Gesture Recognition

Vorlesung
„Computer Vision für Mensch-Maschine Interaktion“

Rainer Stiefelhagen

2014-01-31

Overview

- Introduction
- Hidden Markov Models
- Applications
  - American Sign Language
  - Pointing Gestures
- Miscellaneous

Definition

Gesture:

- a movement usually of the body or limbs that expresses or emphasizes an idea, sentiment, or attitude
- the use of motions of the limbs or body as a means of expression

[Merriam-Webster Online Dictionary]

Automatic Gesture Recognition

- A gesture recognition system generates a *semantic description* for certain body motions
- Gesture recognition exploits the power of *non-verbal communication*, which is very common in human-human interaction
- Gesture recognition is often built on top of a *human motion tracker*
- Related topics: Human Activity Analysis, Intention Recognition, Facial Expression Recognition

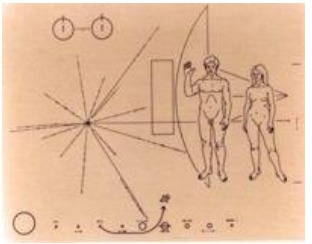
Some Applications

- Virtual Environments
  - manipulation of objects
- Multimodal Rooms
  - interaction with room facilities
  - analysis of a user‘s actions
    → context aware services
- Human-Robot Interaction
  - communicative gestures
  - programming by demonstration
- Understanding Human-Human Interaction
  - level of arousal
  - turn-taking

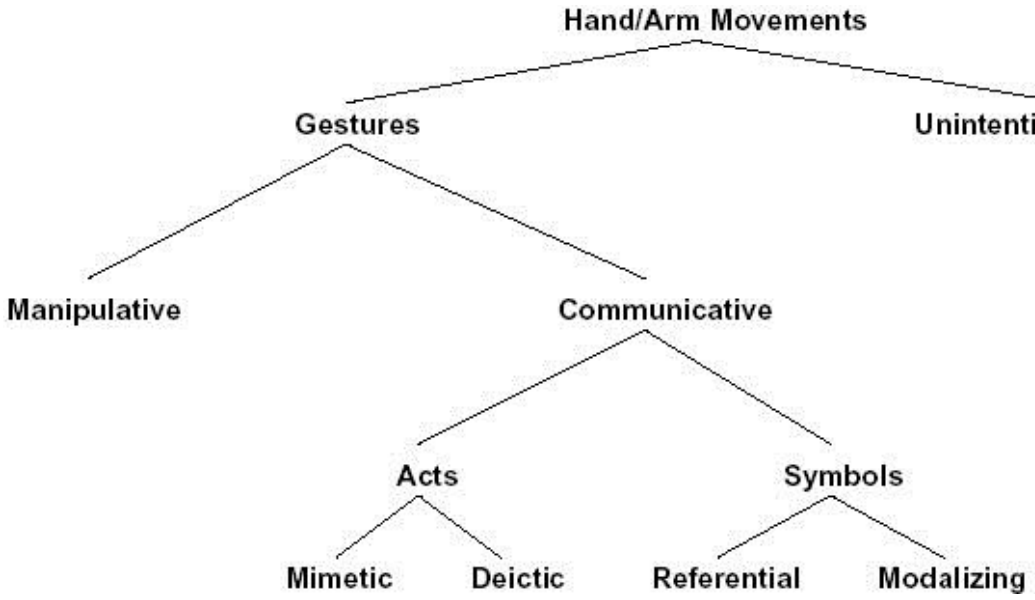
Types of Gestures

- Hand & arm gestures
  - Pointing Gestures
  - Sign Language
  - Hello, Thumbs up/down, victory sign, the finger, call-me, … (Wikipedia lists around 40 hand gestures)
- Head gestures
  - Nodding, head shaking, turning, pointing
- Body gestures
  - Don't know / shrug

- Gestures are mostly culture-specific
  - Pointing gesture probably one of the exceptions
- Some gestures closely coordinated with speech



Gesture Taxonomy



[Pavlović97]

- static vs. dynamic gestures
- different phases of a gesture: preparation, peak, retraction

A Purpose Taxonomy of Gestural Interaction

[Quek, ICMI05]

- Manipulative Gestures
  - "Put that there" [Bolt 1980]
  - Navigation in virtual spaces, game control, …
  - Robot control
- Semaphoric Gestures
  - Precisely defined, specific symbols of an alphabet
  - Static: Finger menu selection, hand pose classification, …
  - Dynamic: Crane operator signals, editing gestures for interactive whiteboard, …
- Conversational Gestures
  - Naturally performed in the course of human multimodal communication

Spontaneous gestures [Cassell98]
(Communicative Gestures)

- ***Iconic* gestures**
  Depict by the form of the gesture some feature/action/event being described

- ***Metaphoric gestures***
  *A*lso representational, but the concept they represent has no physical form; form of the gesture comes from a common metaphor. Example: "the meeting went on and on" + hand indicating rolling motion.

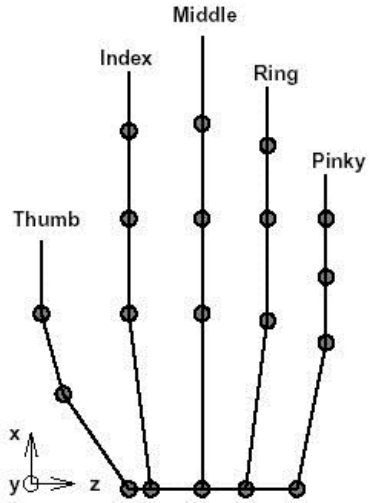
- ***Deictics***
  Spatialize, or locate in the physical space in front of the narrator, aspects of the discourse.

- ***Beat gestures***
  *S*mall baton like movements; do not change in form with content of speech. Pragmatic function, occurring with comments on one's own linguistic contribution, speech repairs and reported speech. Example: "she talked first, I mean second" + hand flicking down and up.

See also:
McNeill, D. 1992. *Hand and Mind: What gestures reveal about thought.* The University of Chicago Press, Chicago IL.

Spatial Models

- 3D model-based
  - 3D hand model. Parameters: angels, palm position



- Appearance-based
  - Parameters:

    Pixel templates, image geometry parameters, Image motion parameters, Fingertip position & motion... [Pavlović]
  - Use templates for modelling gestures

Gesture Recognition

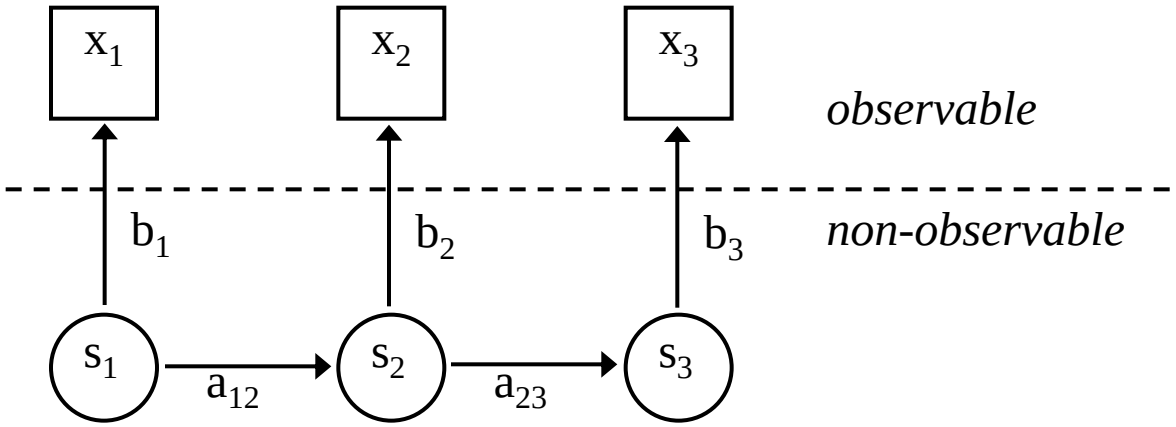
Feature acquisition:
- attached sensors
- visually
  - markers
  - color
  - motion
  - shape
  - …

Classifiers:
- **Hidden Markov Models (HMM)**
- Neural Networks (ANN)
- Finite State Machines (FSM)
- Support Vector Machines (SVM)
- …

Hidden Markov Models

- Gesture recognition ← temporal pattern matching
- HMMs
  - represent templates in a person-independent way (when trained with examples from many persons)
  - provide an inherently “soft” time-alignment mechanism
  - are based on a sound mathematical foundation
- Introduction to HMM → [Rabiner89]

Motivation



- the term "**hidden**" comes from observing observations and drawing conclusions without knowing the *hidden* sequence of states
- **Markov assumption** (1st order): the next state depends only on the current state (not on the complete state history)

HMM Definition

A Hidden Markov Model is a five-tuple (S,π,A,B,V).

