Natural Human-Robot Interaction:
Audio-Visual Perception of Humans and of their communication modalities

Rainer Stiefelhagen
Interactive Systems Labs
Universität Karlsruhe

stiefel@ira.uka.de

Overview

- Natural Human-Robot Interaction
  - Motivation
- Visual Perception of People
  - 3D Body Tracking (Head, Hands, ...)
  - Gesture Recognition
  - Tracking of Head Pose and Focus of Attention
- Acoustic Perception
  - Speech Recognition (State of the Art, Problems, ...)
  - Emotions, Person Identification, Topic Tracking, ...
- Multimodal Dialogue
- Conclusion and Outlook

Natural Human-Robot Interaction

People use a variety of cues during face-to-face interaction:

- Visual:
  - Gestures, Pointing Gestures
  - Gaze / Attention
  - Facial Expressions
  - Body Language
  - Identity
- Acoustic:
  - Speech / Language
  - Tone of Voice / Emotions
  - Topic

We need to perceive and understand the full context:

Who, What, Where, Why, How, To Whom?

Our Scenario

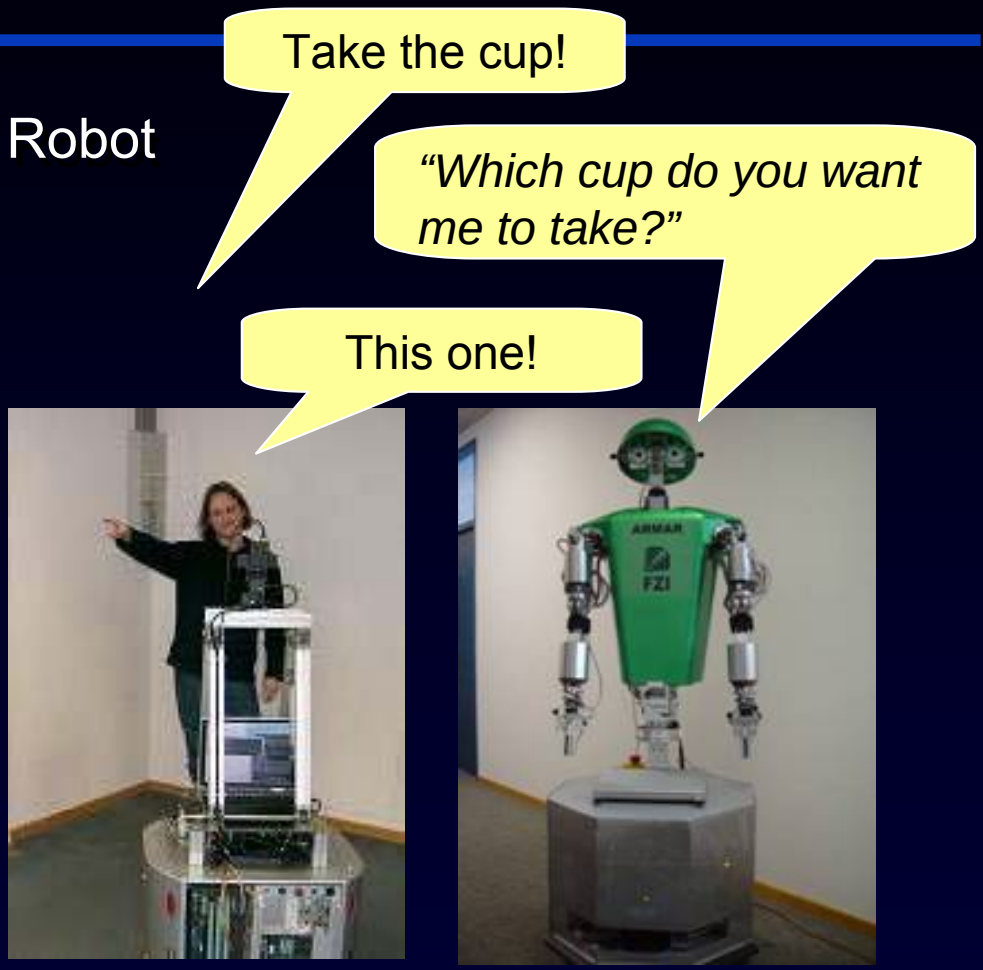
Multimodal Interaction with a Household Robot

Visual:

- Face Detection and Tracking
- Body-Tracking
- Detection of Pointing Gestures
- Head Orientation & Focus of Attention
- Face Recognition

Acoustic:

- Speech Recognition
- Emotion Recognition
- Dialog-Processing
- Speaker Identification

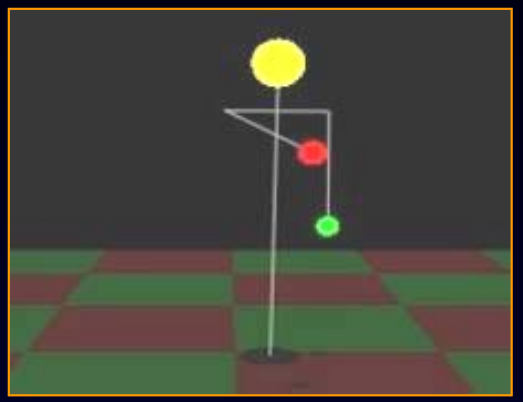


German Research Foundation (DFG):
Collaborative Research Center 588
"**Humanoid Robots - Learning and Cooperating Multimodal Robots**"

3D-Tracking of Head and Hands

Stereo camera

Left/right image



Extracted 3D model

- Combined use of skin-color and disparity features. Benefits:
  - 3D positions of head and hands
  - improved tracking due to a basic 3D-model of human body
- No markers, no manual initialization, no static background modeling
- 10 frames/sec on a 2.6GHz PC (for a single Person)

