MobileHCI, SIMPE Workshop:
Speech in Mobile and Pervasive Environments
Singapore, 9th September 2007

Context-Sensitive Multimodal Mobile Interfaces

Speech and Gesture Based Information Seeking Interaction with Navigation Maps on Mobile Devices

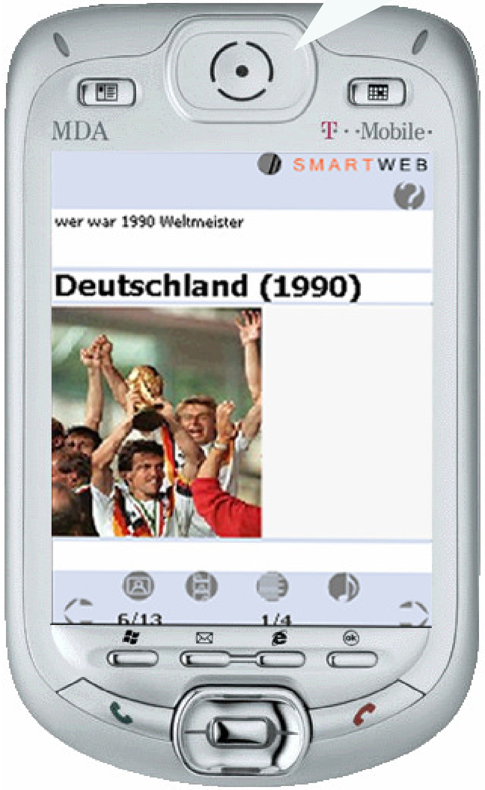
Daniel Sonntag
German Research Center for Artificial Intelligence
66123 Saarbruecken, Germany

daniel.sonntag@dfki.de

Agenda

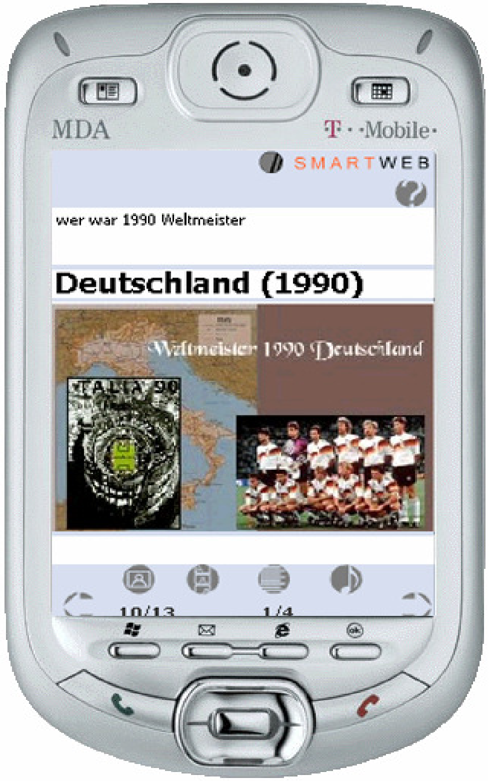
- SmartWeb and Multimodal HCI
- Speech and Gesture Based Navigation
- Conclusions

Who was world champion in 1990 ?



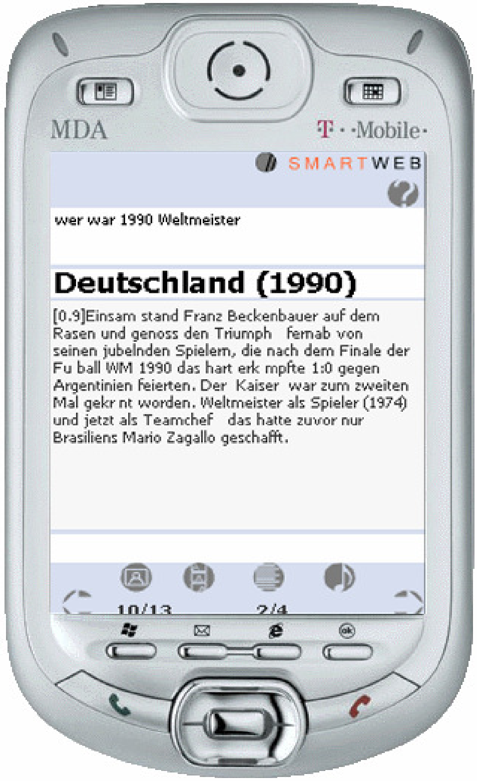
wer war 1990 Weltmeister

Deutschland (1990)



wer war 1990 Weltmeister

Deutschland (1990)



wer war 1990 Weltmeister

Deutschland (1990)

[0.9]Einsam stand Franz Beckenbauer auf dem Rasen und genoss den Triumph fernab von seinen jubelnden Spielern, die nach dem Finale der Fu ball WM 1990 das hart erk mpfte 1:0 gegen Argentinien feierten. Der Kaiser war zum zweiten Mal gekr nt worden. Weltmeister als Spieler (1974) und jetzt als Teamchef das hatte zuvor nur Brasiliens Mario Zagallo geschafft.

Question Answering Functionality

SMARTWEB

- Intuitive multimodal access to a rich selection of Web-based information services.

HCI and dialogue system goals:

- Provide concise and correct **multimedia** answers in a **multimodal** way.
- Show how knowledge retrieval from ontologies and Web Services can be combined with advanced dialogical interaction, e.g., **system** clarifications.
- Provide ontology-based **integration** of verbal and non-verbal system input (fusion) and output (reaction/presentation).

The SmartWeb Consortium

Federal Ministry
of Education
and Research

T - Systems

UNIVERSITÄT
DES
SAARLANDES

SYMPALOG
Speech Technologies

SIEMENS

European Media Lab

Chair for
Pattern Recognition
FAU Erlangen–Nuremberg

Ludwig-Maximilians-
Universität München

IMS Institut für Maschinelle
Sprachverarbeitung,
Universität Stuttgart

LMU
IPSK

AIFB

DAIMLERCHRYSLER

FIRST
Fraunhofer
Institut
Rechnerarchitektur
und Softwaretechnik

ontoprise
SEMANTICS FOR THE WEB

ISI
Berkeley, USA

Funded by the German Government and Industry

Funding: 13.7 M €, Budget: 24 M €

Scientific Director: Wolfgang Wahlster

Project Duration: 2004-2008

More than 60 Researchers and Engineers

Smartweb Requirements

- Multimodal dialogue with question answering functionality.
- Speech is dominant input modality for interaction.
- Multimodal recognition for speech or gestures.
- Modality interpretation and fusion, intention processing.
- Modality fission, result rendering for text, images, videos, graphics, and synthesis of speech.
- Reuse already existing components.
- Control the message flow in the system.

[2] http://www.smartweb-project.de/start_en.html
[3] http://www.w3.org/TR/emma
[4] http://www.w3.org/TR/speech-synthesis
[5] http://www.w3.org/TR/rdf-primer
[6] http://www.w3.org/Submission/OWL-S
[7] http://www.w3.org/TR/wsdl
[8] http://www.w3.org/TR/soap
[9] http://www.chiariglione.org/mpeg

Interactive Mobile Multimodal Semantic Web Access

3G smartphone

?

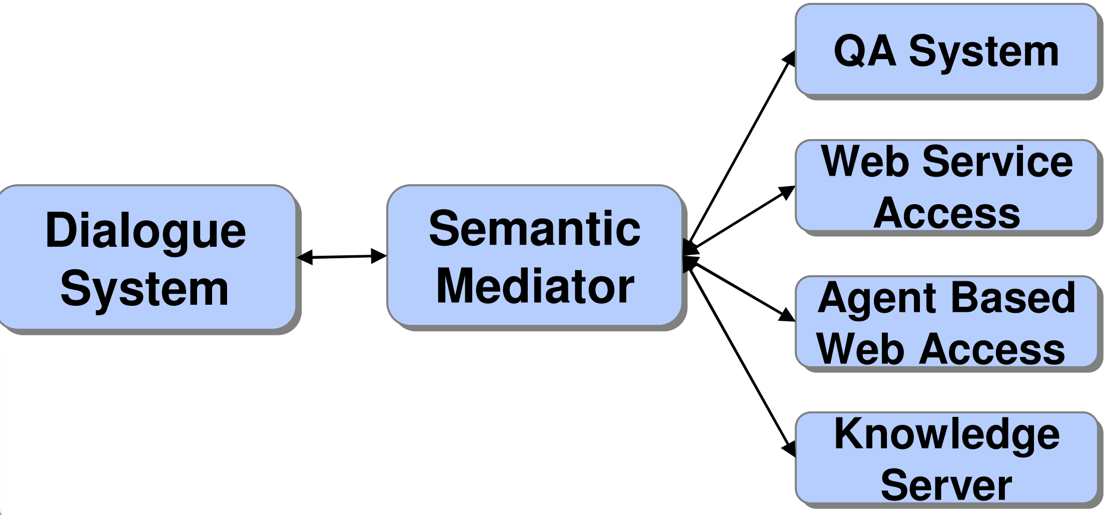
(1) Lack of computational power — Challenges

(2) Small screen

(3) Exploit contextual information — Opportunities

(4) New interaction possibilities

MDA
T··Mobile·
SMARTWEB
zeige mir das Maskottchen der WM
Goleo
Wann wurde Deutschland zum ersten Mal Weltmeister?
1954
Rahn schießt das 3:2



Application Scenarios

- Personal guide at the FIFA Worldcup 2006
- Answer football related and navigation related questions.