It consists of:

- the set of **States** $S = \{s_1,s_2,...,s_n\}$
- the **initial probability** distribution π
  $\pi(s_i)$ = probability of s_i being the first state of a state sequence
- the matrix of **state transition probabilities** A
  $A=(a_{ij})$ where a_{ij} is the probability of state s_j following s_i
- the set of **emission probability** distributions/densities B
  $B=\{b_1,b_2,...,b_n\}$ where $b_i(x)$ is the probability of observing x when the system is in state s_i
- the observable **feature space** V can be discrete
  $V=\{x_1,x_2,...,x_v\}$, or continuous $V=\mathbf{R}^d$

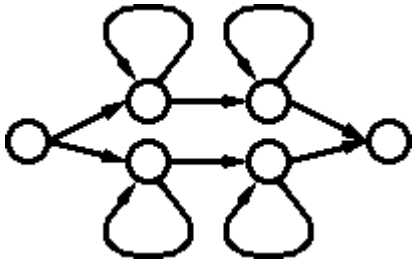
Some Properties of HMMs

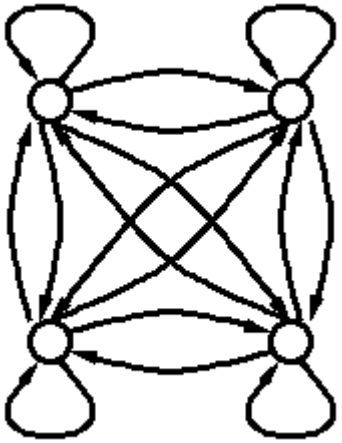
- for the initial probabilities we have: $\Sigma_i \pi(s_i) = 1$
- often things are simplified by $\pi(s_1)=1$, and $\pi(s_i>1) = 0$
- obviously: $\Sigma_j a_{ij} = 1$ for all i
- often: $a_{ij} = 0$ for most j except for a few states
- when $V = \{x_1,x_2,...,x_v\}$ then b_i are discrete probability distributions, the HMMs are called discrete HMMs
- when $V = \mathbf{R}^d$ then b_i are continuous probability density functions, the HMMs are called continuous (density) HMMs

HMM Topologies

left-to-right

alternative paths





ergodic

The Observation Model

Most popular: Gaussian mixture models

$$P(x_t|s_j) = \sum_{k=1}^{n_j} c_{jk} \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma_{jk}|}} e^{-\frac{1}{2}(x_t - \mu_{jk})^T \Sigma_{jk}^{-1} (x_t - \mu_{jk})}$$

n_j: number of Gaussians (in state j)
c_{jk}: mixture weight for k-th Gaussian (in state j)
μ_{jk}: means of k-th Gaussian (in state j)
Σ_{jk}: covariane matrix of k-th Gaussian (in state j)

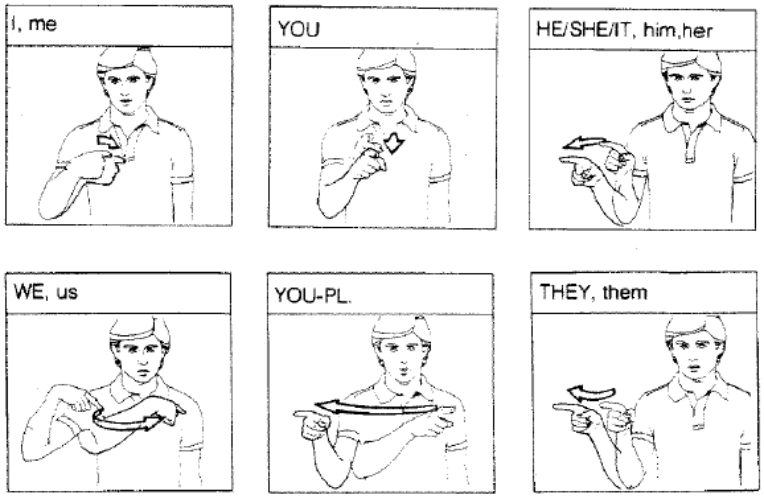
Three main problems of HMMs (That can be solved!)

Given an HMM λ and an observation $x_1,x_2,...,x_T$

- **The evaluation problem:**
  compute the probability of the observation
  $p(x_1,x_2,...,x_T \mid \lambda)$
- **The decoding problem:**
  compute the most likely state sequence $s_{q1},s_{q2},...,s_{qT}$,
  i.e. $\text{argmax}_{q1,..,qT}\, p(q_1,..,q_T \mid x_1,x_2,...,x_T, \lambda)$
- **The learning/optimization problem:**
  find an HMM λ' such that
  $p(x_1,x_2,...,x_T \mid \lambda') > p(x_1,x_2,...,x_T \mid \lambda)$

Sign Language Recognition [Starner98]

- American Sign Language (ASL)
  - 6000 gesture describe persons, places and things
  - Exact meaning and strong rules of context and grammar for each



- Sign recognition
  - Previous with instrumental gloves and neural nets
  - HMM ideal for complex and structured hand gestures of ASL

ASL Test Lexicon

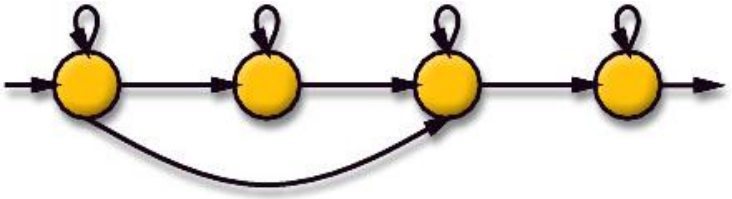
Vocabulary size: 40 words

Part of speech	*Vocabulary*
Pronoun	I, you, he, we, you (pl), they
Verb	want, like, lose, dontwant, dontlike, love, pack, hit, loan
Noun	box, car, book, table, paper, pants, bicycle, bottle, can, wristwatch, umbrella, coat, pencil, shoes, food, magazine, fish, mouse, pill, bowl
Adjective	red, brown, black, gray, yellow

Feature Extraction

- Camera either located as a 1st-person and a 2nd-person view
- Segment hand blobs by a skin color model
- Extracted Features: x, y, Δx, Δy, size, blob ”angle”, eccentricity of ellipse

HMM for American Sign Language



- Four-State HMM for each word
- Training:
  - Automatic segmentation of sentences in five portions
  - Initial estimates by iterative Viterbi-alignment
  - Then Baum-Welch re-estimation
  - No context used
- Recognition
  - With and without part-of-speech grammar (pronoun, verb, noun, adjective, pronoun)
  - All features / only relative features used

ASL Results: Desk-based

- 348 training and 94 testing sentences without contexts
- Accuracy: $$Acc = \frac{N - D - S - I}{N}$$

N: #Words
D: #Deletions
S: #Substitutions
I: #Insertions

Experiment	*Training set*	*Independent test set*
All features	94.1%	91.9%
Relative features	89.6%	87.2%
All features & unrestricted grammar	81.0% (D=31, S=287, I=137, N=2390)	74.5% (D=3, S=76, I=41, N=470)