(Kai Nickel, R. Stiefelhagen, Humanoids 2003)

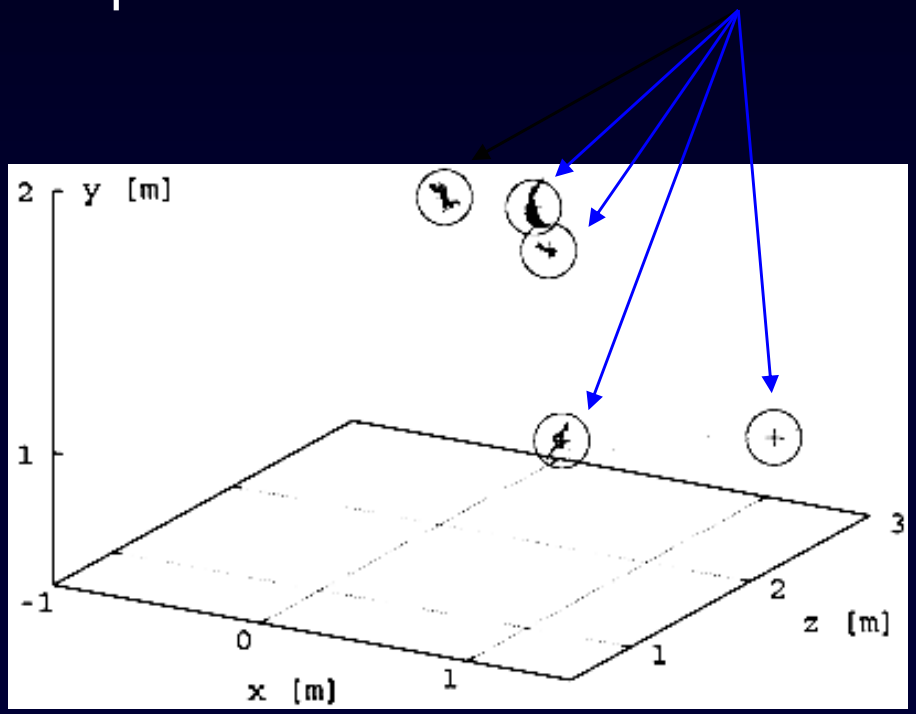
Locate Head-/Hand-Candidates

Dense disparity map

Skin color map

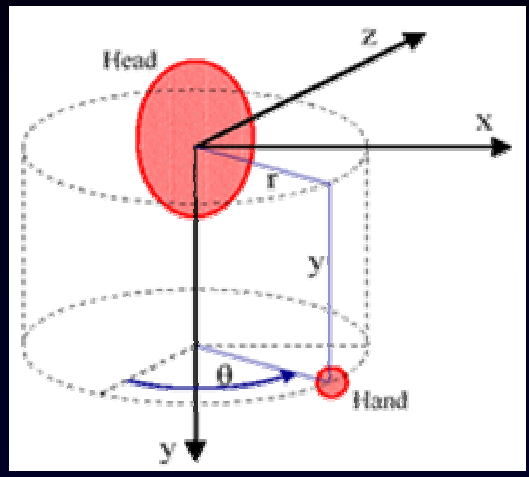
spatial k-means clustering

3D skin-pixel clusters represent possible head/hand locations.