German Telekom Mobility

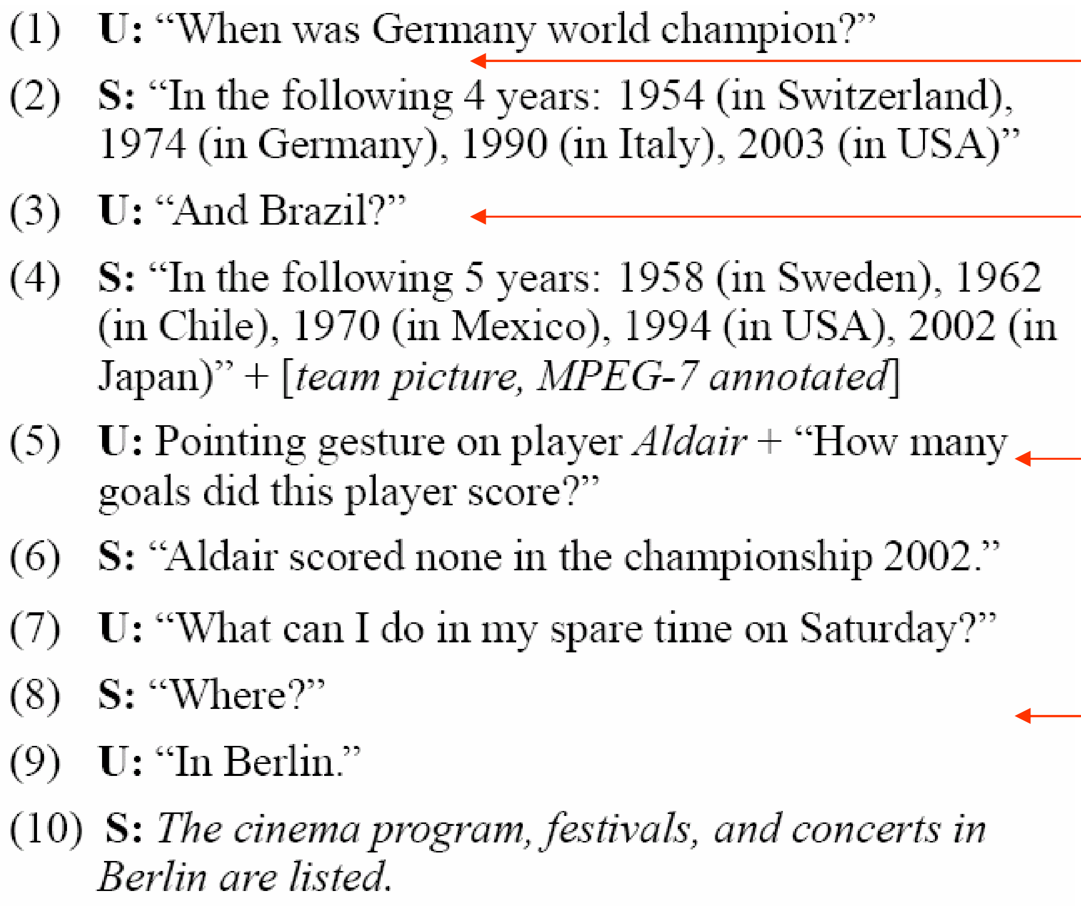
and Navigation Scenario

http://smartweb.dfki.de/SmartWeb_FlashDemo_eng_v09.exe

Presentation Design Guidelines

- Produce useful reactions and give hints or examples to the user so that the use of supported terminology is not insisted, but at least directed.
- Keep acoustic messages short and simple.
- Align speech synthesis to a text fragment.
- Deal with layout as a rhetorical force.

Natural Dialogue Based Mobile Interaction Example



(1) **U:** "When was Germany world champion?"

(2) **S:** "In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy), 2003 (in USA)"

(3) **U:** "And Brazil?"

(4) **S:** "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)" + [*team picture, MPEG-7 annotated*]

(5) **U:** Pointing gesture on player *Aldair* + "How many goals did this player score?"

(6) **S:** "Aldair scored none in the championship 2002."

(7) **U:** "What can I do in my spare time on Saturday?"

(8) **S:** "Where?"

(9) **U:** "In Berlin."

(10) **S:** *The cinema program, festivals, and concerts in Berlin are listed.*

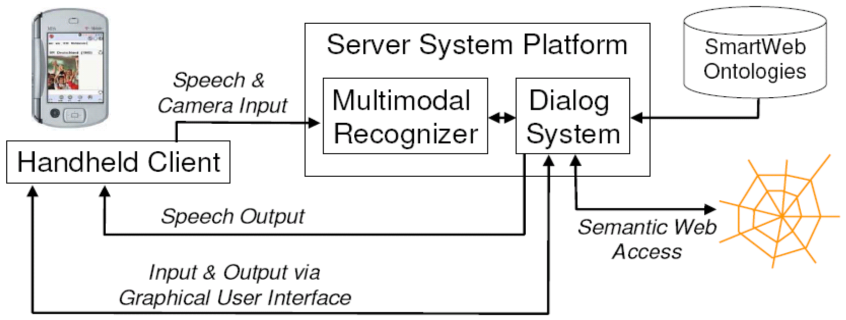
Inducting & deducing enumeration questions

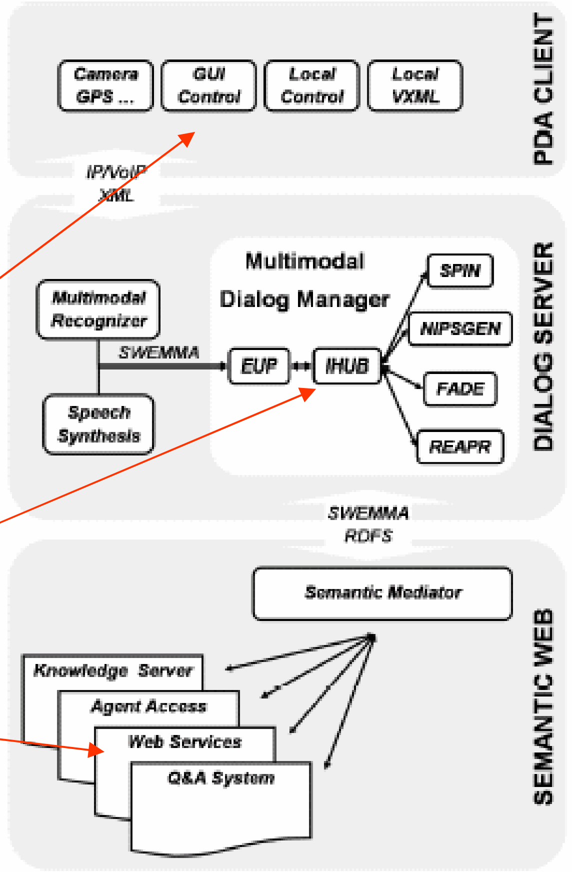
Ellipsis resolution & query completion

Integration of verbal and non-verbal output

System clarifications in Web Service description

Technical Design





Graphical User Interface Control

Information Hub

Web Service Access

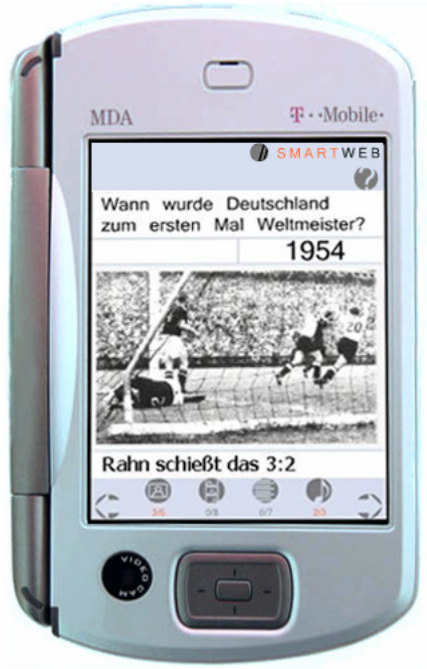
T-Traffic info

amazon.de

SMARTWEB

WM-Guide

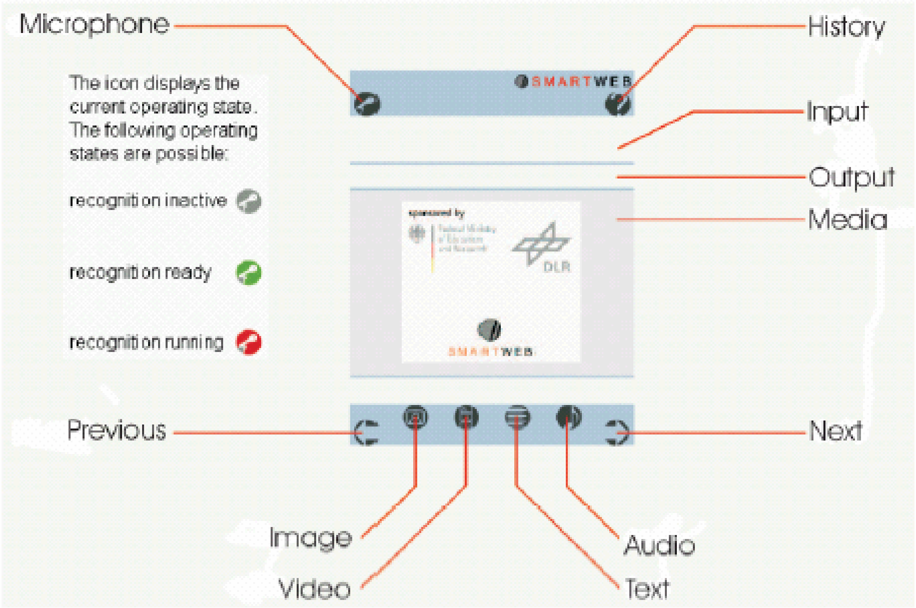
Core User Interface



Wann wurde Deutschland zum ersten Mal Weltmeister?