ASL Results: Wearable-based

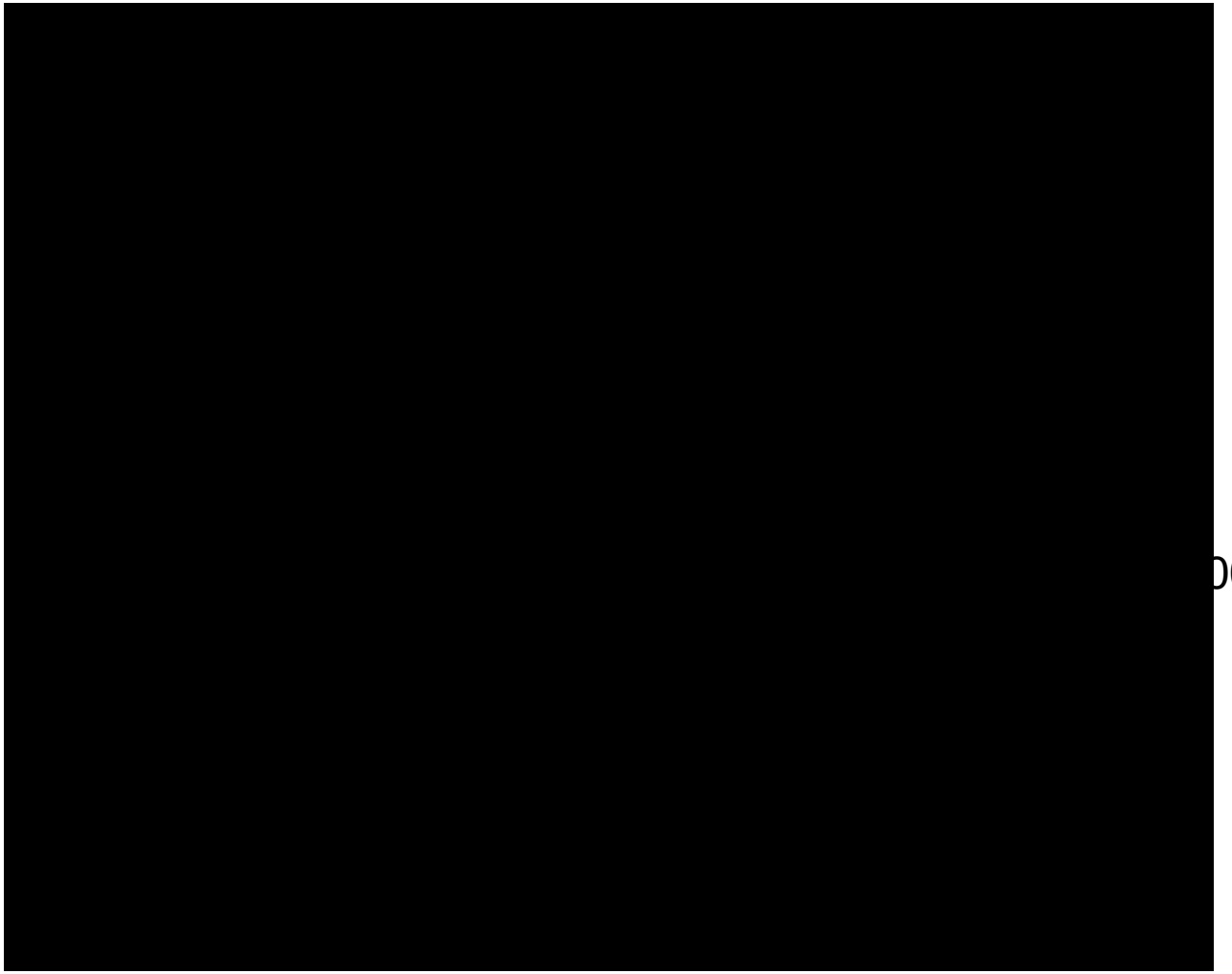
- 400 training sentences and 100 for testing
- Test 5-word sentences
- Restricted and unrestricted similar!

Experiment	*Training set*	*Independent test set*
Part-of-speech	99.3%	97.8%
5-word sentence	98.2%	97.8%
unrestricted	96.4% (D=24, S=32, I=45, N=2500)	96.8% (D=4, S=6, I=6, N=500)

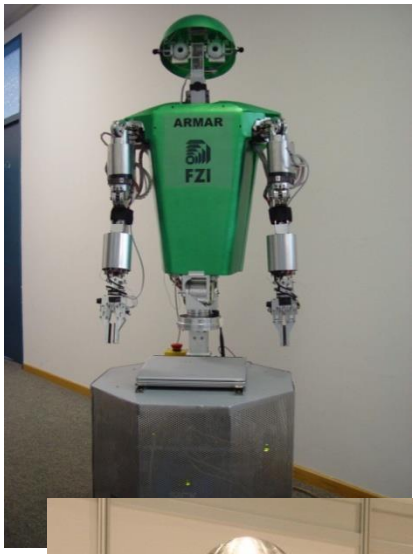
Pointing Gesture Recognition

- Pointing gestures
  - are used to specify objects and locations
  - can be needful to resolve ambiguities in verbal statements:
    „Put that there!“
- Here:
  Pointing gesture = movement of the arm towards a pointing target
- Tasks:
  - Detect occurrence of human pointing gestures in natural arm movements
  - Extract the 3D pointing direction

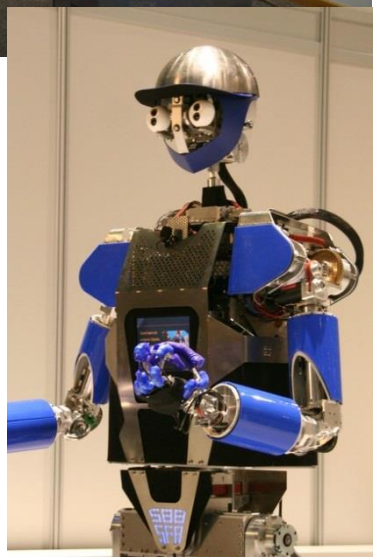
Application: Multimodal Dialogue for Human-Robot Interaction



(2004) Video

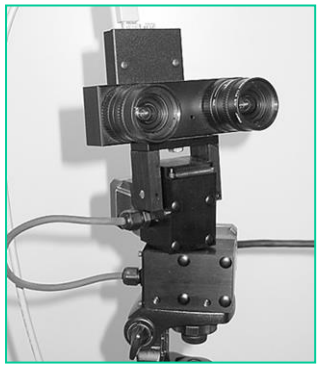


005)



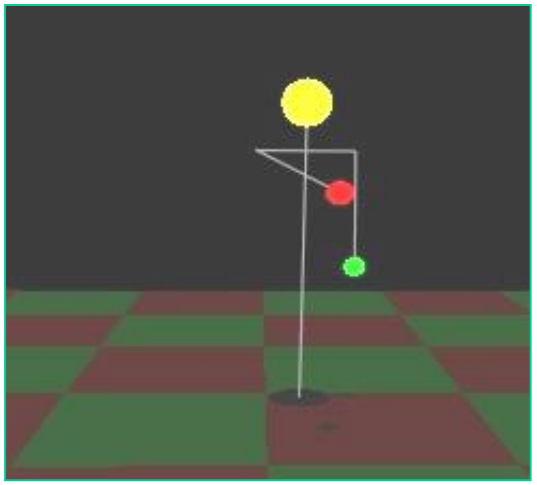
(2006)

Tracking of Head and Hands [Nickel04]



Stereo camera

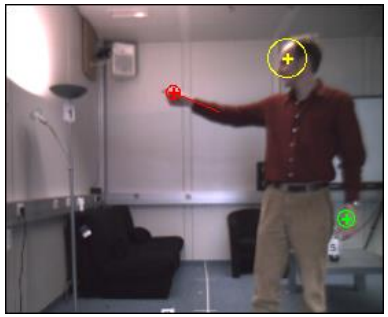
left/right image



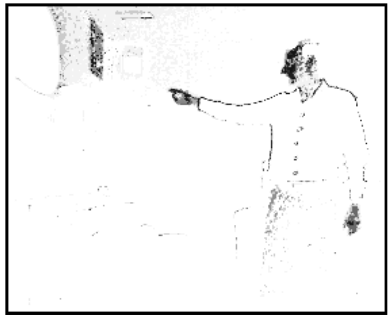
Extracted 3D model

Features: Color + Dense Disparity

Skin-pixel Clustering

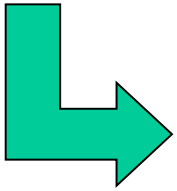


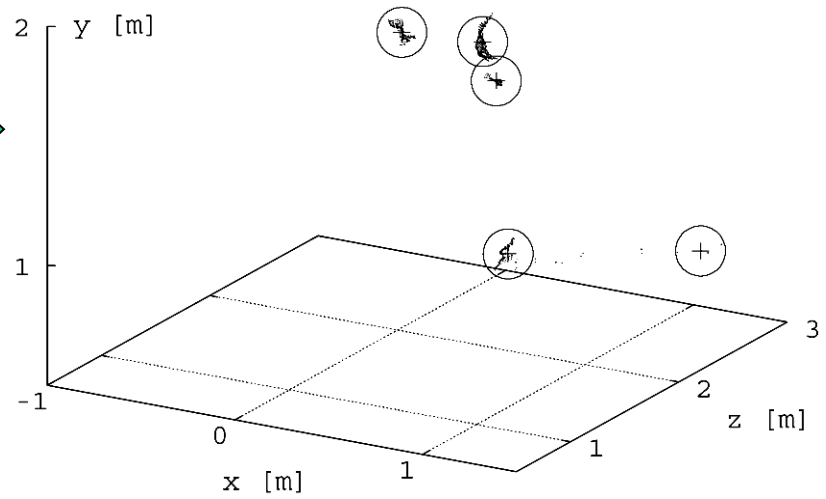
Camera image



Skin color map

Disparity map





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

Tracking Scheme

Given: the observation O_t (cloud of skin color pixels)