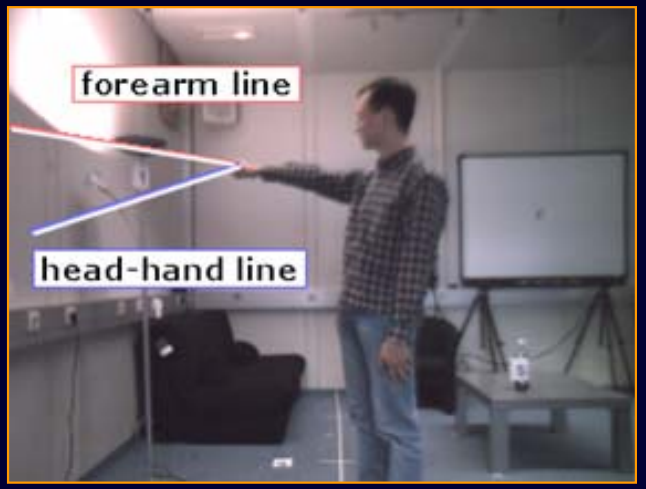


Pointing Gesture Recognition

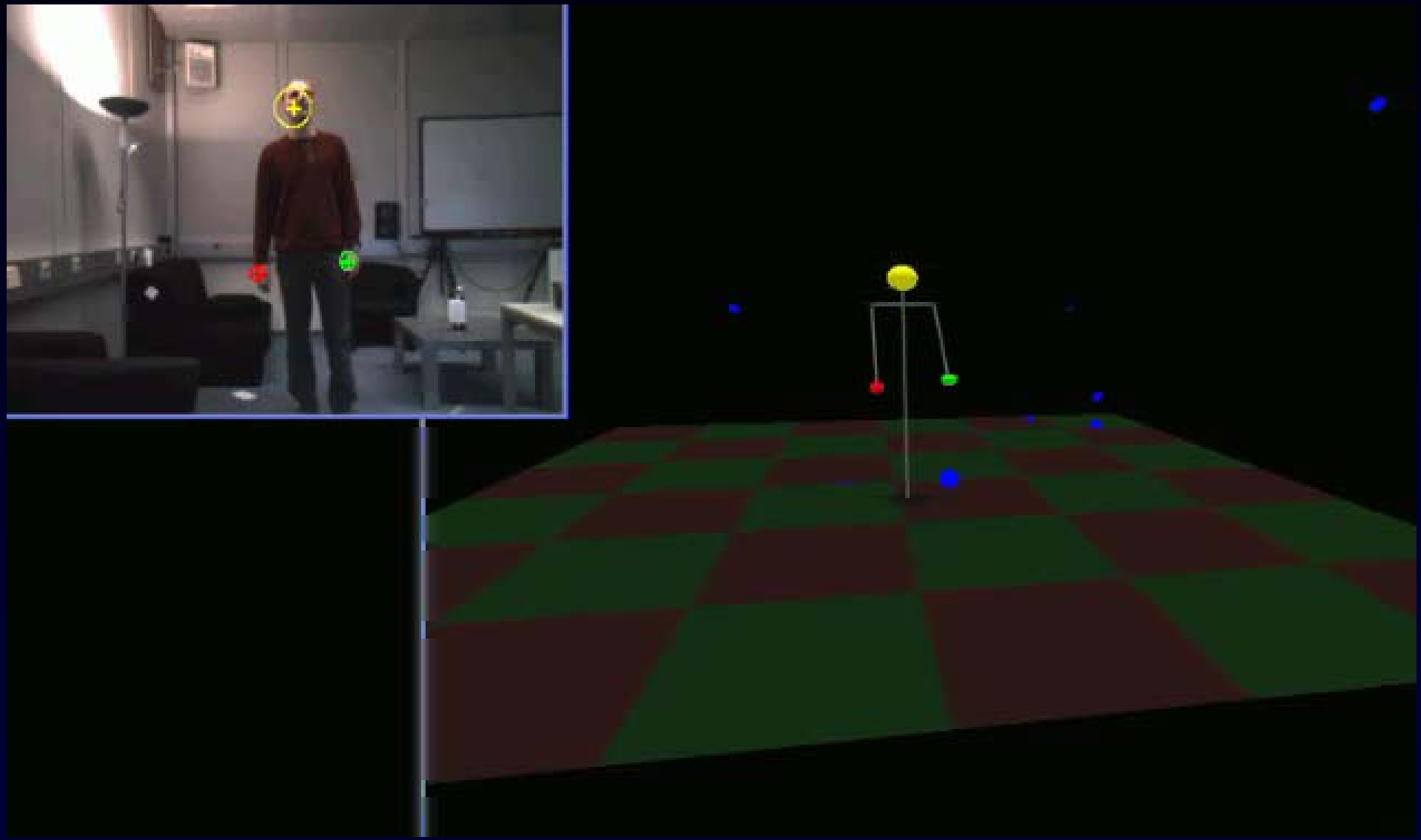
- HMMs to detect pointing gestures
  - 3 Models: Begin – Hold - End
  - Features: **(r, Δθ, Δy)**
  - **Online-Decoding**



- Estimating Pointing Direction
  - Estimated in Hold-Phase
  - A) Head-Hand-Line
  - B) Forearm-Direction