1954

Rahn schießt das 3:2



The icon displays the current operating state. The following operating states are possible:

recognition inactive

recognition ready

recognition running

Multimodal Interaction Guidelines

- *Multimodality*: More modalities allow for more natural communication.
- *Encapsulation:* Encapsulate user interface proper from the rest of the application.
- *Standards:* Re-use own and others resources.
- *Representation*: A common ontological knowledge base eases data flow, avoids transformations, and provide a basis for processing natural language dialogue phenomena.
  - Principles:
    - No presentation without representation
    - No interaction without representation

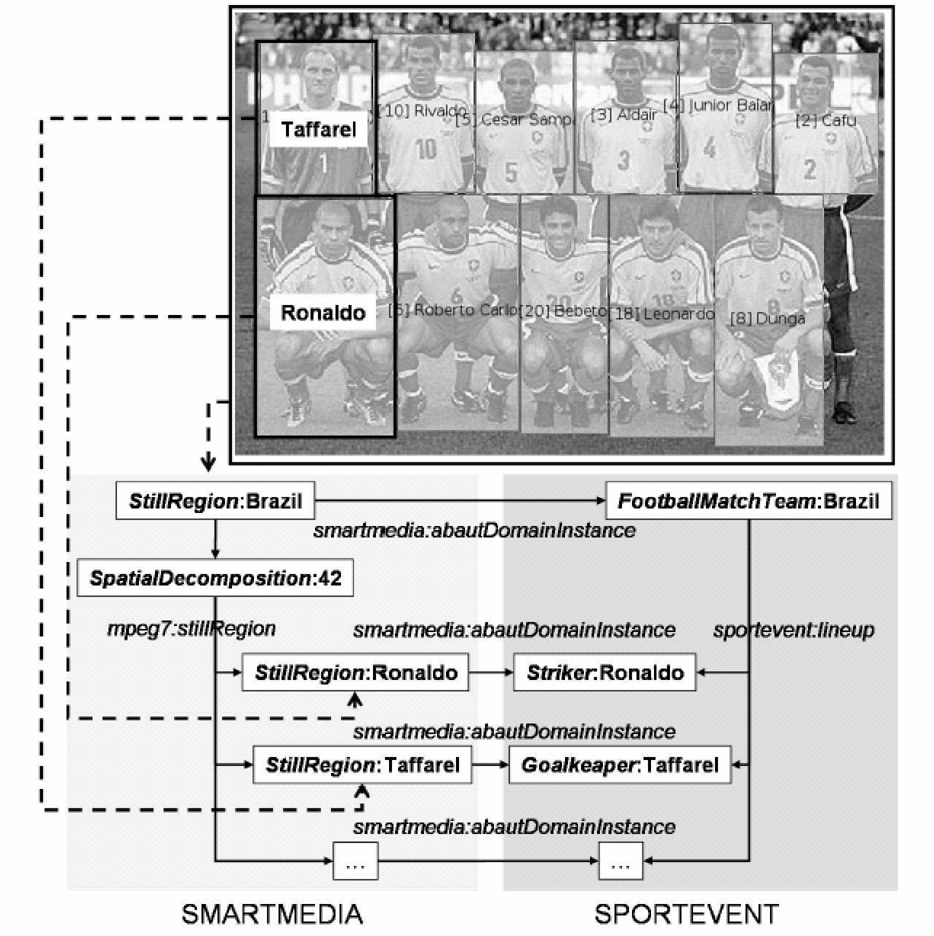
Ontologies

- An Ontology is
  - an explicit specification of a conceptualization [Gruber 93].
  - a shared understanding of a domain of interest [Uschold/Gruninger 96].
  - a **community reference** for applications.
  - **shared understanding** of what particular information means.
  - (language) concepts and facts in relation to each other.

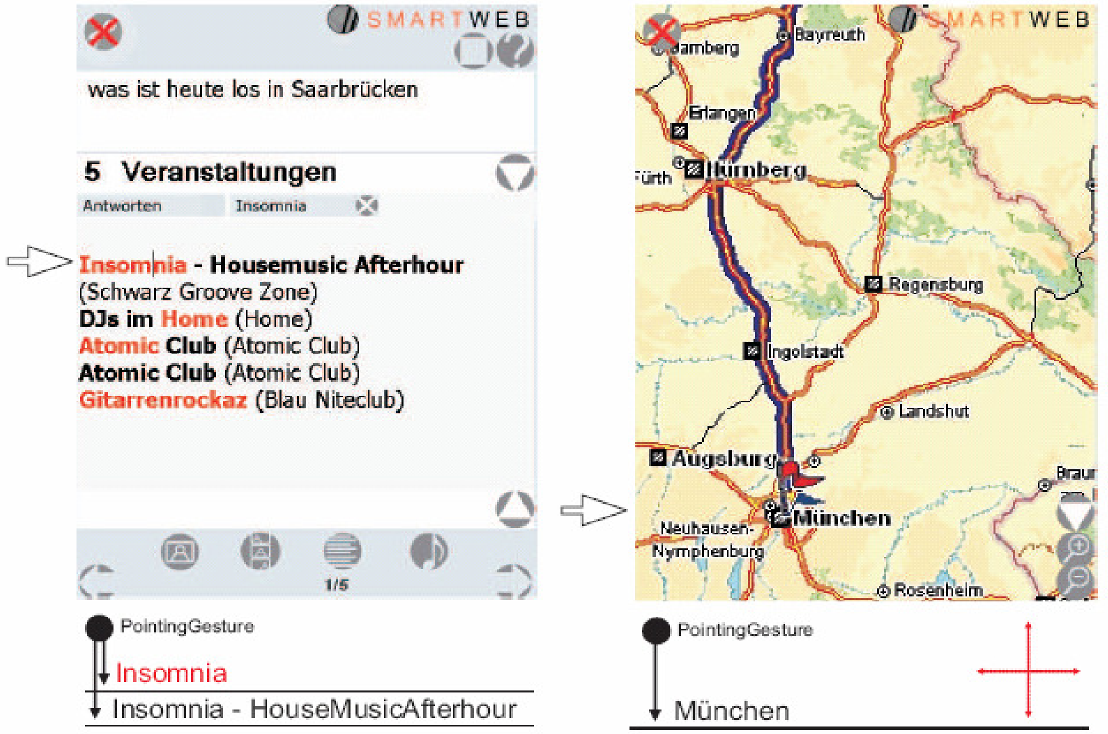
- Ontologies make domain assumptions **explicit.**
  - Separate **domain knowledge** from operational knowledge.
  - Re-use domain and operational knowledge separately.

Ontology Representation and Multimedia

- Framework for gesture and speech fusion
- Multimedia decomposition in space, time and frequency (MPEG-7)
- Link to the Upper Model Ontology to close the *Semantic Gap*



Pointing Gestures



SMARTWEB

was ist heute los in Saarbrücken

5 Veranstaltungen

Antworten | Insomnia

Insomnia - Housemusic Afterhour (Schwarz Groove Zone)
DJs im Home (Home)
Atomic Club (Atomic Club)
Atomic Club (Atomic Club)
Gitarrenrockaz (Blau Niteclub)

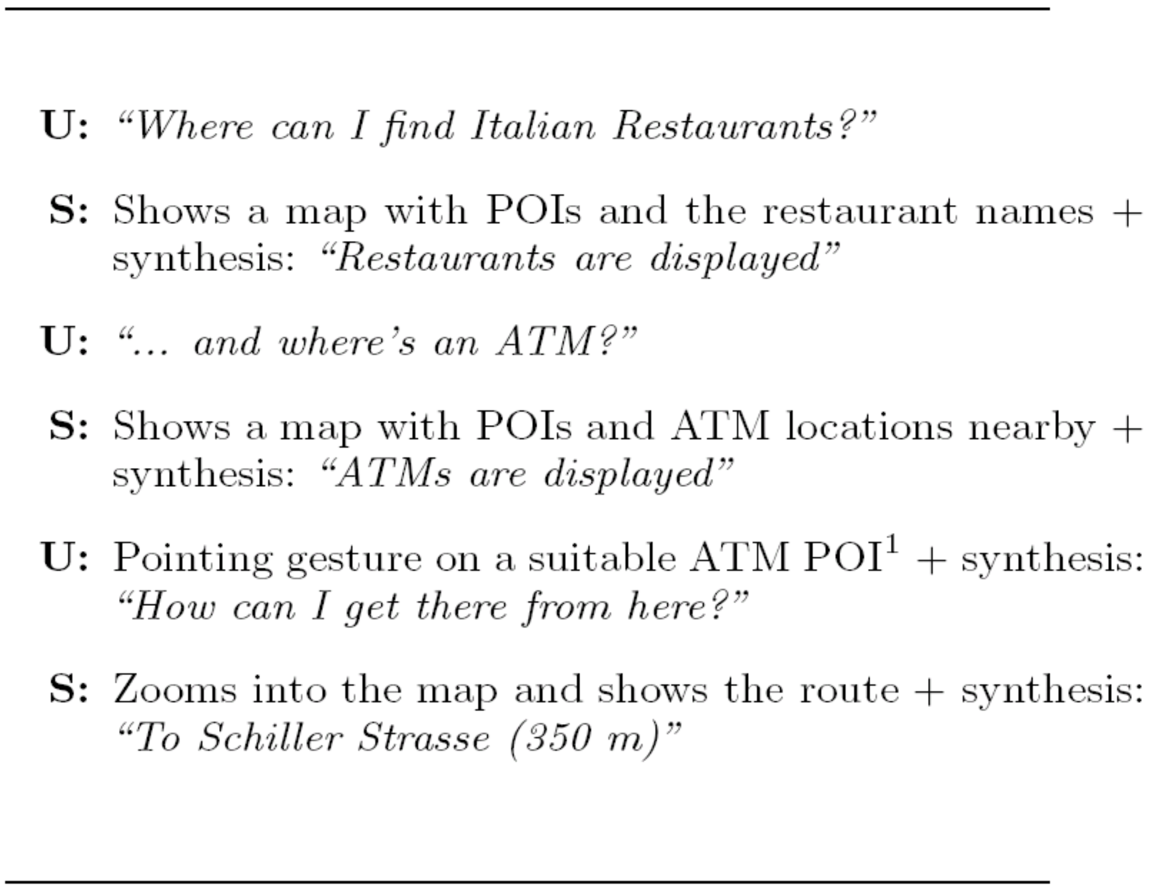
1/5

PointingGesture
Insomnia
Insomnia - HouseMusicAfterhour

PointingGesture
München

Figure 1: Pointing Gestures allow the selection of hyperlinks links, text entities, and POIs. Every pointing gesture should refer to a visual object that is transmitted to an input fusion module.