Wanted: the best hypothesis s^*_t (head + hand positions)

- Build a set of hypotheses $\{s_t\}$ using all permutations of the n strongest skin-color clusters as candidate positions for head and hands.
- Search

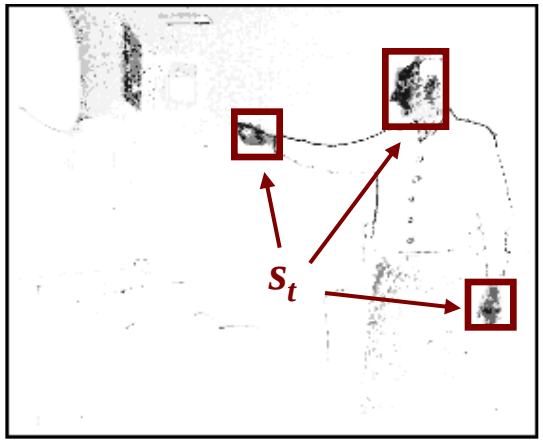
$$s_t^* = \arg\max_{s_t} P(O_t \mid s_t) \cdot P(s_t \mid s_{t-1}^*) \cdot P(s_t)$$

Observation model — Motion model — Posture prior

The Observation Model

$$\mathbf{P(O_t \mid s_t)}$$

- Represents the match between the state s_t and the current observation O_t
- Calculate the ratio of
  - the weights of the skin pixels that are in agreement with s_t
  - the weights of the skin pixels that are not covered by s_t

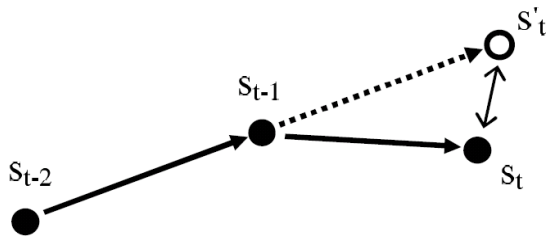


The Motion Model

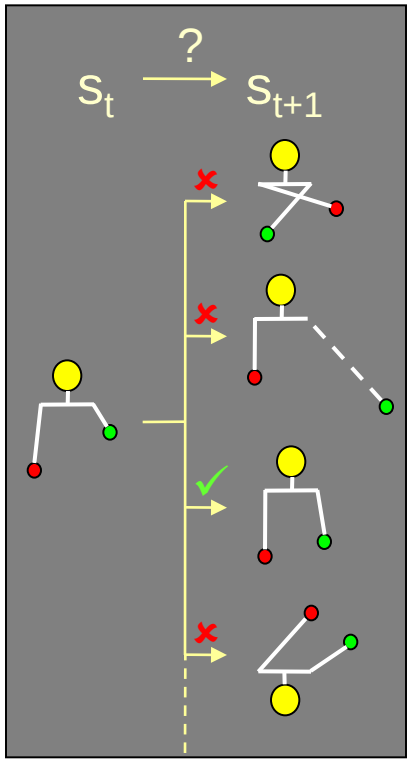
$$\mathbf{P(s_t \mid s_{t-1})}$$

Likelihood of a transition from s_{t-1} to s_t

- The less the body parts move from s_{t-1} to s_t, the higher $P(s_t|s_{t-1})$ will be.



- $P(s_t|s_{t-1}) = 0$, if the distance exceeds a given maximum (speed constraint).

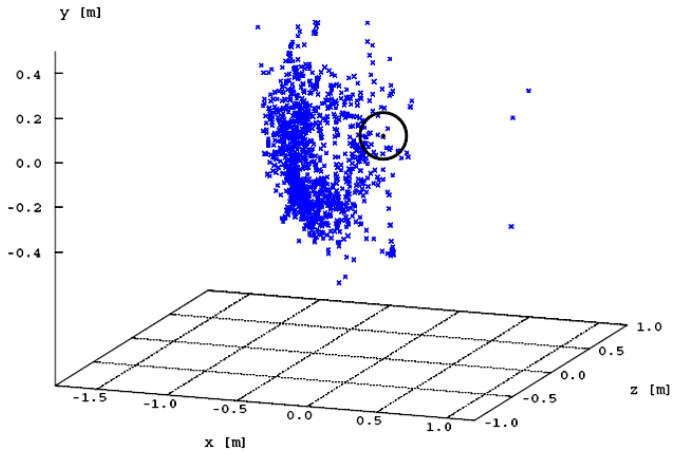


The Posture Prior

$P(s_t)$

Some postures are by nature more likely than others!

- Consider average person height
- Consider hand position relative to head → Gaussian mixture model trained on sample data
- $P(s_t) = 0$, if s_t breaks anatomic constraints

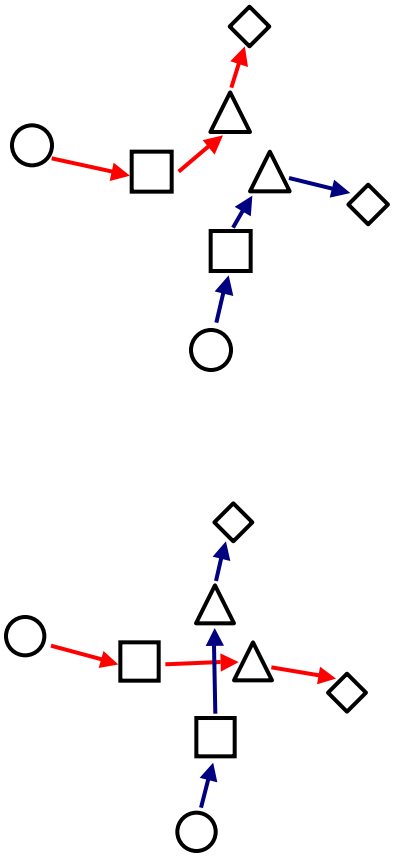


Observed right hand's positions over a time of 2min.

Multi-Hypothesis Tracking

- Often, there is more than one candidate for the current object location.
- A wrong assignment will cause errors in the future (can even lead to a dead end situation).
- From a future point of view, the problem may be not so hard anymore.

→ Defer the final decision. Establish more than one hypothesis at each time step and search for the *best sequence* back in time.



Pointing Gesture Phases

- Looking at persons performing pointing gestures, one can identify 3 phases in the movement of the pointing hand:

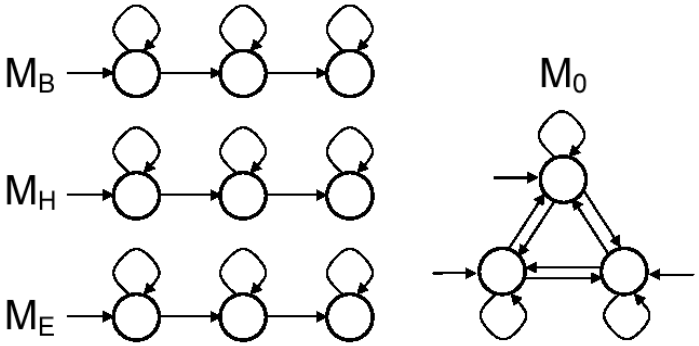
	μ [sec]	σ [sec]
Complete gesture	**1.75**	**0.48**
Begin	**0.52**	**0.17**
Hold	**0.76**	**0.40**
End	**0.47**	**0.12**

Average duration of pointing gesture phases

- For estimating the pointing direction, it is crucial to detect the hold phase precisely!

