage Verbmobil)

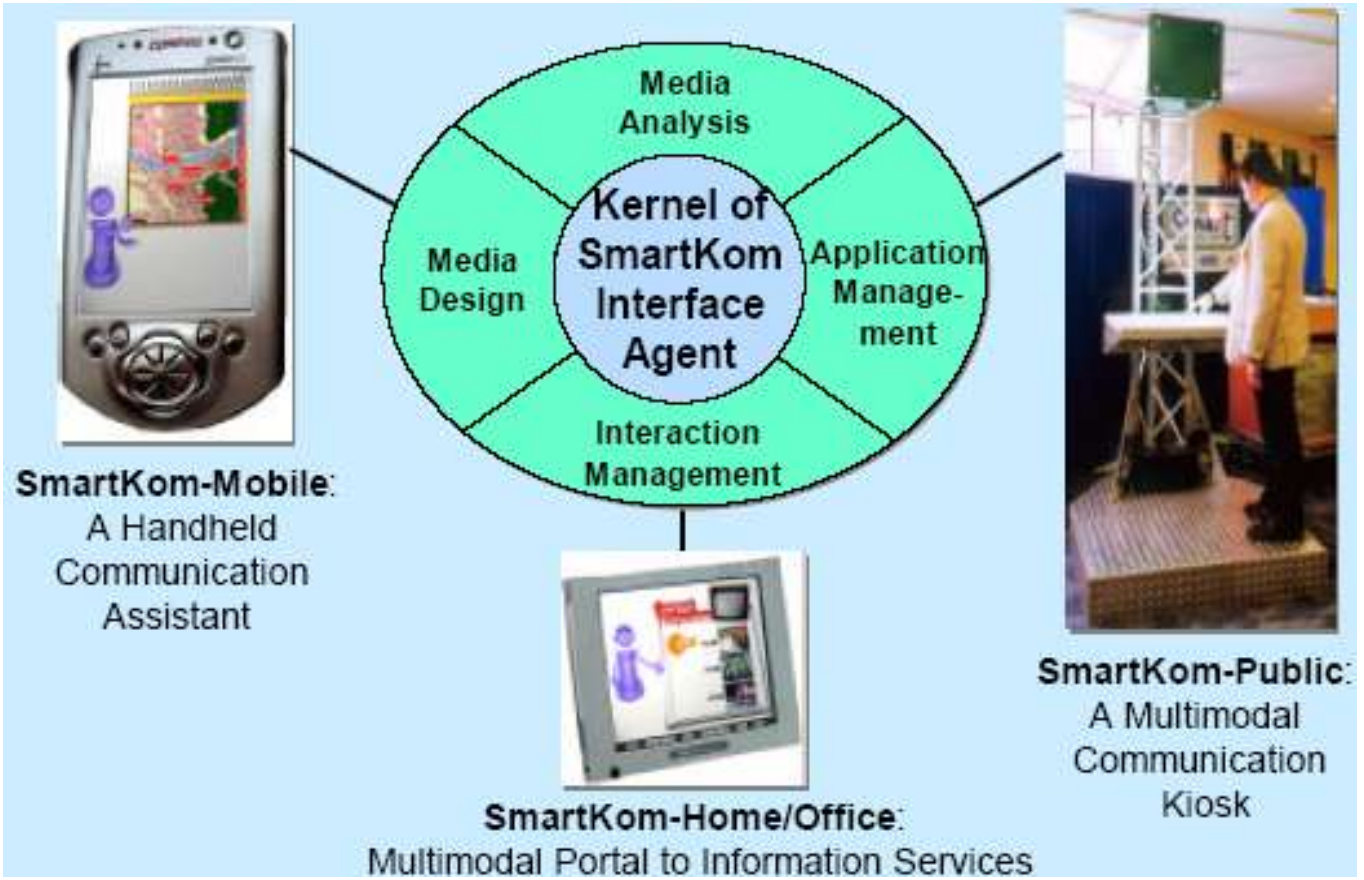
sowohl für den in- als auch für den output des Systems.

Der System output geschieht über den Interface-Agent oder auch virtual-communication-assistant hier genannt “Smartakus”.

Er spricht, zeigt Gesichtszüge (Mimik) und gestikuliert.

Unterstützt wird der Interface-Agent über spezifische grafische Darstellungen.