Pointing Gesture Recognition - Video



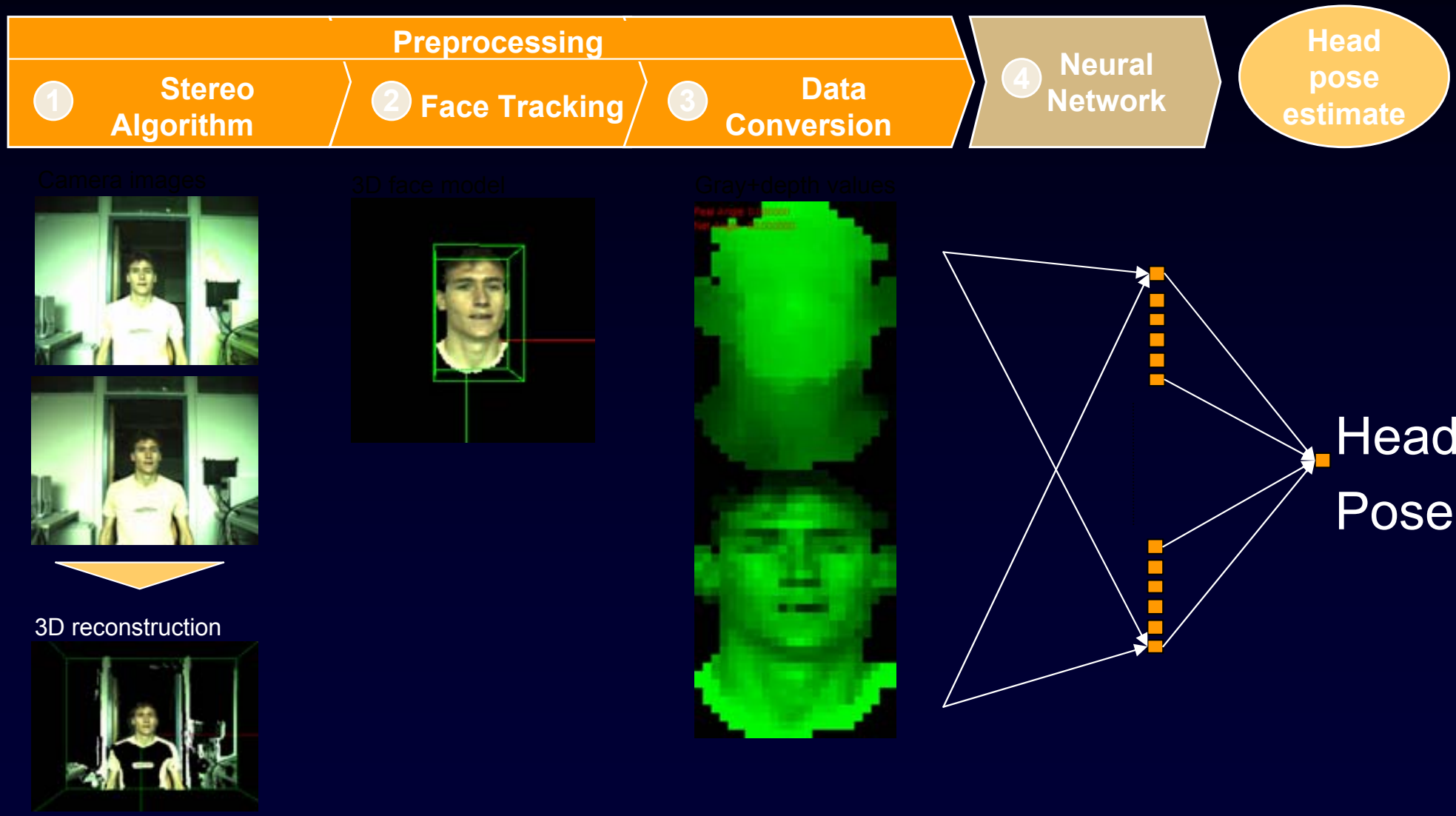
Gaze-Aware Human-Robot Interaction

Focus of Attention tracking:

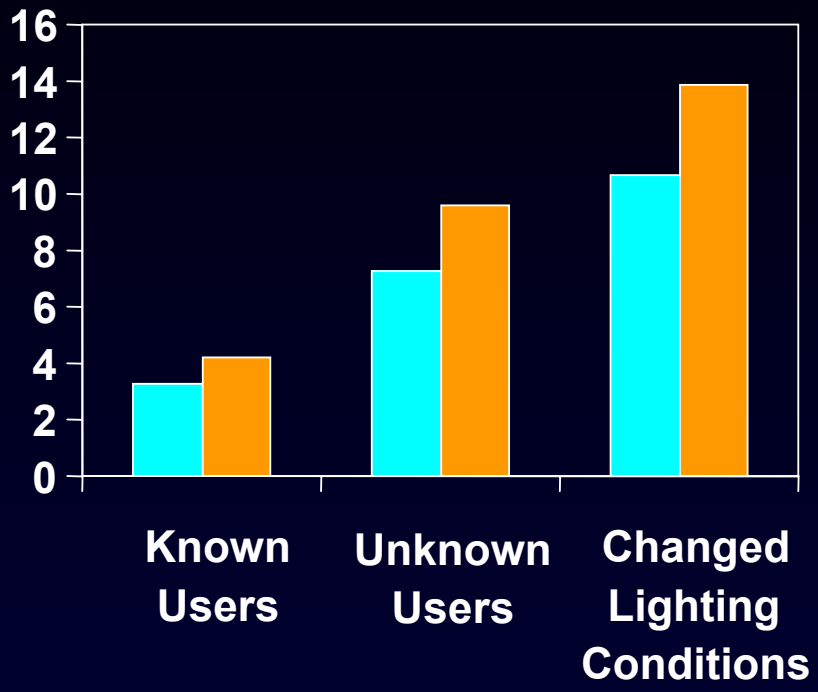
- to understand the user's actions/intentions
- to establish joint/shared attention
- to determine the addressee of a speech act
- „Is the robot addressed or someone else?"

- Head orientation is a good cue for a persons direction of attention
  (and it is much easier to detect than eye-gaze!)
  - Speech-based cues do also help !

Appearance-Based Head Pose Tracking



Head Pose Estimation: Results



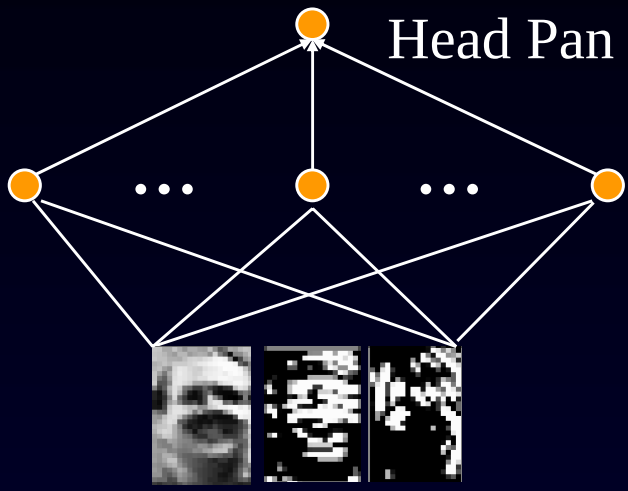
- Known users: ~ 4° mean error
- New Users: ~ 10° mean error
- Systems runs at ~ 10fps

disparity image → 30% relative improvement under different illumination!

(Seemann, Nickel, Stiefelhagen `03/´04)

Audio-Visual Estimation of Addressee

- Scenario: Two humans, one household robot
- Goal: Identify Target:
  - When was the robot adressed?