Modeling the Gesture

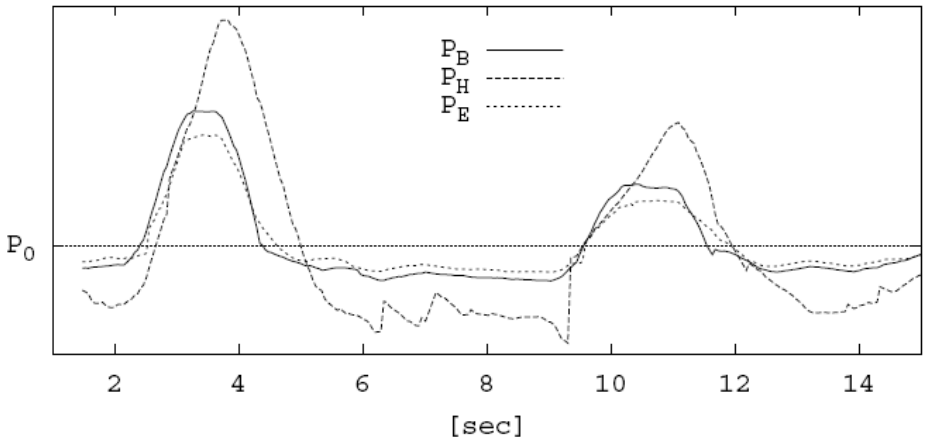
- Separate models for each of the 3 phases:



continuous HMMs with 2 Gaussians per state

- „Garbage model“ acts as a reference for the phase models‘ output
- Models trained on hand-labeled sequences using the Baum-Welch reestimation equations [Rabiner89].

Detection



Output of the phase models during a sequence of 2 pointing gestures (subtracted by P_0)

- the likelihood P_0 of the garbage model M_0 is subtracted from the gesture model likelihoods (P_1,P_2 ,P_3)
  - P_0 thus serves as a threshold

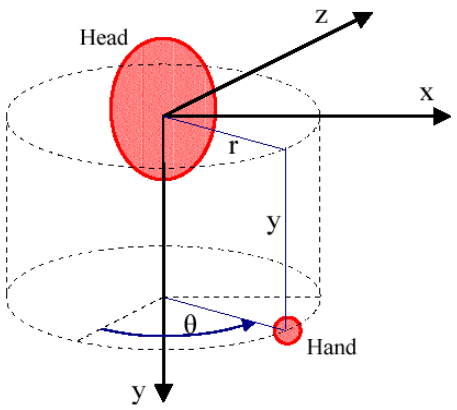
- → Detect a pointing gesture, whenever there are 3 points in time $t_B < t_H < t_E$, so that
  - $P_E(t_E) > P_B(t_E)$ and $P_E(t_E) > 0$
  - $P_B(t_B) > P_E(t_B)$ and $P_B(t_B) > 0$
  - $P_H(t_H) > 0$

Features

Hand position

(r, Δθ, Δy)

Coordinates of the hand in a cylindrical head-centered coordinate system



[see also Campbell, 96]

+

Head orientation

$(|\theta_{Head}-\theta_{Hand}|, |\Phi_{Head}-\Phi_{Hand}|)$

- Difference between head's and hand's azimuth/elevation angle

  → 0, if head and hand are "in-line"

- Measured with a magnetic sensor

- Visually Estimated (ANN)

Pointing Direction

Head-hand line

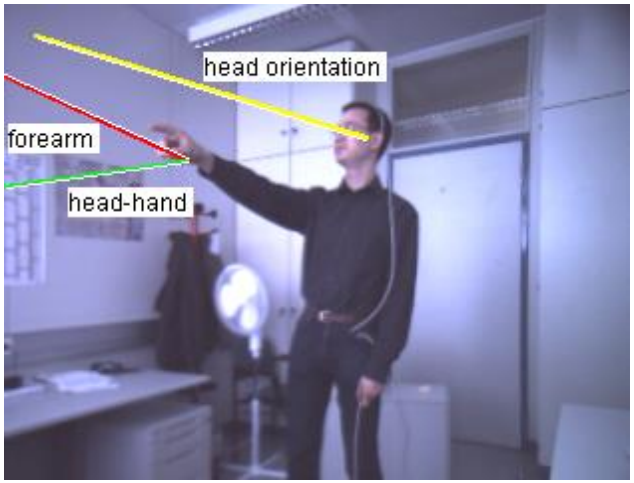
- Line of sight between head and hand → provided by the tracker

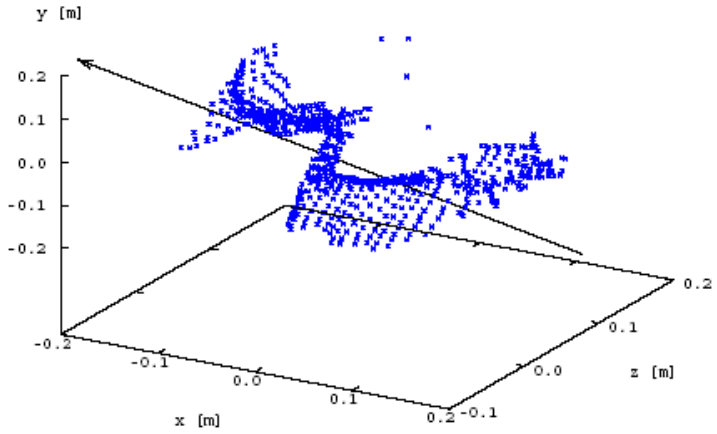
Forearm orientation

- PCA of the point cloud around the hand → forearm orientation

Head orientation

- Visually estimated (ANNs)
- (Magnetic sensor)