Momentan mögliche Applikationsfelder



SmartKom-mobile

Hardwarebasis: iPAQ Pocket PC

Anwendungsbereiche:
- Kopplung mit Autonavigationssystemen zur Routenplanung
- interaktive Fußgängernavigation z.B. durch Städte

input: Sprache, pen-based-pointing

output: (einfache Smartakus version) Sprache, Gestik, Mimik

SmartKom-home/office

Hardwarebasis: Fujitsu Stylistic 3500X portable webpad

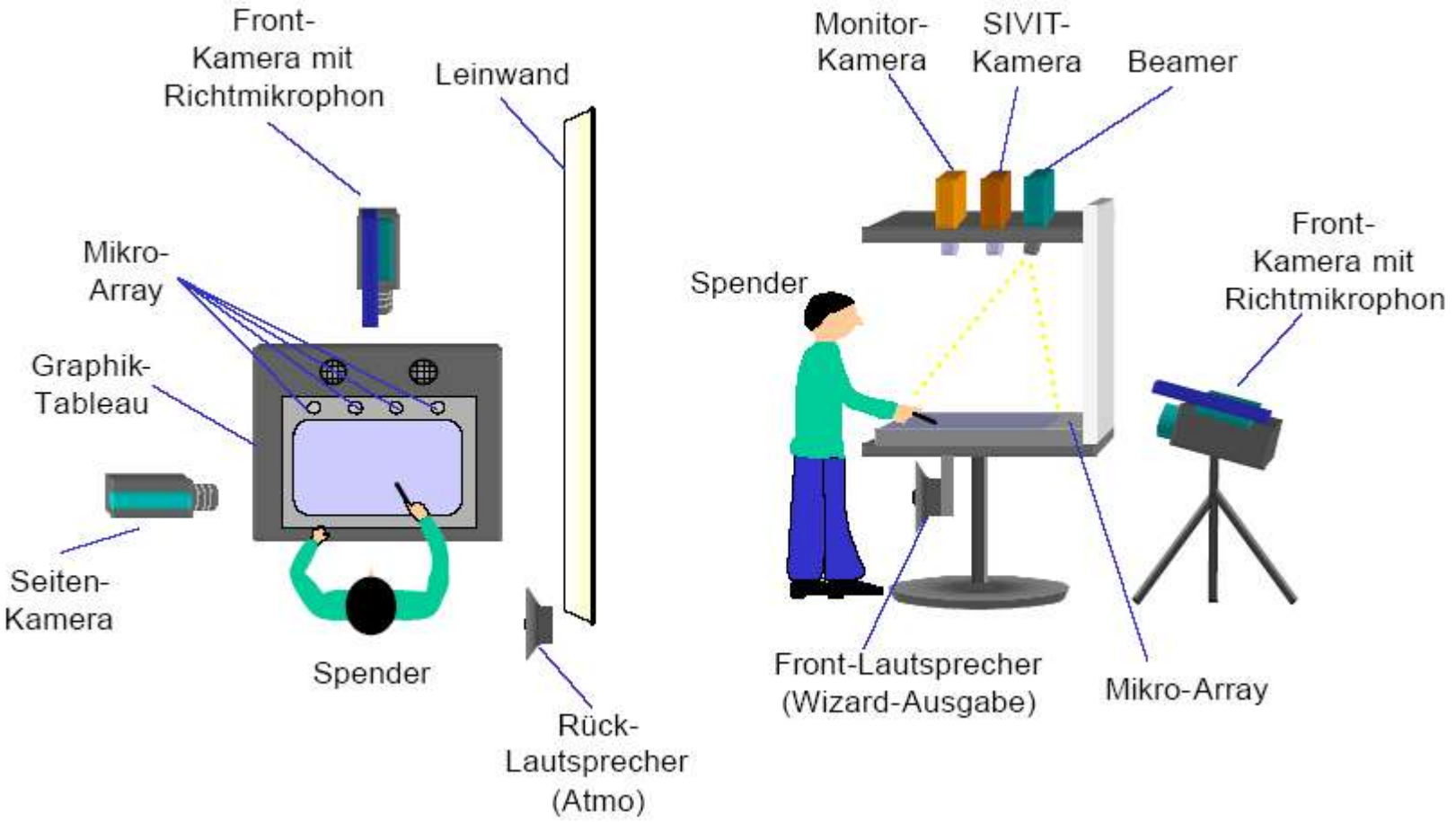
Anwendungsbereiche:
- z.B. Fernsehprogrammabfrage
- kontrollieren/programmieren von Elektrogeräten wie z.B. TV, Video, DVD
- telefonieren, e-mail Verwaltung

zwei Bedinungsarten:
lean-forward: in-/output über Sprache, Gestik (Stift) und virtuelle Anzeigen

lean-back: in-/output nur Sprache

SmartKom-public

Ein multimodaler Kummunikationskiosk



SmartKom-public
Ein multimodaler Kummunikationskiosk

Anwendungsbereiche:
Flughafen, Bahnhof, auch Hotel, Kino etc.