- Audio-visual estimation of Addressee
  - Visual: Based on Head Pose Estimation
  - Speech-based:
  - Utterance lenght, occurence of „Robot“, syntactical and semantical differences, sentence structure and parseability features

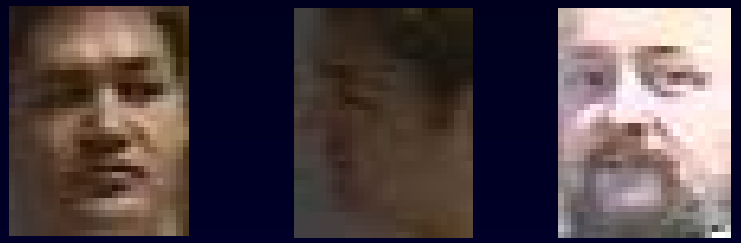
(M. Katzenmaier et al., ICMI 2004)

Estimation	Precis.	Recall	F-Meas.	Accu.
Acoustic	0.19	0.91	0.31	0.49
Head Pose	0.57	0.81	0.67	0.90
Combined	0.65	0.81	0.72	0.92

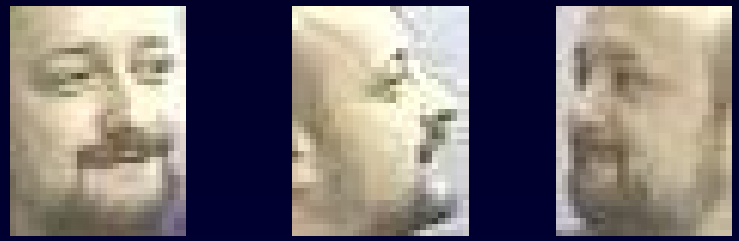
People Identification: Challenges

Low quality

Illumination



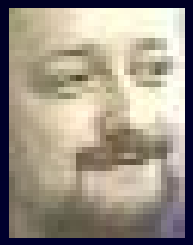
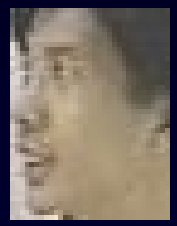
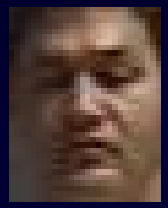
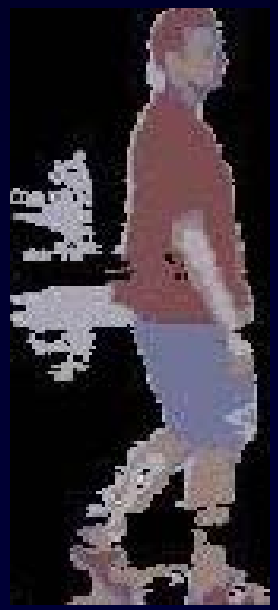
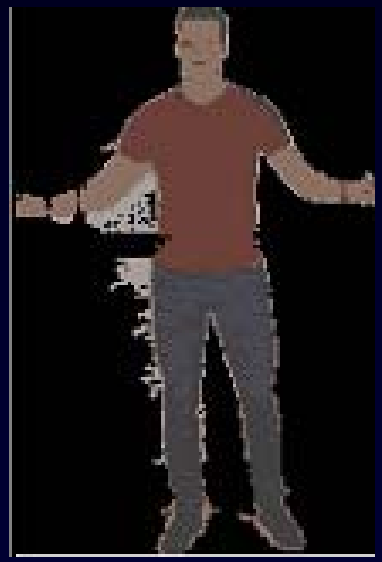
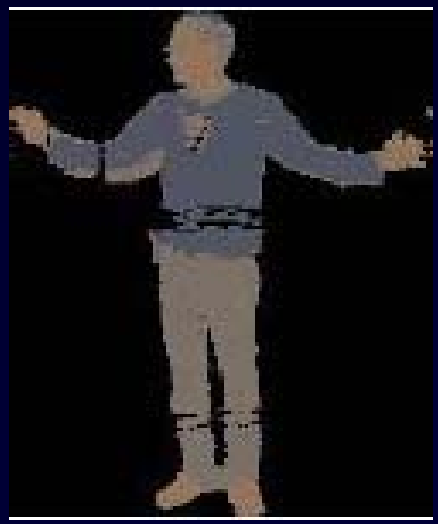
Head pose