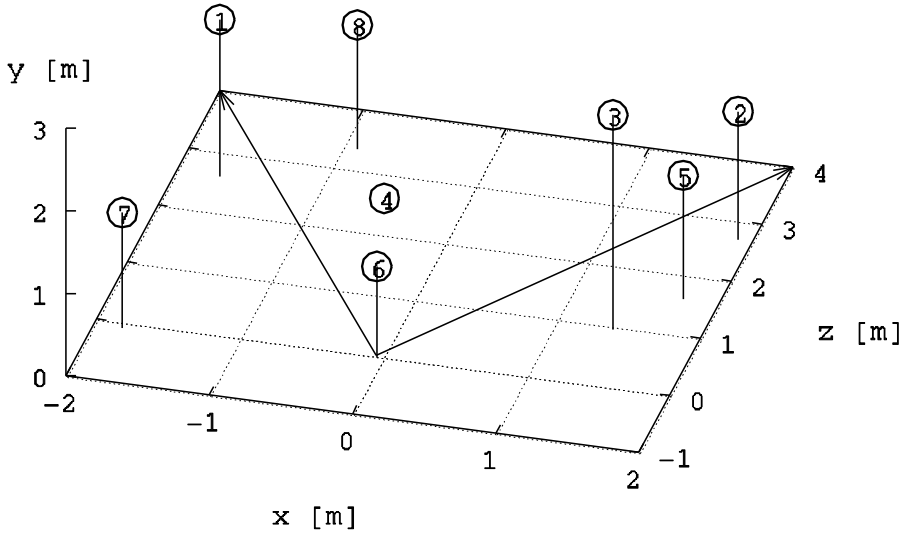


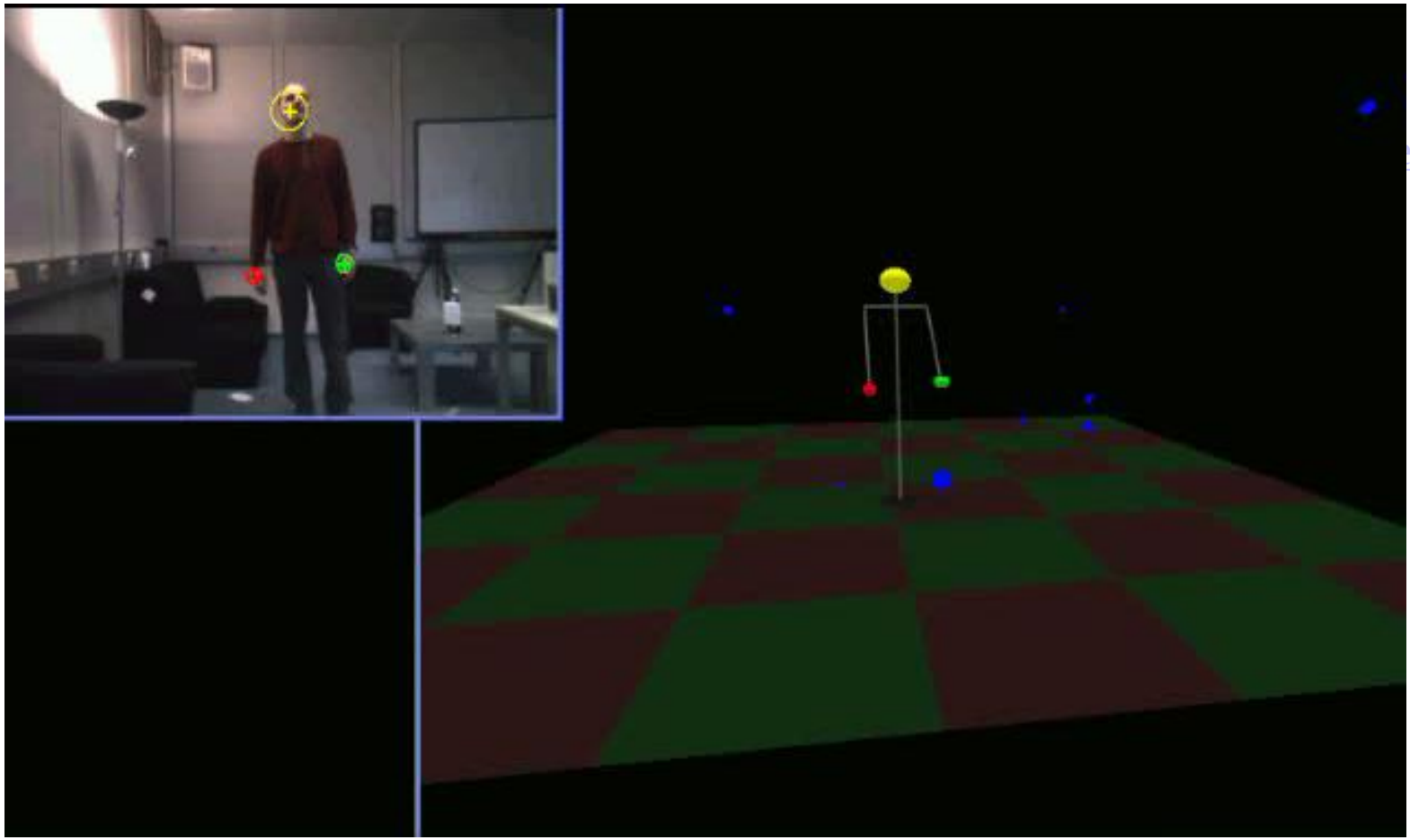
Experiments and Results

Scenario:

Household Robot

- 118 gestures
- 8 targets

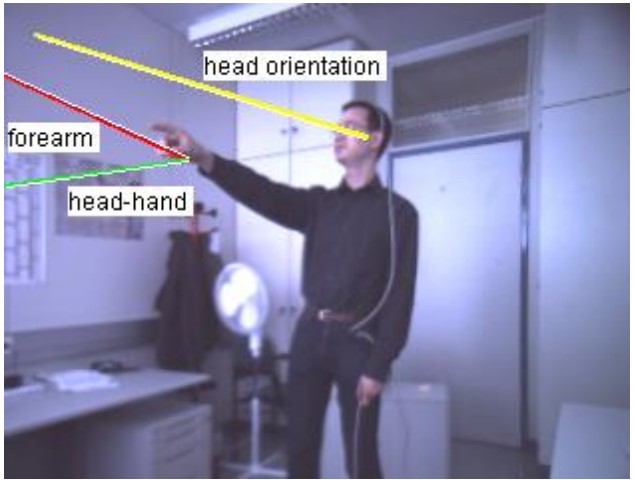




- (Nickel & Stiefelhagen, 2004)

Results Pointing Direction

Accuracy of pointing direction estimation on manually labeled hold-phases



	Error angle	Targets identified	Availability
Head-hand line	**25°**	**90%**	**98%**
Forearm line	**39°**	**73%**	**78%**
Head orientation	**22°**	**75%**	**(100%)**

Results in Gesture Recognition

Performance of the automatic gesture recognition and pointing direction estimation:

	Recall	Precision	Error angle*
No Head Orientation	**79.8%**	**73.6%**	**19.4°**
Visually *estimated* Head-Orientation	**78.3%**	**87.1%**	**16.9°**
True (Sensor) Head Orientation	**78.3%**	**86.3%**	**16.8°**

*Head-hand line

Multimodal Human-Robot Interaction

Multimodal Interaction in a Household Scenario

- Person Tracking & Gesture Recognition
- Speech Recognition
- Multimodal Dialog Manager
  - Initiates clarification dialog for missing parameters
- Speech Synthesis
- Integrated on Mobile Robot
  - able to find & follow person

SFB 588 Humanoide Roboter - Lernende und kooperierende multimodale Roboter

- **Weitere Anwendung: Zeigegestenerkennung in einem Smart Room**

Ziele

- Projekt: “Visuelle Perzeption für die MMI – Interaktion in und mit aufmerksamen Räumen”
  - Projekt am Fraunhofer IOSB (ehemals IITB)
  - Beginn: Oktober 2007, Laufzeit 5 Jahre, 5 Mitarbeiter

- Projektziel:
  - Maschinensehen für die Unterstützung der Interaktion des Menschen in und mit »aufmerksamen« Räumen (»Smart Rooms«) und großflächiger Display-Umgebung
  - Erstes Anwendungsziel: »Smart Control Room« für Krisenreaktion und –Management

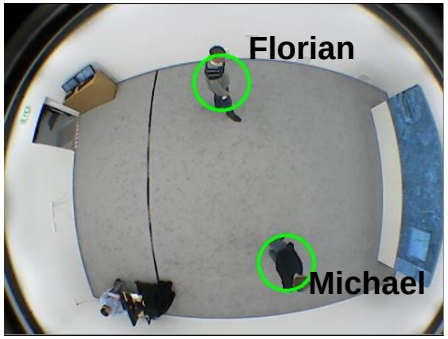
- Zielfunktionalitäten des Raumes:
  - Erfassung aller Personen: Position, Identität, Blickrichtung und Aufmerksamkeit, Bewegungen
  - Erkennung, welche Information wahrgenommen wurde
  - Unterstützung personalisierter Arbeitsumgebungen
  - Unterstützung multimodaler Interaktion mit großen Displays

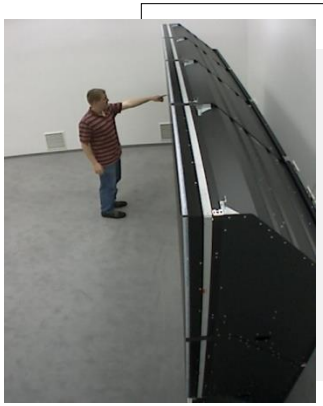
Der Smart Room: Sensoren & Geräte

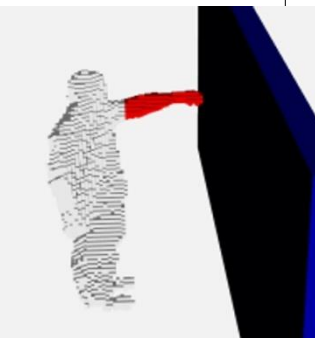
- Videowand:
  - 8 Rückprojektionswürfel (4096x1536 px)
  - Können als ein Display angesteuert werden
- Sensoren
  - 4 Kameras in den Raumecken,
  - 1 Weitwinkelkamera an der Decke
  - 2 (4) Kameras über der Videowand
  - 2 aktive (pan-tilt-zoom) Kameras
  - Mikrophone
  - Lautsprecher
- Weiterer Smart Room am KIT
  - Ähnliches Setup, keine Videowand
  - Mikrophonarrays zur Lokalisierung

Smart Room: Perzeption & Funktionalität

- Perzeptionskomponenten (alle echtzeitf.)
  - 3D Personentracking und Berechnung der Visuellen Hülle
  - Erkennung von Zeigegesten und Berührungen
  - Erkennung v. Kopfdrehungen und Aufmerksamkeit
  - Identifikation durch Gesichtserkennung
  - Spracherkennung (Kooperation mit KIT)
- Situationserfassung
- Erste Funktionalitäten
  - Personalisierte und lokalisierte Arbeitsplätze und Nachrichten
  - Multimodale Interaktion mit der Videowand (Sprache, Gesten, Berührung)
  - Erkennung der Aufmerksamkeit auf Videowand