Allgemein sind auch persönliche Webdienste und telefonieren möglich.

Integration von Sprache, Gestik und Mimik
allgemeine Modalitätenverarbeitung

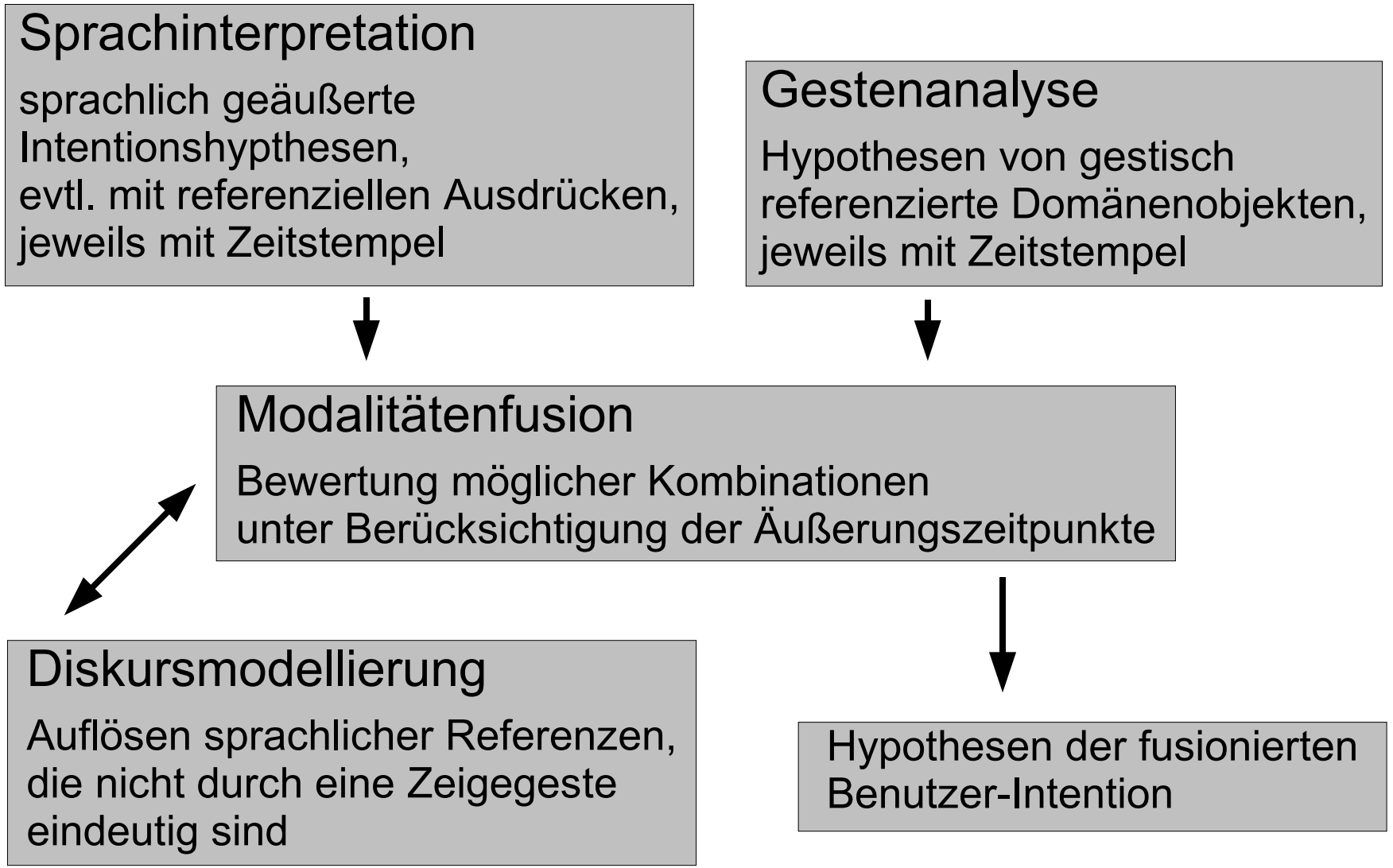
Leitvorstellung: einheitlicher Verarbeitungsansatz für alle Anwendungen und Modalitäten

mit folgenden Grundsätzen:
1. keine Speziallösungen für Anwendungen: Verarbeitung mittels übergreifenden, wissenbassiert parametisierten Ansätzen