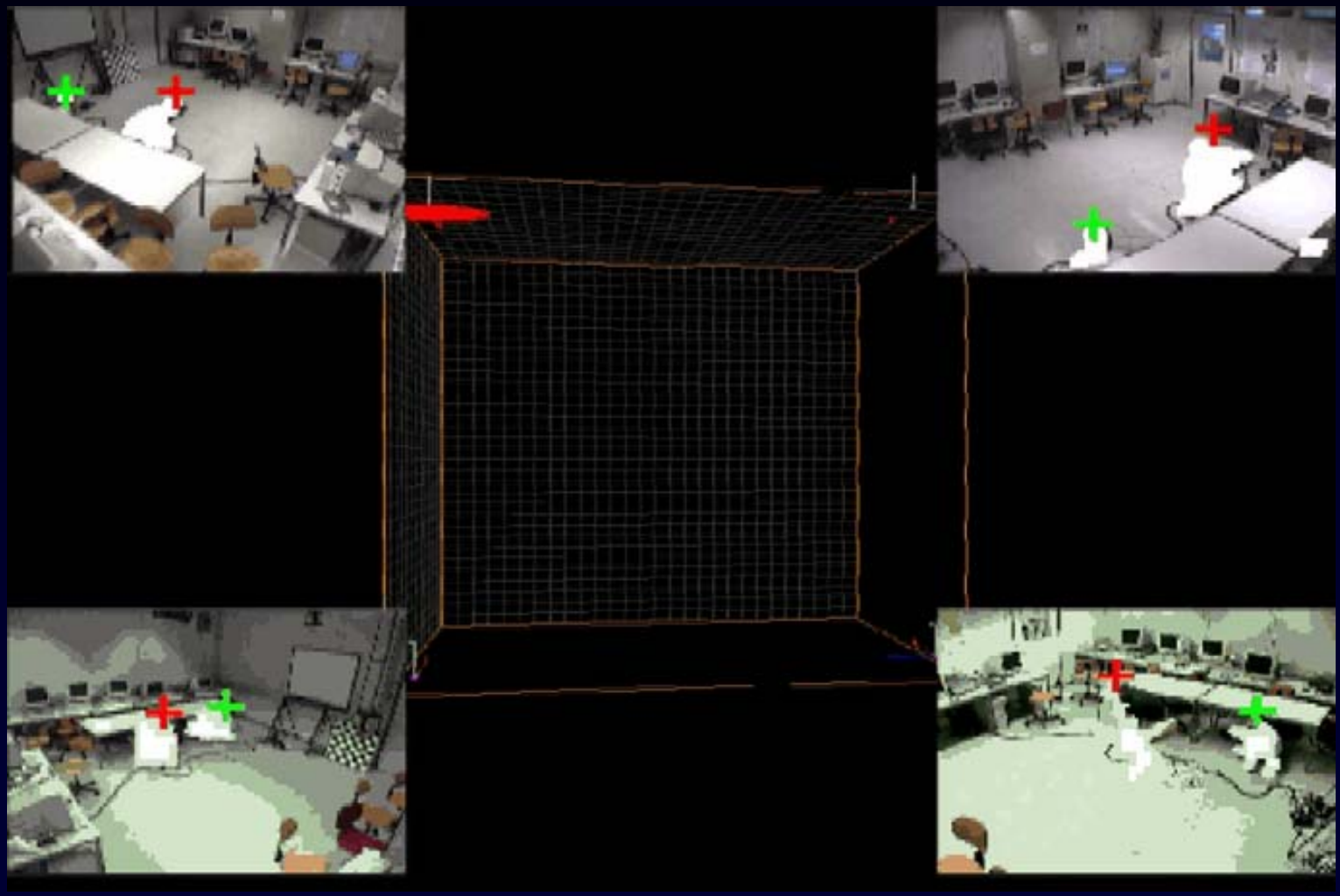
Occlusion

Tracking the Identities of People

- Combined use of
  - Face recognition
  - Speaker identification
  - Tracking of people‘s appearance (color models)

Where ? : 3D Person-Tracking with external cameras



- 4 calibrated cameras used
- Adaptive foreground segmentation
- probabilistic tracking (Kalman-Filter) (Focken & Stiefelhagen 2002)

What was said and meant ?

- **Large Vocabulary Speech Recognition**
  - Issues:
    - Sloppy Speech
    - Distant Microphones
    - Mismatch in Vocabulary
    - (Other Languages)
  - Many Other Aspects: Topic Detection, Named Entity, Translation, Discourse, ….
- **Multimodal Dialog**
  - Fuse Speech, Pointing, Gesture, Handwriting
  - Fusion Usually at Semantic Level
- **Audio-Visual Speech**
  - Combine Speech and Visual Info

JANUS-Speech Recognition Toolkit (JRTk)

- Unlimited and Open Vocabulary
- Spontaneous and Conversational Human-Human Speech
- Speaker-Independent
- High Bandwidth, Telephone, Car, Broadcast
- Languages: English, German, Spanish, French, Italian, Swedish, Portuguese, Korean, Japanese, Serbo-Croatian, Chinese, Shanghai, Arabic, Turkish, Russian, Tamil, Czech
- Best Performance on Public Benchmarks
  - DoD, (English) DARPA Hub-5 Test ‘96, ‘97 (SWB-Task)
  - Verbmobil (German) Benchmark ’95-’00 (Travel-Task)

Non-Verbal Cues in Speech

Transcript: Onune baksana be adam!

Turkish	Language ID
Bus Station	Acoustic Scene
Angry	Emotion ID
Negotiation	Discourse Analysis
Umut	Speaker ID
Chemicals	Topic ID
Istanbul	Entity Tracking

Handsfree, Always-On

- No Headset
  - Can’t Control the Microphone
  - Can’t Control the Recording Condition
  - Cross-Talk, Multiple Speaker
  - Noise
- No Push-to-Talk Button
  - Don’t know when to start and stop; always on !
  - Don’t know who is meant
  - Knowledge of the State of the World Becomes Important
  - Don’t Know the Topic and Goal of Speech
  - Input and Commands always Change

Speech Recognition – Distant Speech

Adapting the acoustic model using MLLR to different speaking distances (but not to the speaker!)