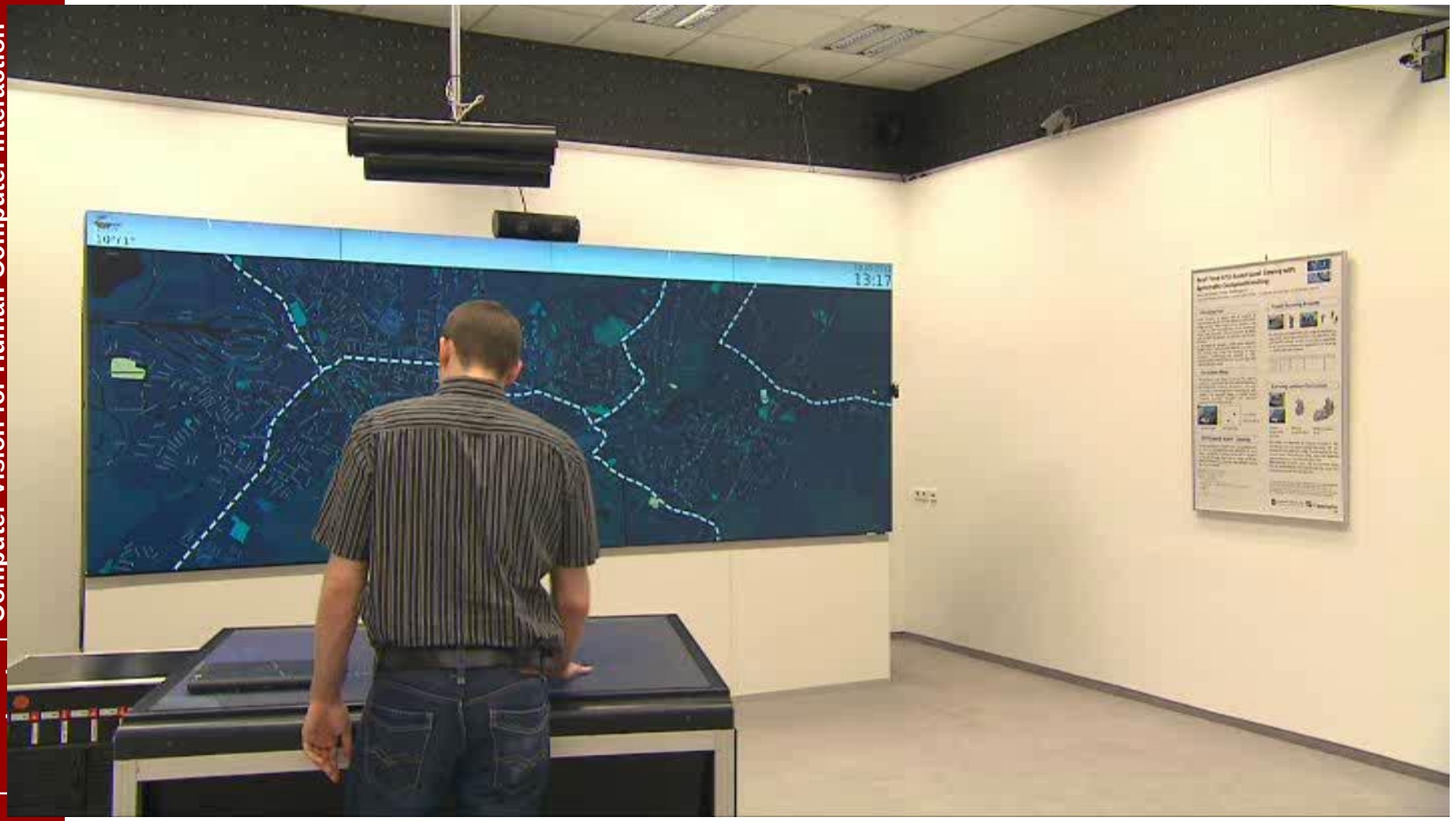


Video:

Interaktive Räume

Fraunhofer IOSB

Computer Vision for Human-Computer Interaction



Perzeptionskomponenten

3D Mehrpersonentracking

→ Lokalisierung ist eine grundlegende Komponente für alle weiteren Analysen

- Merkmale:
  - Kopf- und Oberkörperdetektoren, Vordergrund, Farbe
  - Audio: Time-delay of arrival (TDOA), Gen. Cross-Corr.
- Partikelfilter für Tracking und Fusion

Multi-person tracking in a smart room (4-5 cameras)

Identifikation

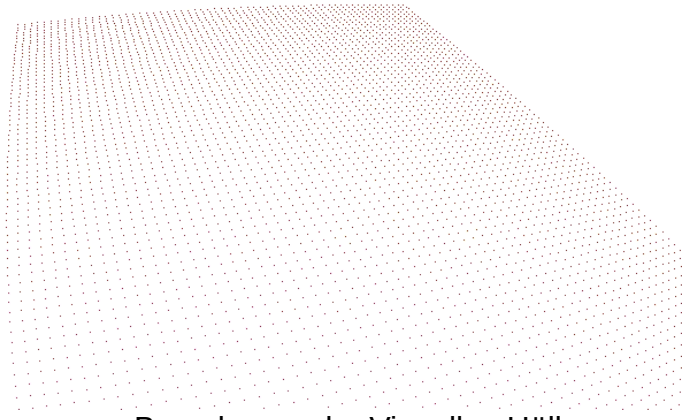
→ Notwendig für Personalisierung

- Lokaler ansichtsbasierter Ansatz auf Basis von DCT
  (Forschungspreis 2008 des Europ. Biometric Forum)
- Kann mit Sprechererkennung und aktiven Kameras kombiniert werden
- Gemeinsames Tracking und Identifikation bringt Vorteile

Combined tracking and identification

Perzeptionskomponenten (2)

- Extraktion der Visuellen Hülle:
  - → Liefert 3D Repräsentationen der Personen
  - → Hilfreich zur Erkennung der Körperhaltung und für Zeigegesten
  - In Echtzeit möglich, mit Voxel-Carving Ansatz



Berechnung der Visuellen Hüllen

- Erkennung von Kopfdrehungen
  - → Liefert wichtige Information über die Aufmerksamkeit
  - Gelöst durch videobasierte Schätzung der Kopfdrehungen (Neuronale Netze)
  - Fusion über alle Ergebnisse aus den Kamerasichten (Partikelfilter)
  - Abbildung der Kopfdrehungen auf wahrscheinliche Ziele, bspw. andere Personen, bestimmte Objekte, Regionen der Videowand

Erkennung von Kopfdrehungen

Interaktion mit großen Videowänden

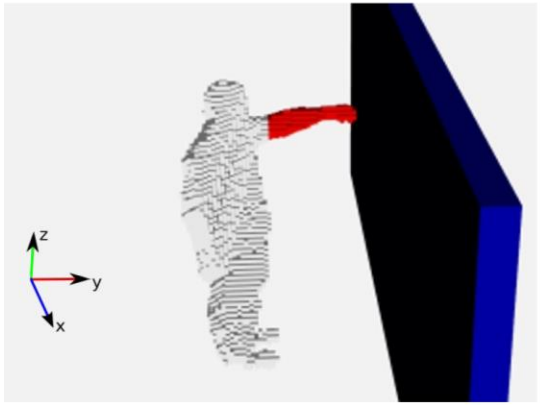
- Benutzer erwarten Berührungserkennung!
- Berührungen reichen aber nicht aus
  - Objekte sind zum Teil nicht erreichbar (Entfernung / Höhe)

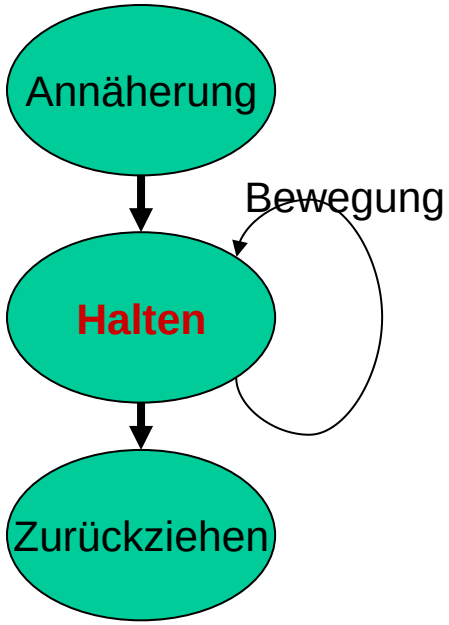
→ Erweiterung: Nutzung von Zeigegesten

- Vorteile der videobasierten Lösung
  - Vorhandene Geräte und Oberflächen können erweitert werden
  - Skaliert gut in der Größe
  - Kameras sind günstig
  - Nahtloser Übergang zwischen Berührung und Zeigegesten ist möglich