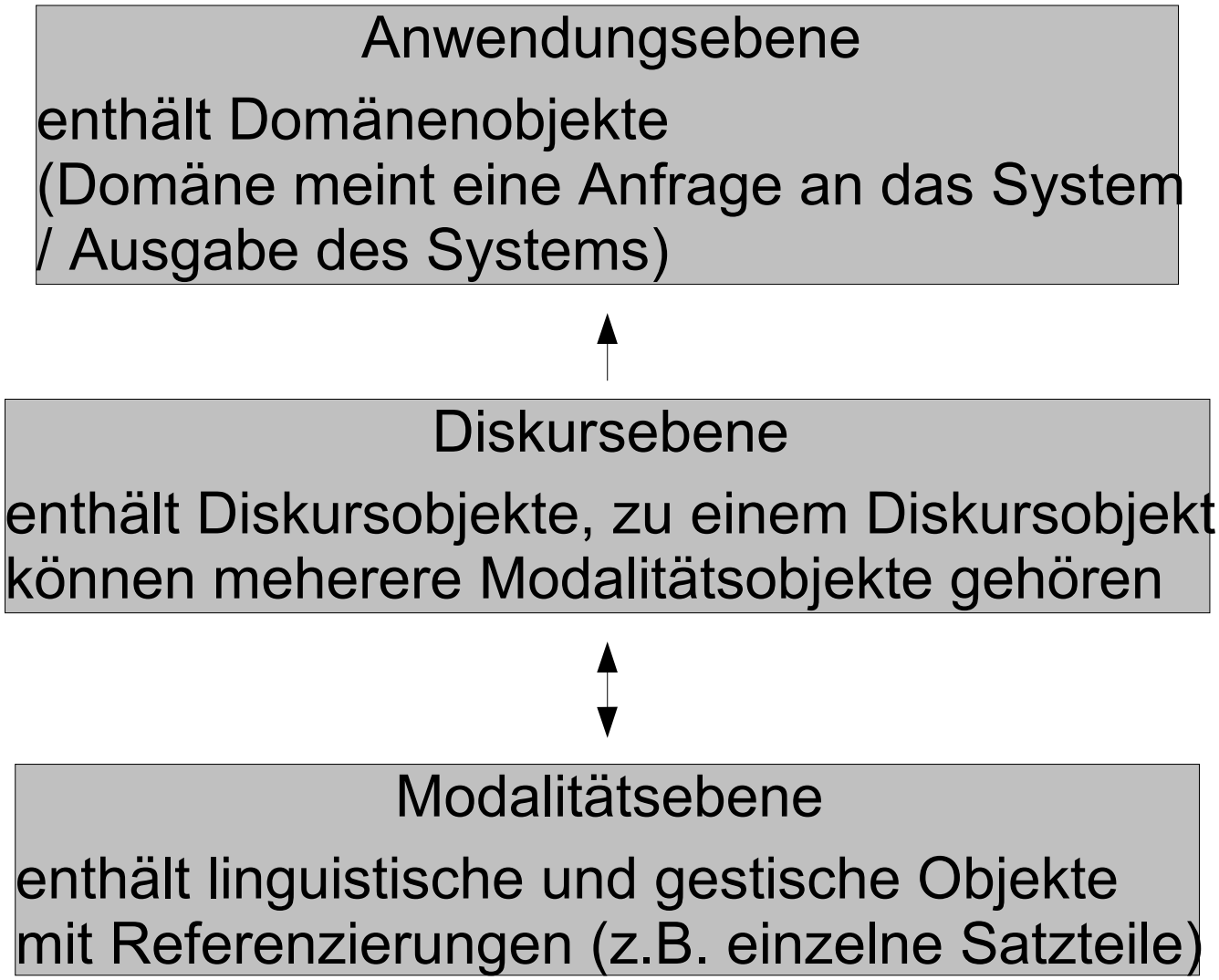
2. Für alle Ein- und Ausgaben einheitlicher, allgemein im System verwendeter Repräsentationsformalismus