Navigation Scenario



U: *“Where can I find Italian Restaurants?”*

S: Shows a map with POIs and the restaurant names + synthesis: *“Restaurants are displayed”*

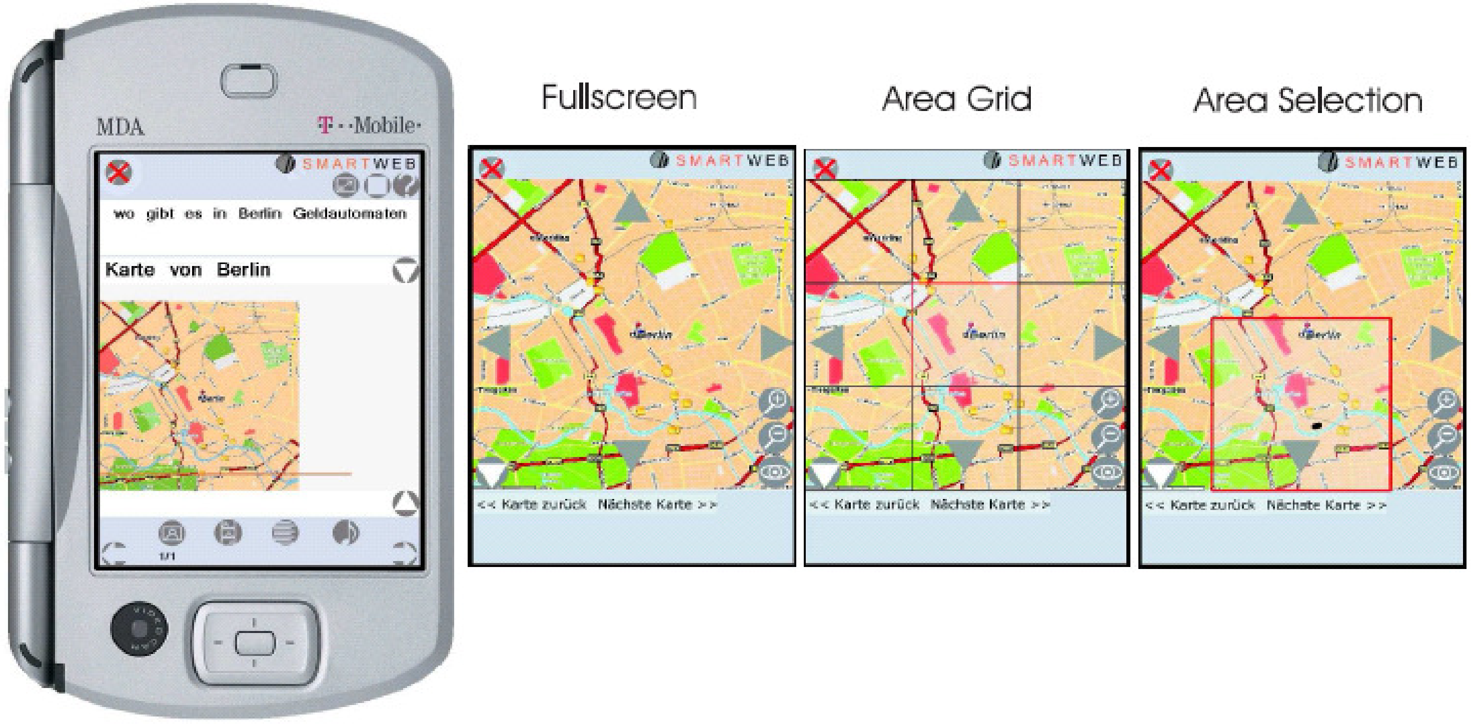
U: *“... and where’s an ATM?”*

S: Shows a map with POIs and ATM locations nearby + synthesis: *“ATMs are displayed”*

U: Pointing gesture on a suitable ATM POI[1] + synthesis: *“How can I get there from here?”*

S: Zooms into the map and shows the route + synthesis: *“To Schiller Strasse (350 m)”*

Navigation Map Result Presentation



wo gibt es in Berlin Geldautomaten

Karte von Berlin

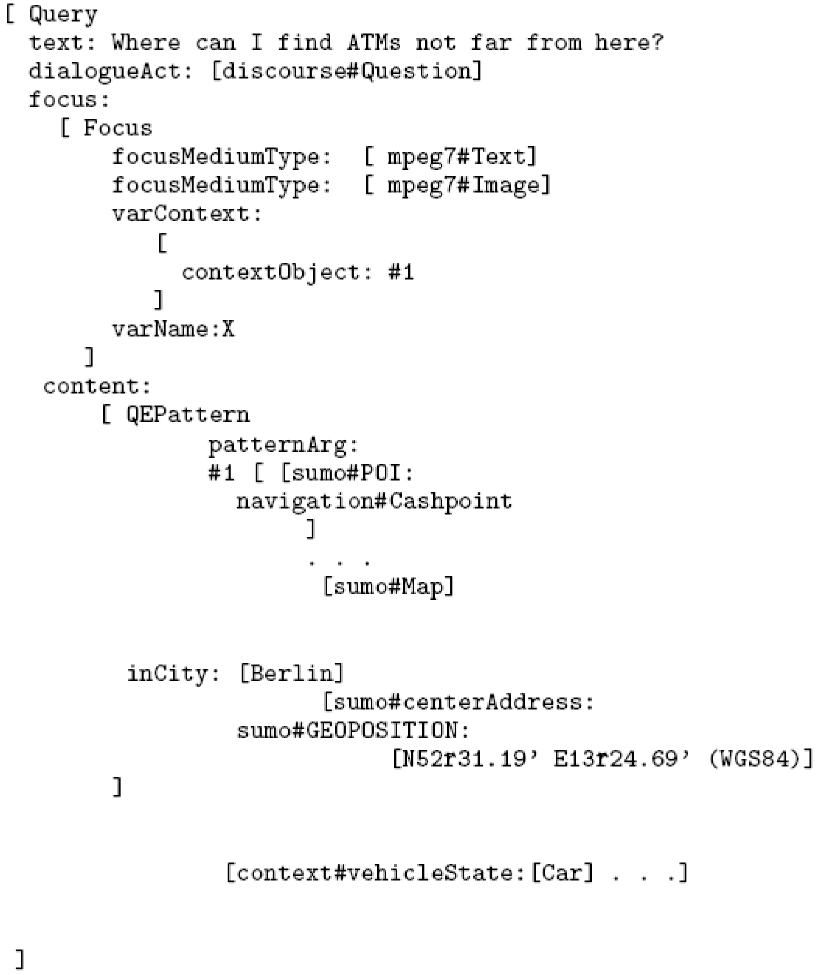
<< Karte zurück Nächste Karte >>

Use new graphical surface to indicate narrowed dialogue context.

Use graphical screen transitions as system dialogue act.

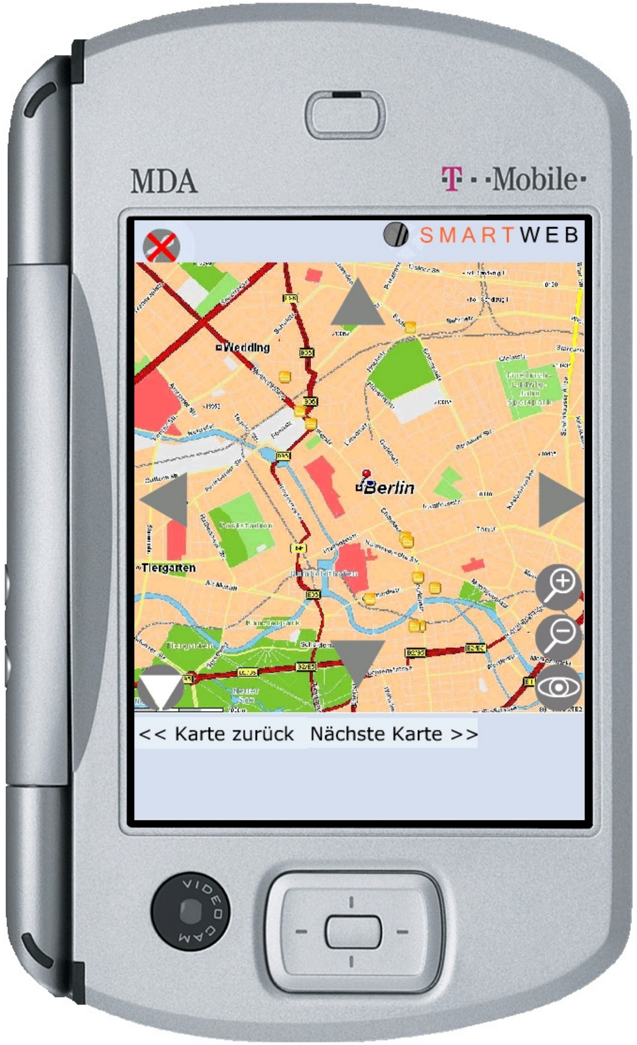
Input Fusion and Semantic Query Construction