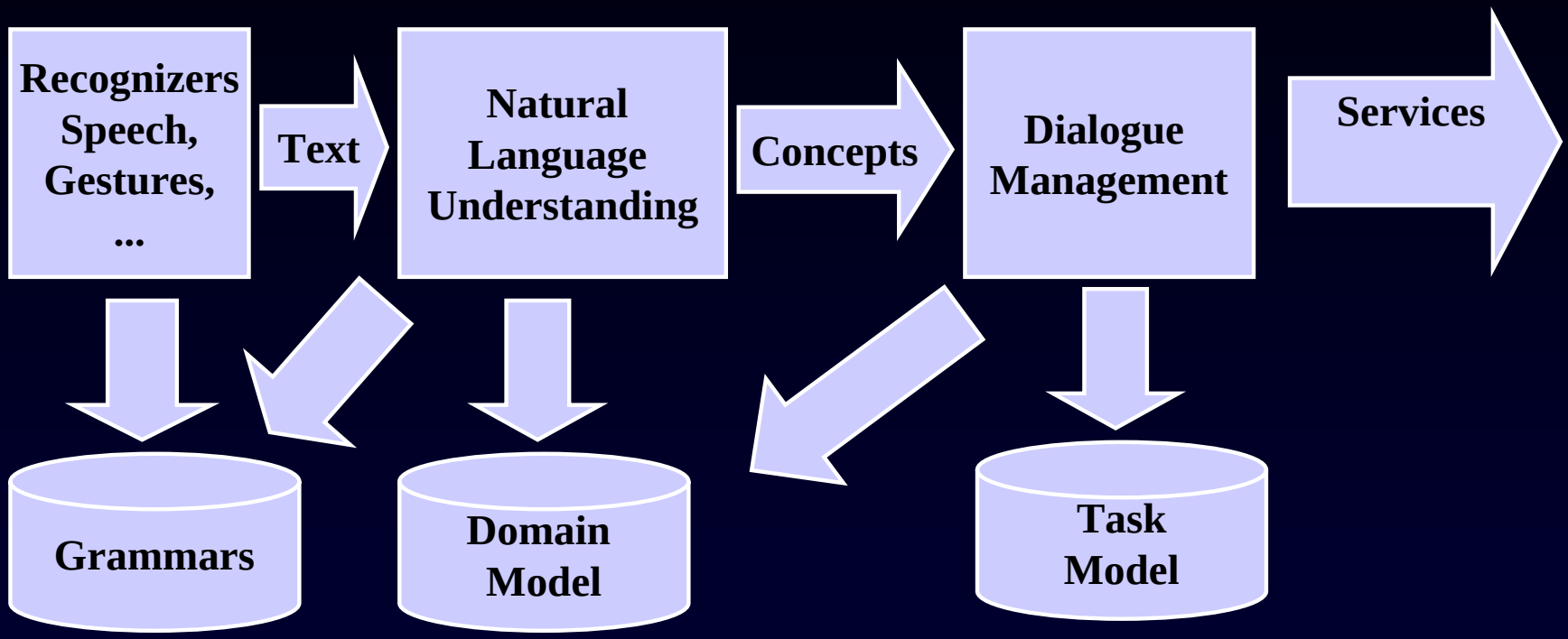
- Effect of unsupervised adaptation

WER	close	lapel	1.2m	1.5m	1.8m	2.4m
Unadapted	26.6%	29.7%	47.7%	51.9%	66.1%	69.3%
Adapted	26.5%	28.4%	42.5%	44.7%	59.7%	60.1%