3. Das System basiert auf einer Multi-Blackboard-Systemarchitektur -> weiterentwickelte Integrationsmiddleware von Verbmobil

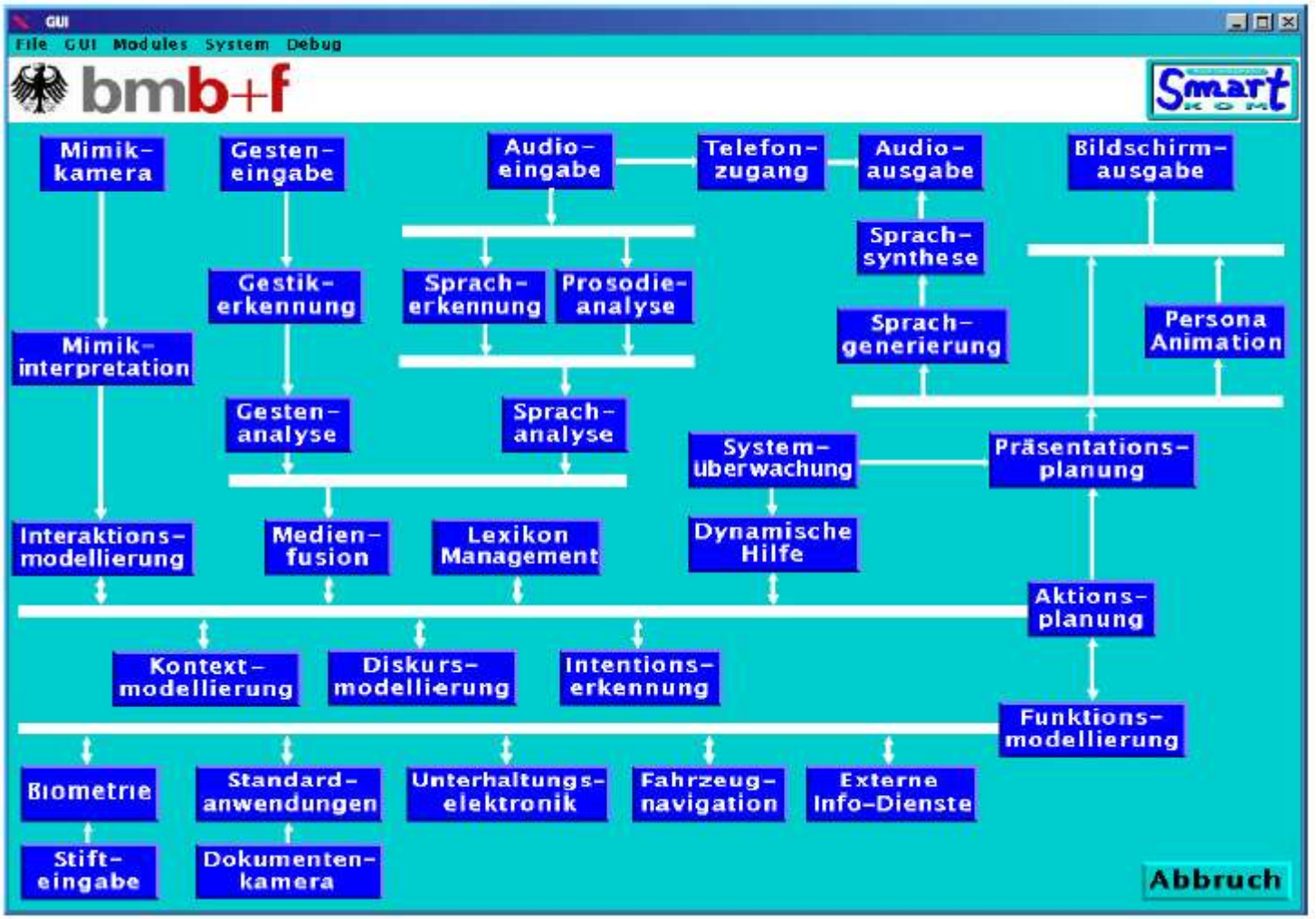
Modalitätenfusion im Kontext von SmartKom



Multimodales Dialoggedächtnis



Der Kontrollbildschirm von SmartKom



GESTIK:

- Gesten in SmartKom
- Wizard-of-Oz
- Zussamenfassung

- Neben Spracheingeben kann der Benutzer auch Geste benutzen um mit SmartKom zu sprechen.
- Man kann natürliche Geste benutzen.
- Ein vordefinierter Lexikon ist einen quasi-natürlichen Lexikon.
- Im SmartKom Geste sind als Zeigen und Einkreisen definiert.
- Gesten werden mit Virtual Touch Screen aufgenommen und anschlißend analysiert.
- Natürliche Gesten zeigen inneren Zustand des Benutzers.
- Verhalten der Geste spielt daher wichtige Rolle.