- *Where can I find ATMs not far from here?*



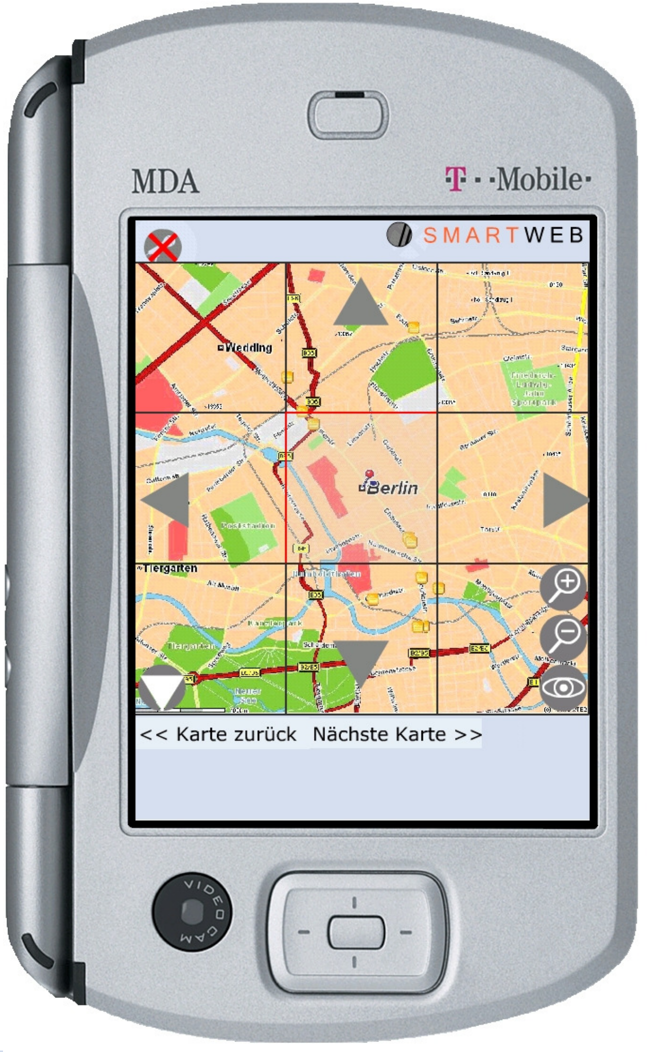
```
[ Query
  text: Where can I find ATMs not far from here?
  dialogueAct: [discourse#Question]
  focus:
    [ Focus
        focusMediumType:  [ mpeg7#Text]
        focusMediumType:  [ mpeg7#Image]
        varContext:
           [
             contextObject: #1
           ]
        varName:X
      ]
   content:
       [ QEPattern
               patternArg:
               #1  [ [sumo#POI:
                 navigation#Cashpoint
                      ]
                      . . .
                       [sumo#Map]


         inCity: [Berlin]
                       [sumo#centerAddress:
                 sumo#GEOPOSITION:
                            [N52r31.19' E13r24.69' (WGS84)]
        ]


                [context#vehicleState:[Car] . . .]


 ]
```