- Sensibility of already adapted system

WER	1.2m	1.5m	1.8m
Adapted on 1.5m	42.4%	44.7%	60.1%

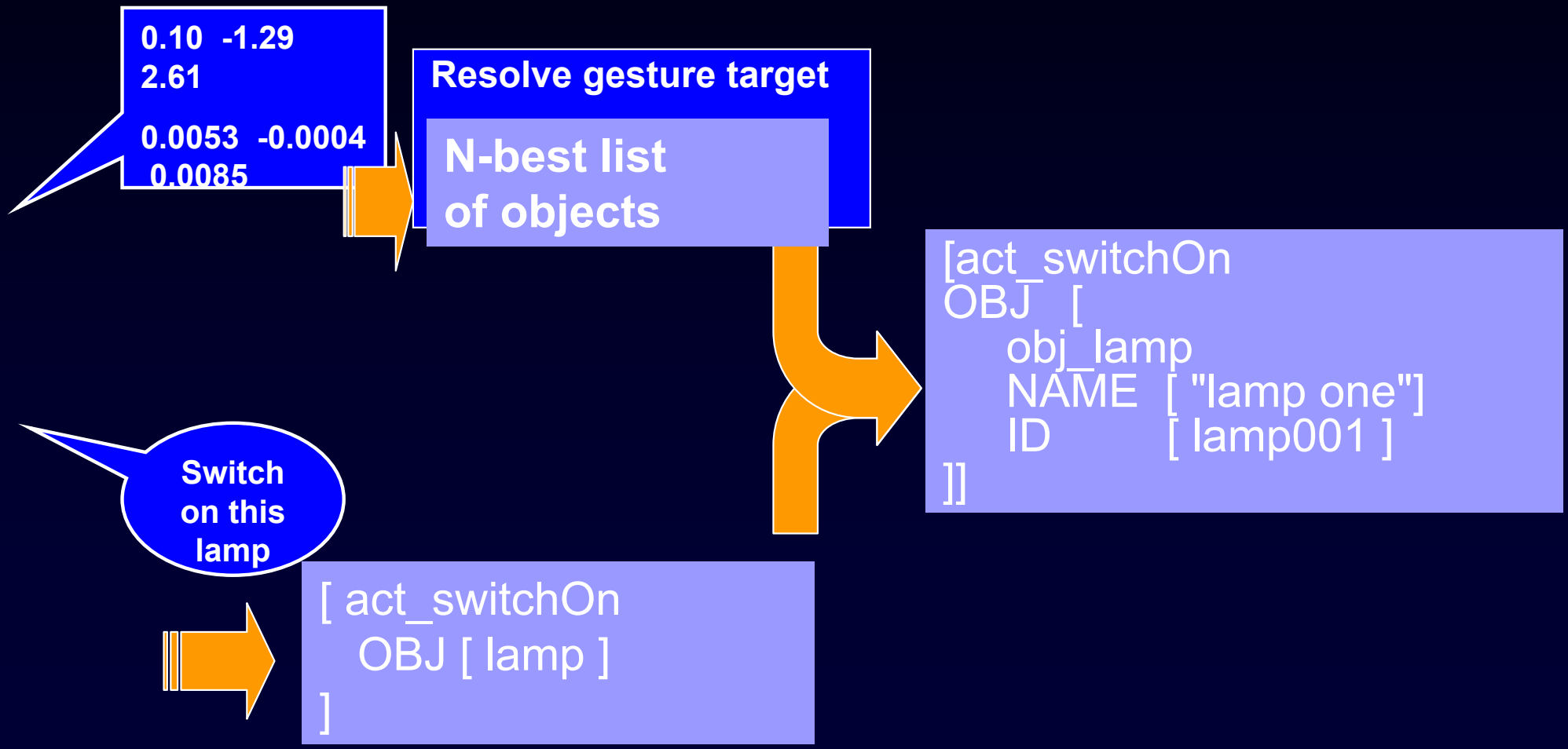
Human-Robot Dialog Processing



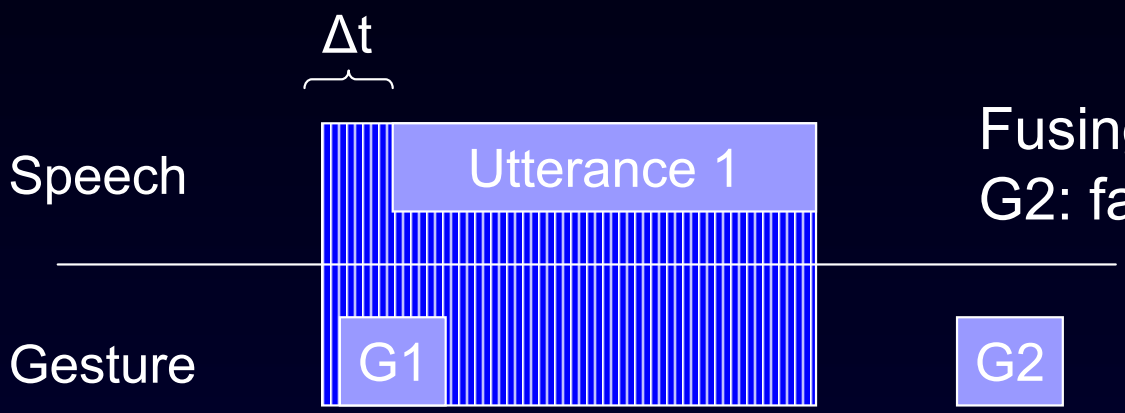
Integration & Fusion of several modalities is necessary:
Speech, Gestures, Focus of Attention, Emotions

Gieselmann, Fügen, Holzapfel, Schaaf, Waibel. Humanoids 2003.

Fusing Speech and Pointing Gestures

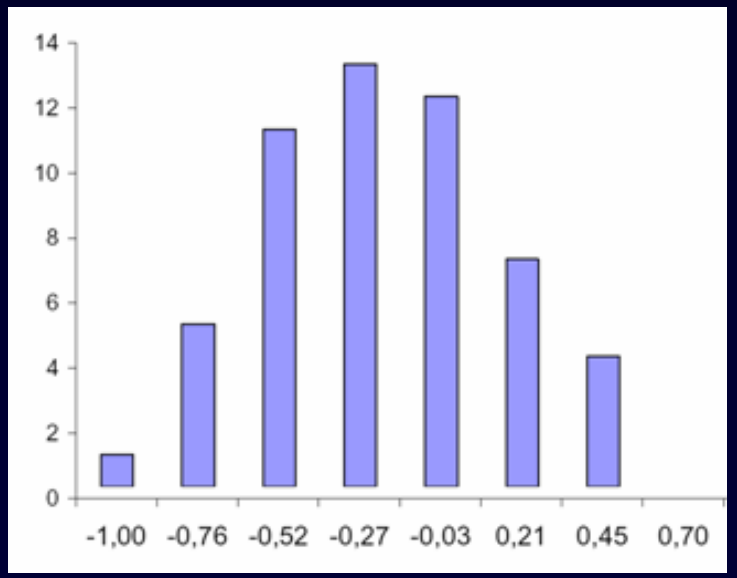


Multimodal Fusion



Fusing utterance 1 and G1
G2: false positive

Temporal correlation between
Speech and pointing gesture



Conclusions

- Interface on a Humanoid Robot Should
  - Operate Naturally around Humans
  - React to *Explicit and Implicit* Input
- The full context must be perceived and interpreted:
  - Who, What, Where, Why, How ?
  - Necessary technologies include: Person/Body Tracking, Identification, Head Pose / Attention, Gesture Recognition, Speech, Emotions, Language Understanding, Dialogue, …
- These technologies must improve with respect to
  - Robustness (noise, lighting conditions etc.)
  - Naturalness

Outlook

- **Long-term goal**:
  - Real-time perception and understanding of scene and user
  - Natural Human-Robot Interaction and Cooperation

- **Short / Mid-Term**:
  - More detailed body-tracking without any markers
  - Improved tracking of user‘s focus of attention
    - Head / Body Orientation
    - Gestures
    - Dynamic Scenes
  - Robust Person Recognition (Face, Speech, Tracking)
  - Object Recognition
  - Attentional Mechanism / Orienting Behavior of Robot

Acknowledgements

- **Thank you for your attention !**

- The team:
  - K. Nickel: Person Tracking, Pointing Gestures
  - H. Ekenel: Face Recognition
  - C. Fügen: Speech Recognition & Dialogue
  - P. Gieselmann: Dialogue Modelling
  - H. Holzapfel: Dialogues & Fusion
  - E. Seemann, M. Katzenmaier, D. Harres, D. Kern (Students, Vision)
  - Alex Waibel (Director of the ISL)

- **Contact:**
  - stiefel@ira.uka.de, http//isl.ira.uka.de/~stiefel

- This work is partly funded by the DFG under Sonderforschungsbereich 588 „Humanoide Roboter“