Merkmale der Geste:

- Geschwindeigkeit
- Beschleinigung
- Kinetische Energi
- Varianzamplitude
- Werden durch ein Hidden-Markov-Modell(HMM) analysiert.

Wizard-of-Oz :

- Aufgabe
- Coding Concept(Begriff der Codierung)
  - Interaktions Gesten
  - Unterstützende Gesten
  - Übrige Gesten
- Label(Schielder)
  - verschiedene arten von Labels(F-, R-, U- Labels)
- Probleme und weitere Arbeit

- Am Institut für Phonetik und Sprachliche Kommunikation in München wurden Wizard-of-Oz Aufnahmen gemacht.
- Versuchspersonen interagieren mit simuliertem multimodelen Dialogsystem und können beliebige Gesten einsetzen.
- Auf einem Video werden Geste bezüglich Anfangs- und Endzeitpunkt markiert und in Kategorien eingeteilt.
- Benutzer haben meist nur Zeigegesten benutzt.
- Das kann zur Analyse des Benutzerszustandes eingesetzt werden.

Sammeln die multimodale Daten:

- Gesammelte Daten wurden für drei verschiedene Hauptzwecke benutzt:

- Training von Sprache, Gesten und Erkennung von Emotionen.
- Etwicklung die Benutzer-, Sprache-, Dialog-Modellen und Sprachsysynthese Modul
- Allgemeine Bewertung des Verhaltens von Benutzern, während Interaktion mit der Maschine

Für Aufnahmen wurde benutzt :

- Digitale Kamera um Gesichtsausdrücke aufzunehmen.
- Zweite digitale Kamera für Geste. Sie nahm Seitenansicht des Benutzers auf.
- Infrarot empfindliche Kamera für Handgeste und graphische Ebene.
- Mit SiVit wurden die Koordinaten von Gesten aufgenommen.

Für Gestecodierung relevante Aufnahmen sind:

- Schwarz-weiße Bilder von infrarote Kamera und Seitenansicht des Benutzers und Ausga
- Seitenansicht des Benutzers
- Ausgaben des Beamers

Fragen bei der Definition von Gesten:

- Welche funktionale Geste kann man identifizieren?
- Welcher Weg der beste ist, um beobachtete Elemente zu kategorisieren?
- Welcher Weg der beste ist, um diese Elemente zu beschreiben?
- Welche klare morphologische Form haben die Elemente?
- Wie kann man Anfang und Ende von gegebenen Elementen definieren?
- Ist ein hinweisendes Wort im Audio enthalten,wenn ja welches dann?

Überblick über Codierungskonzept

- Label ist einen exakten Typ von Gesten.
- Wenn er zu einer von drei Kategorien gehört , wird er dann zu jeder identifizierte Geste zugewiesen.
- Wird von Modifizierer ergänzt.
- Drei Modifizierer weisen auf Zeit hin
- Andere drei auf Inhalt.
- Anfang und Ende sind Zeitpunkte, Takt ist ein Zeitperiode.
- „Geste“ als Folge von Segmenten.

- Wenn Geste eine Interaktion mit dem System ist, dann ist sie eine „Interaktions Geste“
- Unterhaltung mit sich selbst – „unterstützende Geste“
- Etwas anderes – „übrige Geste“
- Geschielderte Geste legen „Cubus“ fest.
- „Cubus“ sind Felder, die auf dem Display oder im Raum über dem Display liegen (wo SiVit die Daten aufnimmt).

Definition von Gesten:

- Drei Kategorien:
  - Interaktions Geste (I – Geste)
  - unterstützende Geste(U – Geste)
  - übrige Geste (R – Geste)
- Kriterium für Zuordnung ist Absicht des Benutzers.

I - Gesten:

- Sind kostruktiv.
- Wenn Benutzer macht eine Bewegung um eine Kommande dem Computer zu geben, dann heißt so eine Bewegung Interaktions Geste.
- Jede Bitte als eine Kommande.
- Wenn eine Kommande nicht ausfürlich ist, bleibt bittende Geste trotzdem eine Interaktions Geste.
- Bestätigung der Systemfrage ist auch eine Interaktions Geste.