Navigation 1/6

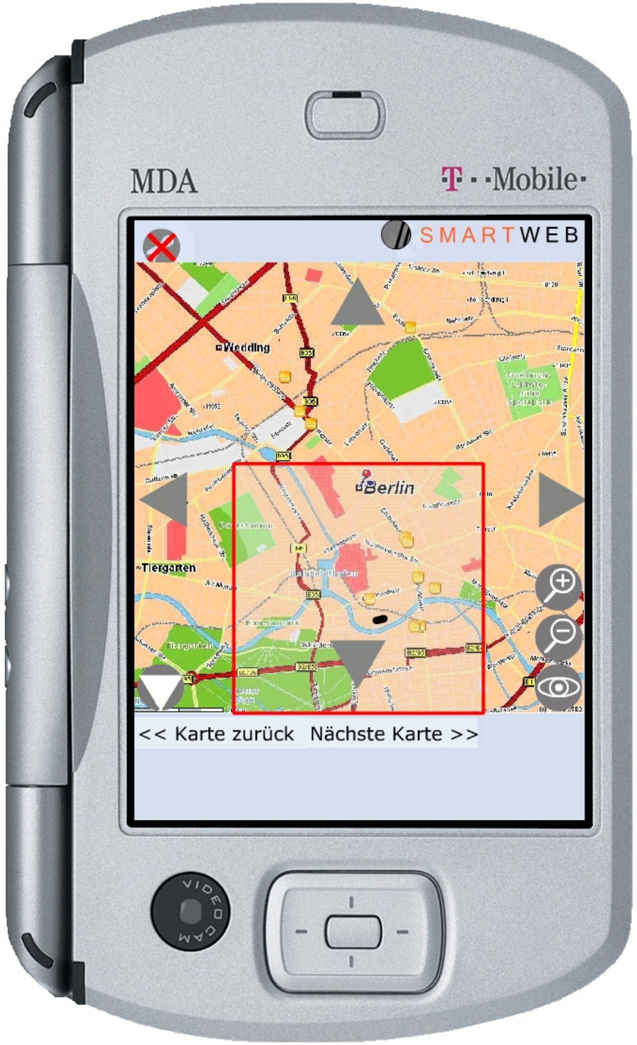


<< Karte zurück Nächste Karte >>

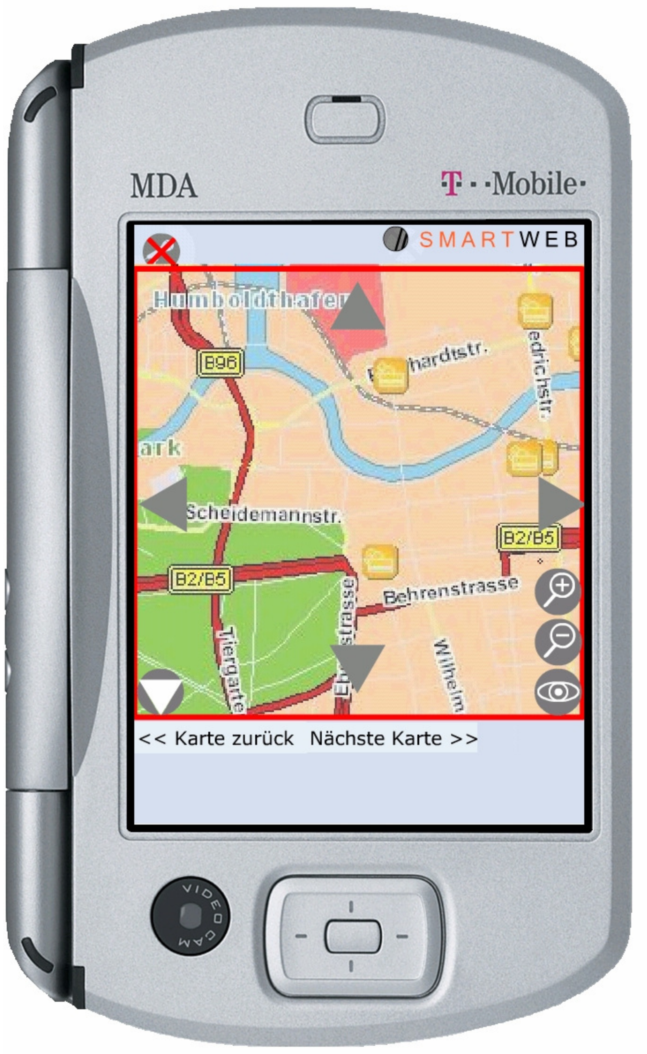
Navigation 2/6



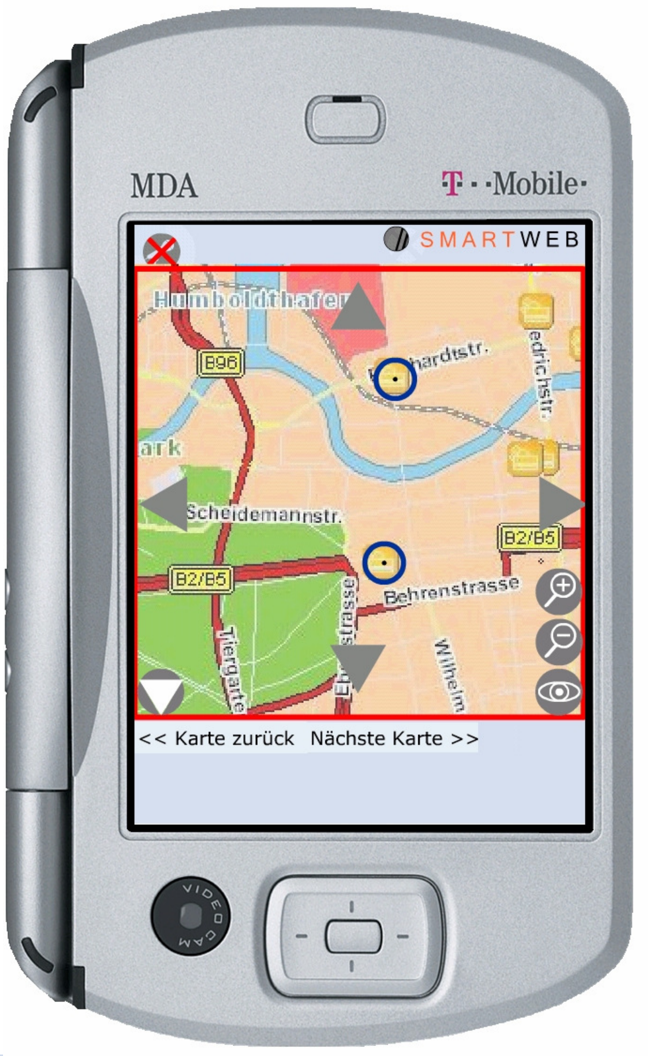
Navigation 3/6



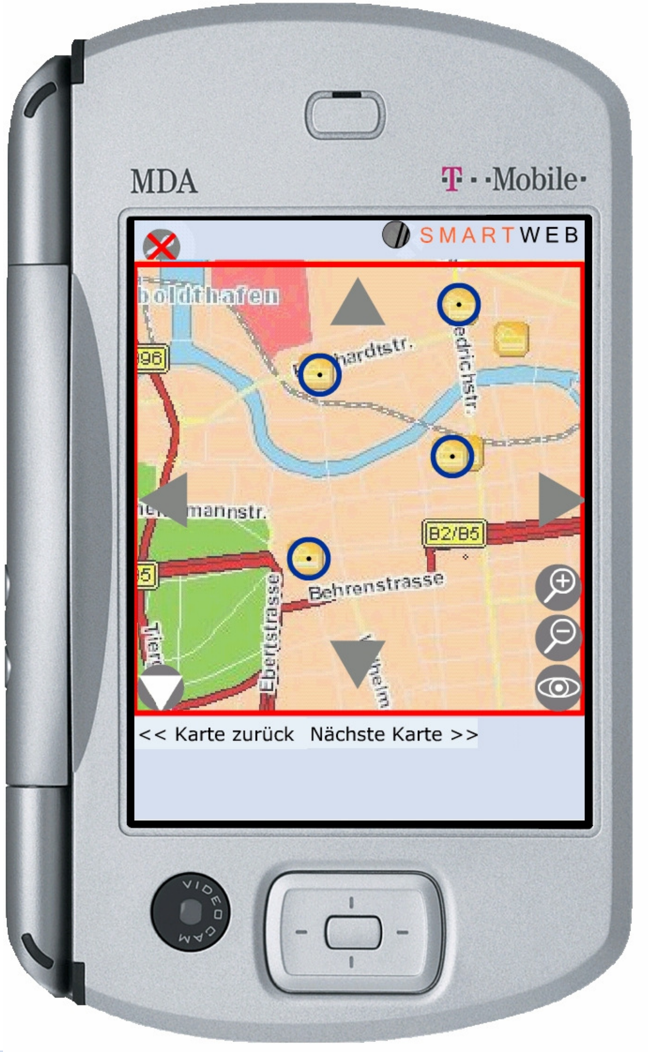
Navigation 4/6



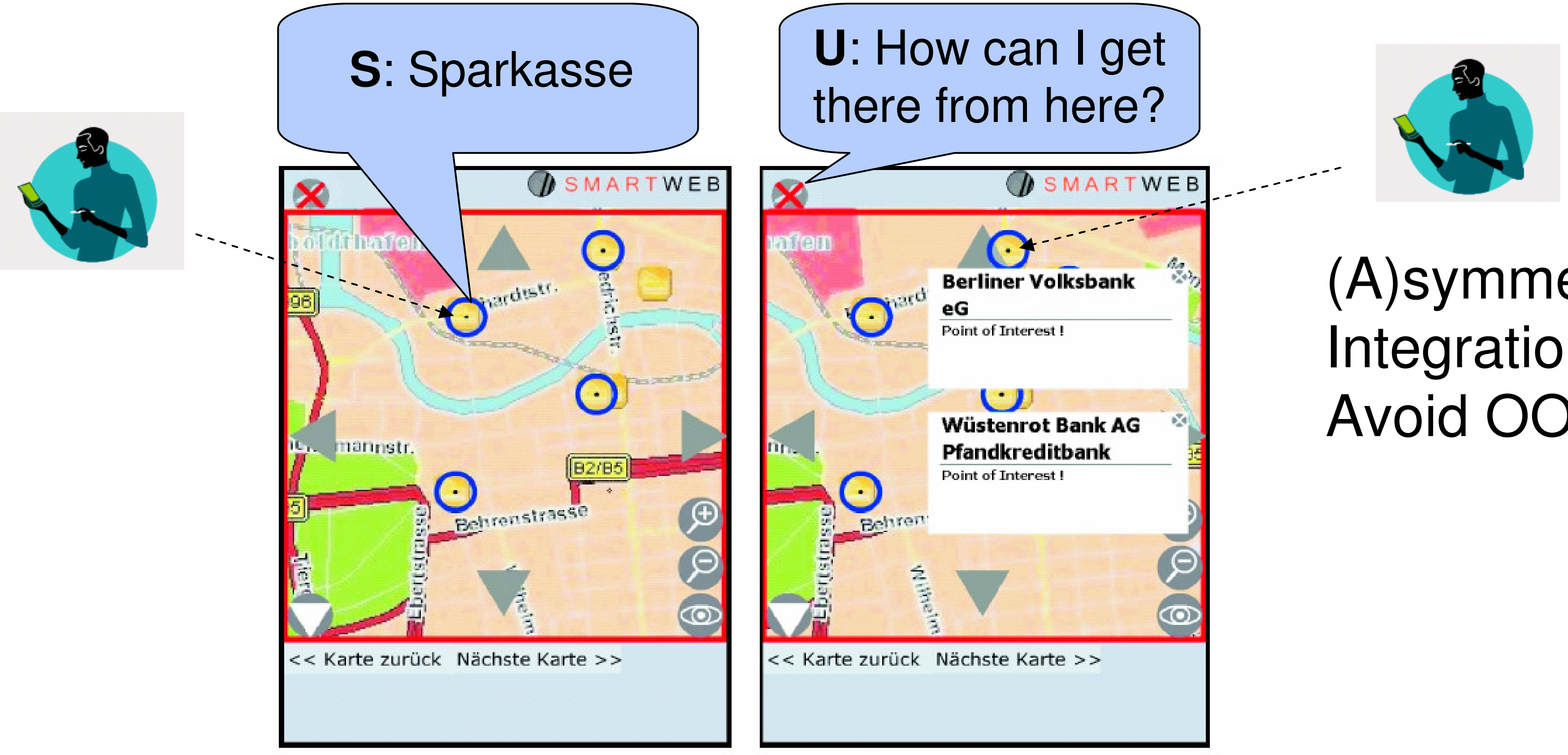
Navigation 5/6



Navigation 6/6



Narrowed Dialogue and Fusion Context in Composite Multimodal Interaction



S: Sparkasse

U: How can I get there from here?

Berliner Volksbank eG
Point of Interest !

Wüstenrot Bank AG Pfandkreditbank
Point of Interest !

<< Karte zurück Nächste Karte >>

Figure 4: POI and additional textual POI info selection

(A)symmetric Integration.
Avoid OOV.

Screen Transition as Dialogue Act

(U1) *Are these all hotels nearby?*	(S1) Changes to Area View

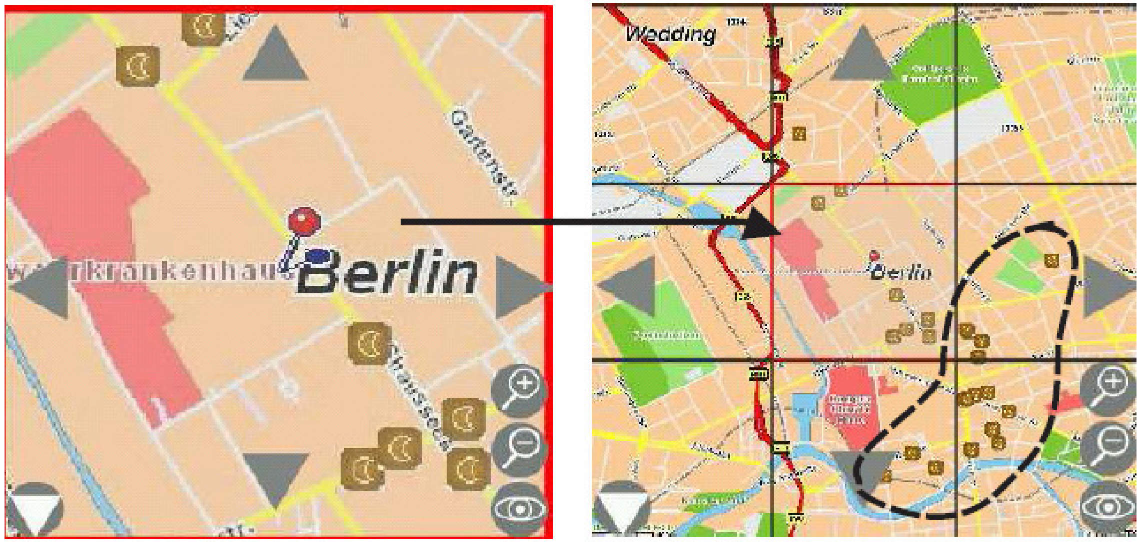
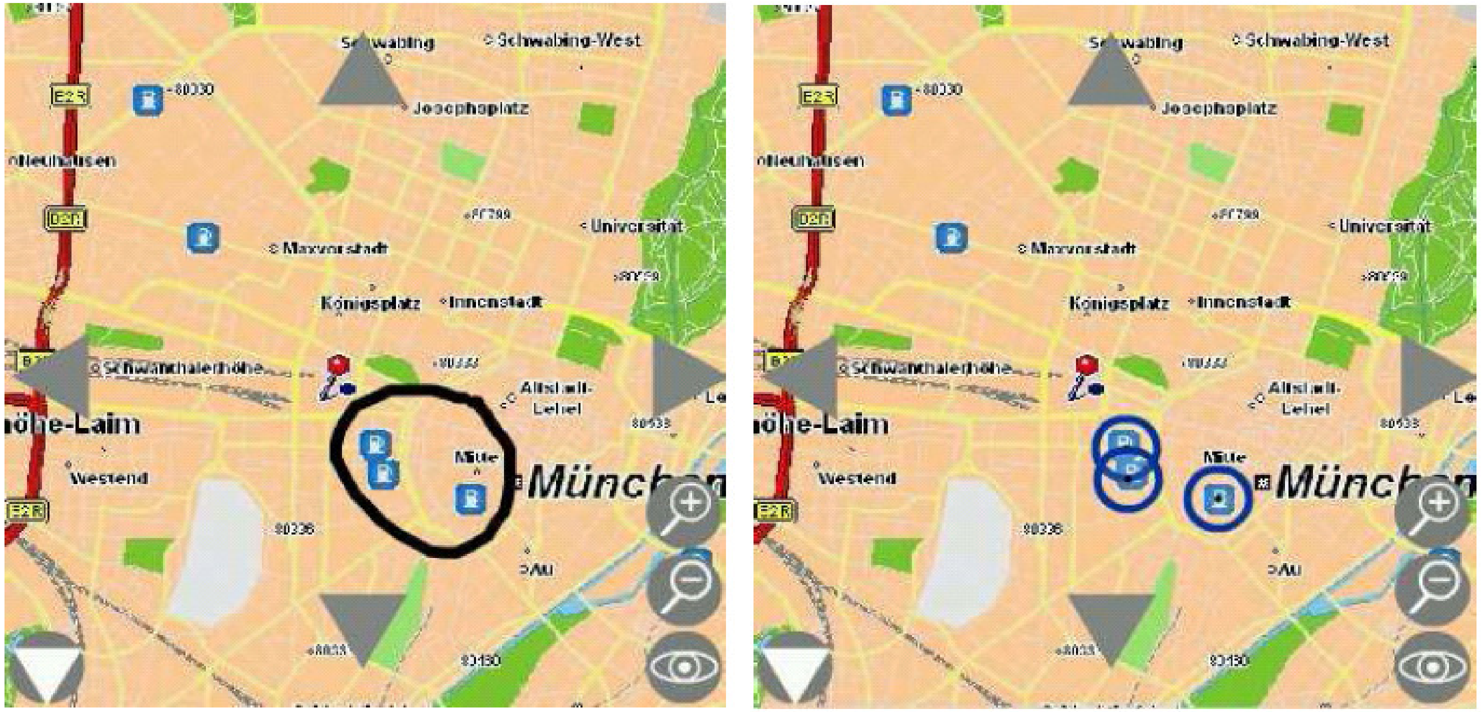