U - Gesten:

- Sind konstruktiv.
- Kommen dann, wenn Phase der Bitte schon abgeschlossen ist.
- Sind gestikallische Unterstützung von „Solo – Aktion“ des Benutzers.
- Enden mit dem Ende von Kommanden nicht.
- Kommande nach der U – Geste wird getrent von ihr als Interaktions Geste geschieldert.

R - Gesten:

- Alle die zu „Cubus“ gehören aber sind nicht zu einer von beiden Kategorien zugeordnet.
- Sind nicht konstruktiv.
- Sind keine Bitte und keine Bestätigung.
- Sind emotionale und auch unbekannte Geste.

Labels:

Es gibt drei Kategorien von labels:

- F -
- U -
- R -

Folgende F- Labels existieren:

- F – Kreis(+)
- F – Kreis(-)

- F – Punkt(long +)
- F – Punkt(long -)
- F – Punkt(short +)
- F – Punkt(short -)
- F – frei(free)

U – Labels:

- U – Krei(read)
- U – Krei(search)

- U – Krei(count)
- U – Krei(ponder)
- U – Punkt(read)
- U – Punkt(ponder)

R – Labels:

- R – emotionell(+ cubus)
- R – emotionell(- cubus)
- R – unbekannt(+ cubus)

Beschreibung des Labels:

F – Label:

- F- Kreis(+)

- Ist eine ununterbrochene Bewegung.
- Die einen Objekt durch einkreisen spezifiziert.
- Muss nicht unbedingt eine Kreibewegung sein.
- Display ist berührt.
- Blick des Benutzers ist auf gewählten Objekt gerichtet.
- Mögliche Absicht des Benutzers ist etwas zu wählen.

- F – Kreis (-):

- Wie F – Kreis (+) nur Display ist nicht berührt.

- F – Punkt (long +):

- Eine selektive Bewegung, die einen Objekt auf dem Display spezifiziert.
- Display ist berührt.
- Taktdauer ist gedehnt.
- Blick des Benutzers ist auf gewählten Region gerichtet.
- Absicht ist etwas zu wählen.

- F – Punkt (long -):

- Wie F – Punkt (long +) aber Display ist berührt.

- F – Punkt (short +):

- Wie F – Punkt(+) ,nur Taktdauer der Bewegung ist sehr kurz.

- F – Punkt (short -):

- Wie F – Punkt (short +) aber Display ist nicht berührt.

- F – frei (free):

- Bewegung weist auf Wunsch oder Kommande hin.
- Man kann die Geste hinsichtlich ihre Morphologie variieren.
- Blick des Benutzers ist gerade auf den Display gerichtet.

U – Labels:

- U – Kreis (read):

- Ununterbrochene Bewegung.
- Kein Objekt ist ausewählt.

- Benutzer liest.
- Blick ist direkt auf den Text gerichtet.

U – Kreis (search):

- Ununterbrochene Bewegung.
- Benutzer sucht etwas.
- Bewegung kann gerade oder krümmelig sein.
- Kein Region ist eingekreist.
- Blick ist auf den Display gerichtet.

- U – Kreis (count):

- Ununterbrochene Bewegung.
- Kein Einkreisen des bestimmten Regiones ist vorhanden.
- Benutzer zählt.

- U – Punkt (read):

- selektive Bewegung möglich aber nicht notwendig.
- Benutzer will kein Text auswählen.
- Blick ruhrt auf einem bestimmtern Punkt.
- Benutzer liest.

- U – Punkt (ponder):

- selektive Bewegung möglich aber nicht notwendig.
- Benutzer will kein Objekt auswählen.
- Blick ruhrt auf einem bestimmtern Punkt.
- Benutzer hat keine Absicht etwas zu wählen.
- Benutzer liest.

R – Labels

- R – emotionell (cubus +):

- Bewegung im cubus mit einem emotionellen Inhalt.
- Geste drückt bestimmte Emotionen aus.

- R – emotionell (cubus -):

- Wie R – emotionell (cubus -) aber außerhalb des cubus.

- R – unbekannt (cubus +):

- Bewegung innerhalb des cubus, die zu keinem anderen Label zugeordnet werden kann.

Probleme und zukünftige Arbeit:

- Konzept ist erster Schritt im Entwicklungsprozesse.
- Nächster Schritt ist Bewertung und Verbesserung.

- Qualitätsmaß soll festgesetzt werden.
- Zeitspannen klarer definieren.
- Jetzt noch nicht klar wie man Takt zuverlässig definieren kann.
- Beobechtete Gesten sind stark von Displaybenutzung beinflusst.
- Deswegen zusätzliche Labels einfügen.

- Konzept ist immer noch zu komplex.
- Aufgetauchte Kategorien sind gute Kandidate für selektive Labels.
- Man soll ihren Qualitätskriterium im 2. Schritt teste.

Mimik

- Einführung
- Der Gesichtsausdruck
- Erkennen von Gesichtsausdrücken
- Beispieltypen von Gesichtsausdrücken
  - Beispiel: Mimik in SmartKom
- Anwendung

Einführung

- Dialogsysteme sind für Normal-Benutzern (keine Fachleute) konstruiert.
- Benutzer wollen keine lange Beschreibung über die Funktionalitäten durchlesen.
- Mensch-Maschine Dialoge sollen in Mensch-Mensch Dialoge verbessert werden.
- Aber wie soll ein Mensch-Mensch Dialog aussehen?
- Folgende Input-Informationen werden benutzt:
  - Ohren für Wahrnehmen der Wörter
  - Augen zum Erkennen von Körperbewegungen
  - Nase zum Riechen
  - Haut zum Erkennen von physischen Kontakte

Der Gesichtsausdruck

- Gesichtsausdrücken sind auch innere Zustände (Benutzerzustand), z.B. Hilflosigkeit oder Ärger.
- Idee zum Erkennen von Gesichtsausdrücken:

  - ein ärgerlichen Benutzer zu erkennen

  → Änderung der Dialogstrategien vom System
  → Unterstützung
- 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.
- Bestimmen des emotionalen Zustandes einer Person ist die Aufgabe für das Erkennen von Gesichtsausdrücken.

Erkennen von Gesichtsausdrücken

- Das Modul arbeitet in zwei Schritten:
  - Lokalisation des Gesichtes
  - Klassifizierung der mimischen Ausdruck des Gesichtes

- **Lokalisation des Gesichtes :**
- Alle nicht hautfarbenen Bereiche werden ausgeblendet.
- Mit einem Klassifikator wird die Position mit der größten Übereinstimmung gesucht.

- Vorteile der Hautfarbensegmentierung:
  - Reduzierung des Suchraums
  - Elimination der gesichtsähnliche Texturen im Hintergrund

- **Klassifikation des Gesichtsausdrucks:**

- Hier wird ebenfalls von Klassifikatoren durchgeführt.
- Die verschiedenen Gesichtsausdrücken mehrere Personen sind 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

- Kategorien von Gesichtsausdrücken:
  - Freude
  - Ärger
  - Hilflosigkeit
  - nachdenklich
  - überrascht
  - neutral

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

Freude

- Kriterien:
  - 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

Ärger

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

Nachdenklich

- Kriterien:
  - Stirn runzeln
  - an den Lippen beißen
  - auf die Decke schauen
  - langsame Bewegung
  - Mund teilweise offen
  - zögern
  - leises Gemurmel
  - stottern

Hilflosigkeit

- Kriterien:
  - Stirn runzeln
  - Falte am Stirn
  - offener Mund
  - zögern
  - stottern
  - Augenbrauen nach oben

- Problem:
  Verwechslung mit „nachdenklich“
- Unterscheidungskriterien:

1. Frage: „Wie fühlt sich der Benutzer?“
   „unter Kontrolle“ → „nachdenklich“
   „außer Kontrolle“ → „hilflos“
2. Gesichtteile:
   Mund → „nachdenklich“
   Augenbrauen, Stirn → „hilflos“

Allgemeine Probleme

- Das System kann die Intensität der Emotionen nicht erkennen.
- Die Beurteilung von Emotionen ist manchmal schwer in Kategorien zuzuordnen, da viele Merkmale ähnlich sind.
- z.B. das Merkmal „Kopf schütteln“ kann Ärger oder auch hilflos bedeuten.

Beispiel: Mimik in SmartKom

- 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?**

Anwendung

- Ein Beispielanwendung ist ein Fernsehprogramm.

- In den Versuchen zeigen die Benutzer keine Verwirrung wegen des Kameras.
- Es gibt keine Ergebnisse über die emotionalen Zustandsinformation.
- Der Grund ist, die Benutzer wussten nicht, 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

- Die Benutzer sind mit dem System zufrieden und würden es gerne weiterbenutzen.