Figure 5: Multimedia presentation and transition as crucial part of a multimodal response dialogue act. The dashed line illustrates the set of additional hotels visible in Area View.

Mobility in Munich

(U2) Draws a circle on the screen + *Where can I get the cheapest diesel fuel?*

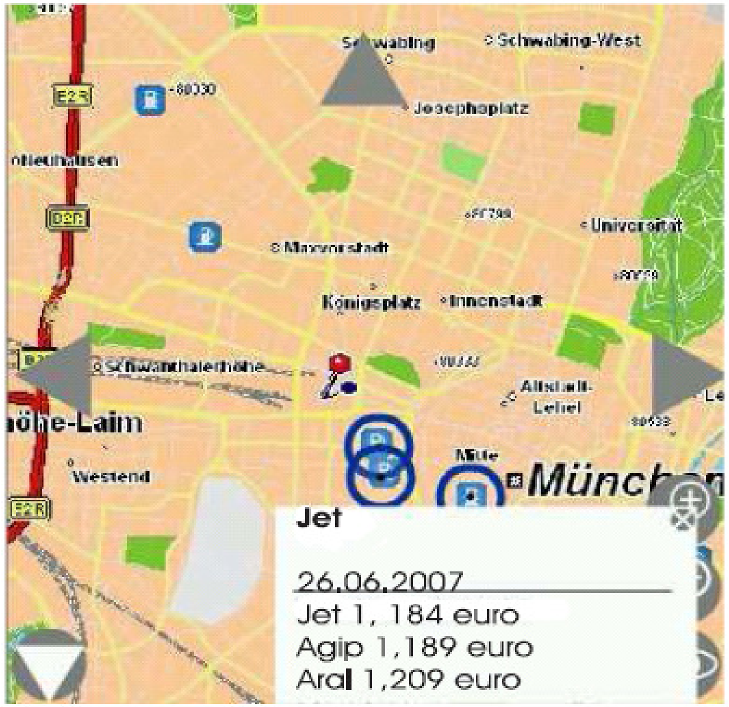
(S2) Draws circles as feedback and syntheses all diesel prices.



Mobility in Munich

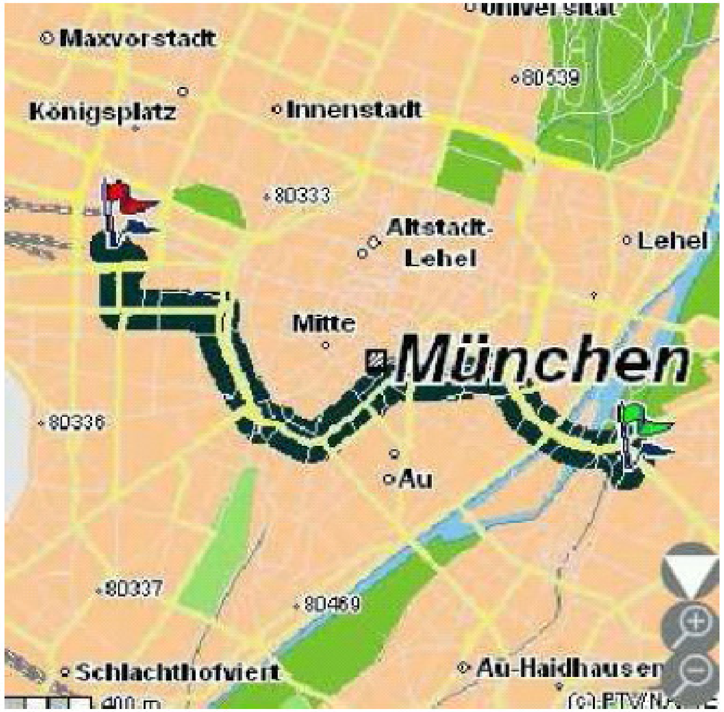
(U3) clicks on one of the circles to see the name of the gas station and the ranking.
(U4) chooses a station even further away and asks: *How can I get there (by car)?*

(S3) synthesises: Calculated route from Bayer Strasse to Hoch Strasse, München (3.6 km).
(S4) shows route in the map.

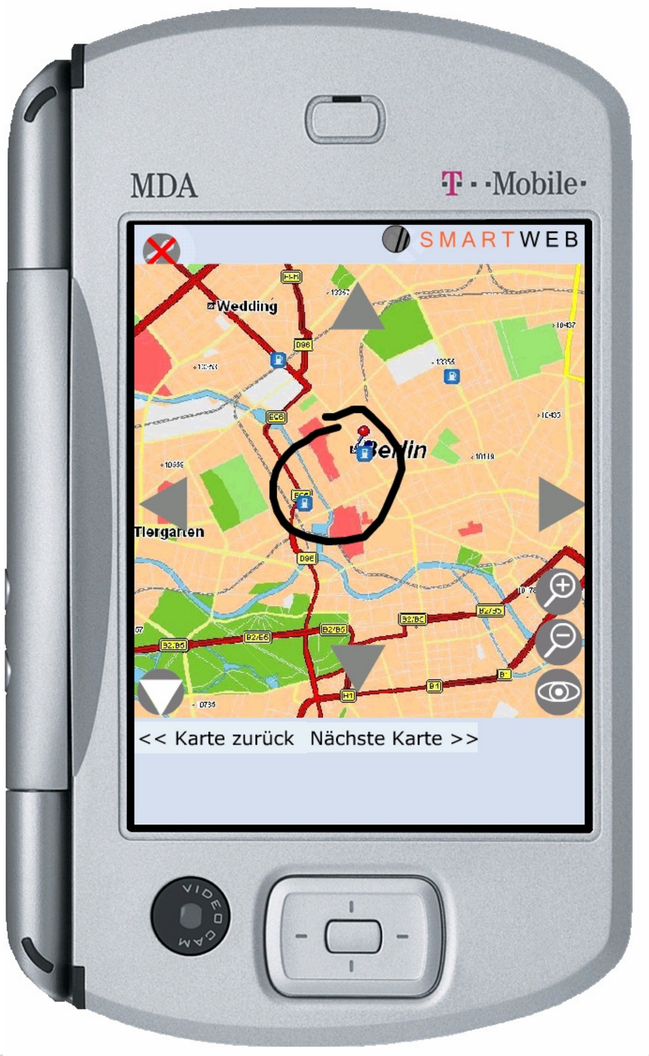


Jet

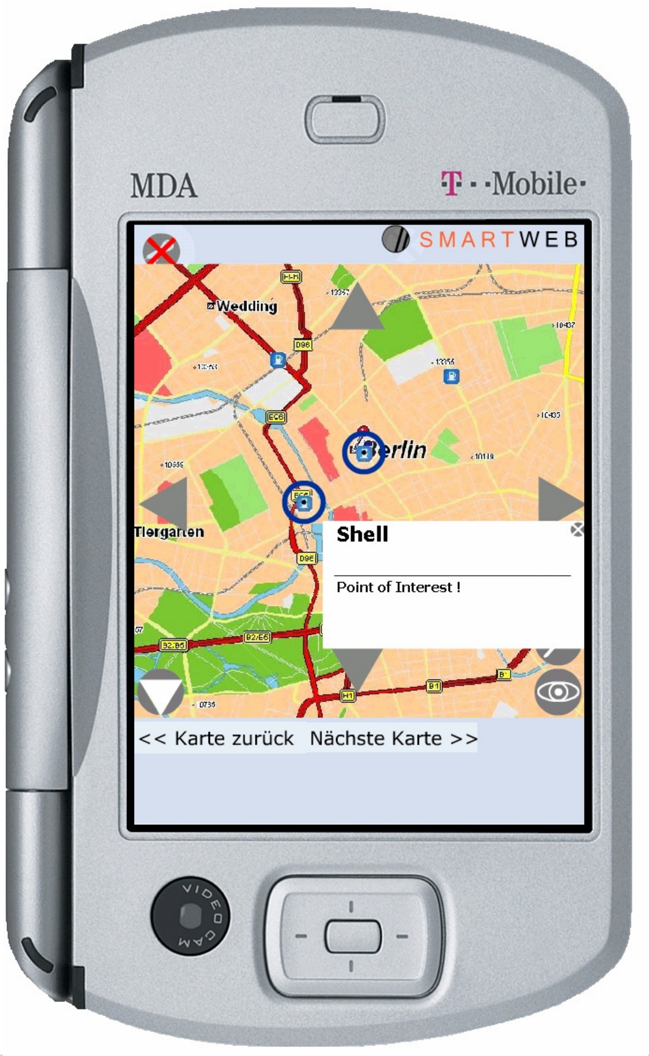
26.06.2007
Jet 1, 184 euro
Agip 1,189 euro
Aral 1,209 euro



Gesture 1/5


MDA
T · · Mobile ·
SMARTWEB
Wedding
Berlin
Tiergarten
<< Karte zurück
Nächste Karte >>
VIDEO CAM

Gesture 2/5

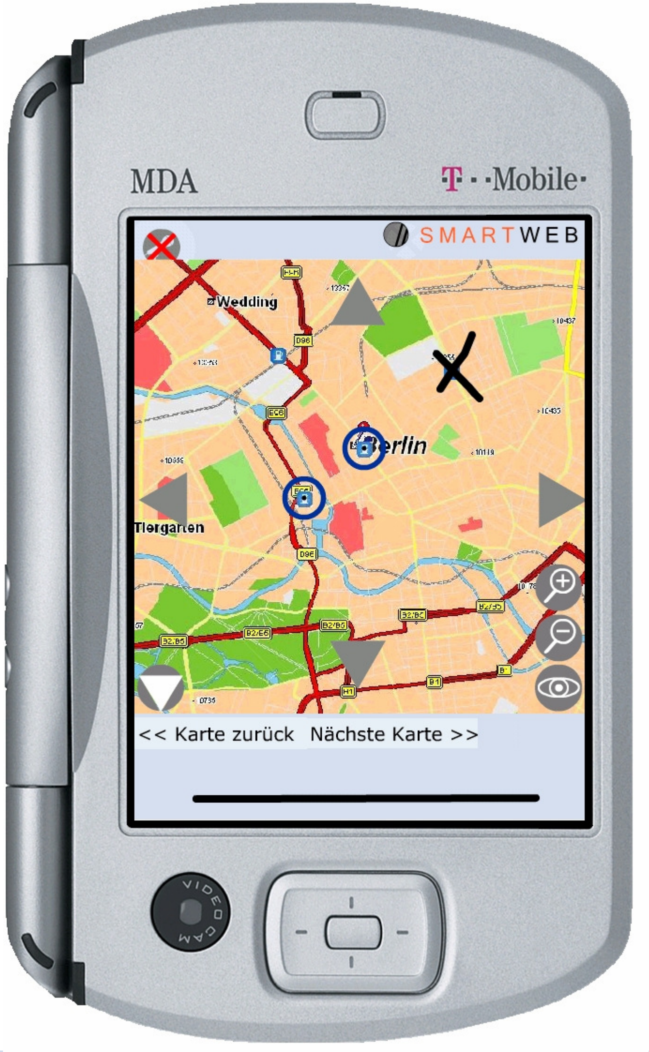


Shell

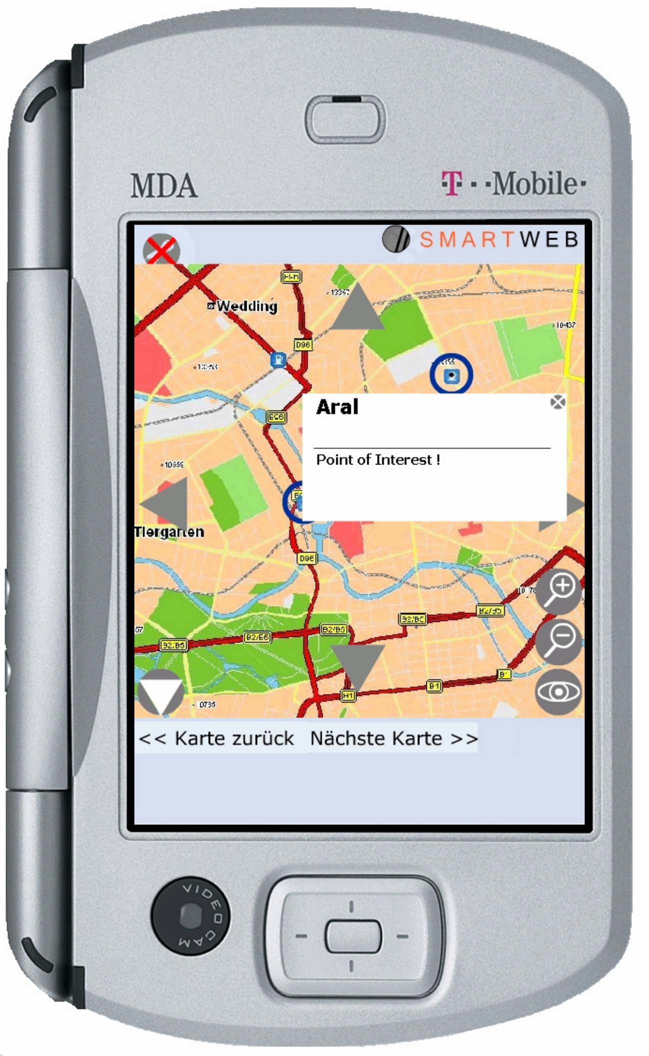
Point of Interest !

<< Karte zurück Nächste Karte >>

Gesture 3/5


MDA
T · · Mobile·
SMARTWEB
Wedding
Berlin
Tiergarten
<< Karte zurück
Nächste Karte >>
VIDEO CAM

Gesture 4/5

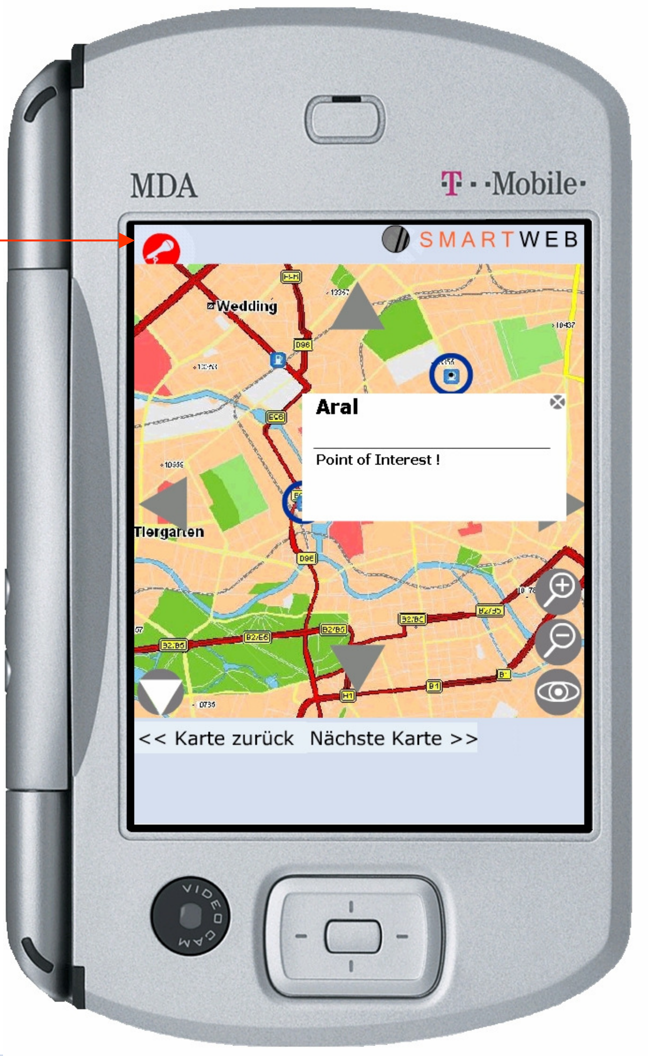


Aral

Point of Interest !

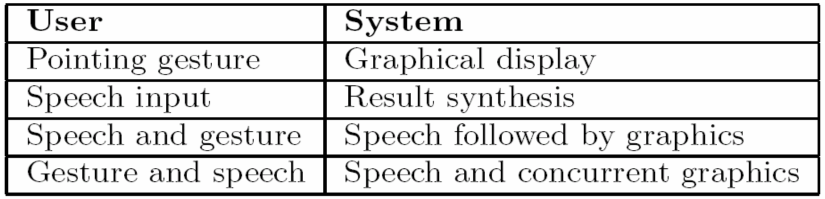
Gesture 5/5

Micro's open. Ask!



Conclusions

- We presented speech and gesture-based interaction with navigation maps.
  - Mobile interfaces in context-sensitive information-seeking scenario
  - Symmetric multimodal presentation behaviour (feedback)



User	System
Pointing gesture	Graphical display
Speech input	Result synthesis
Speech and gesture	Speech followed by graphics
Gesture and speech	Speech and concurrent graphics

- User utterances are quite predictable in map presentation context. That helps ASR and NLU.
- Extensions:
  - Explore more fine-gained co-ordination and synchronisation in multimodal/multimedia presentations.
  - Implement editing functions via concurrent pen and voice.

- Thank You!