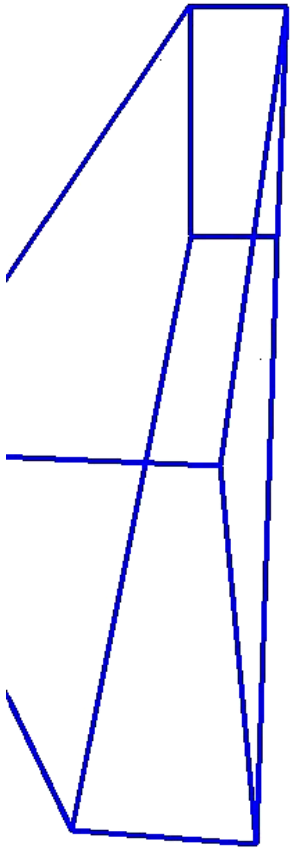
Technische Realisierung

- 3D Extraktion der Visuellen Hülle
  - mit Voxel-Carving Ansatz
- Detektion und Tracking des Arms
  - Detektion eines zylinderförmigen Objekts vor der Videowand
  - Tracking durch paarweises Matching
  - Interaktionspunkt wird durch 3D Rekonstruktion berechnet
- Erkennung der Interaktion (Berührung / Geste)
  - Als binäre Interaktionsmodalität modelliert
  - Zustand wird durch Handbewegung determiniert
  - Nahtloser Übergang zwischen Berührung und Zeigen





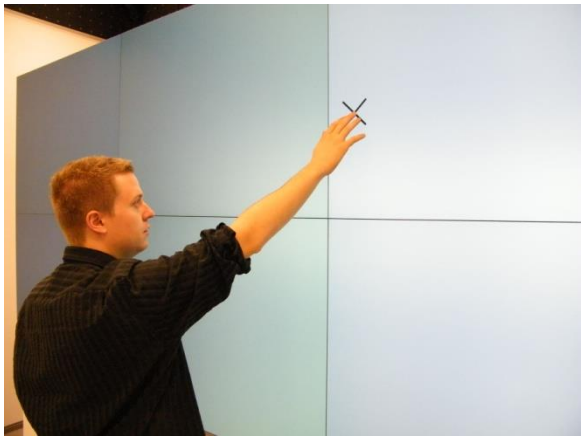
3D Armrekonstruktion und Tracking



Evaluation der Komponenten

- Laufzeit: < 20ms / frame
  - 3 GHz Core 2 Duo, NVIDIA GTX280

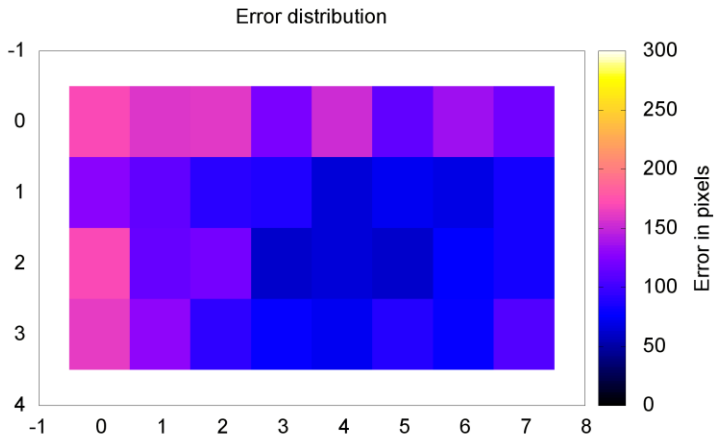
Komponente	Dauer [ms]
Vordergrundsegmentierung	6.50
Voxel carving	8.26
Berührung und Zeigen	1.89



- Genauigkeit
  - Wie genau trifft die Berührungs- / Zeigedetektion den vom Benutzer anvisierten Punkt

Modalität	Mittl. Fehler [cm]	Standardabweichung [cm]
Berührung	9.49	6.56
Zeigen	16.61	9.16

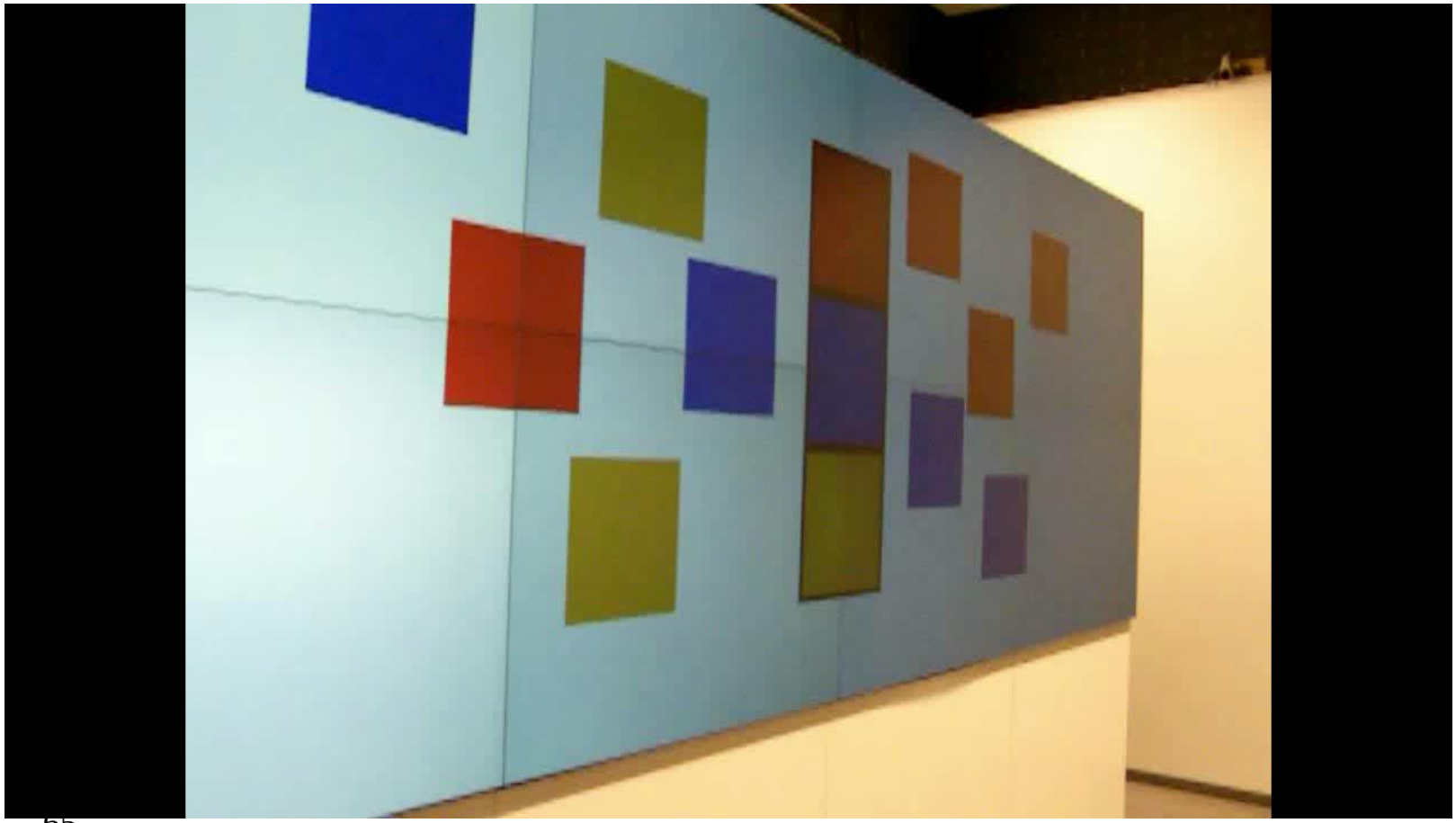
(Stand September 09, inzwischen deutlich genauer)



Verteilung der Fehler auf der Videowand
(Berührung / Zeigen kombiniert)

Benutzerstudie

- Aufgabe: Bewege die farbigen Blöcke schnellstmöglich ins Ziel
- Zwei Konditionen: “Touch-only” vs. “Touch-and-pointing”
- Teilnehmer: 15 Männer, 4 Frauen



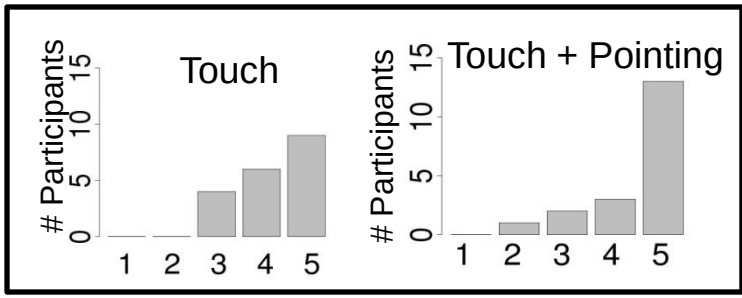
Benutzerstudie: Ergebnisse

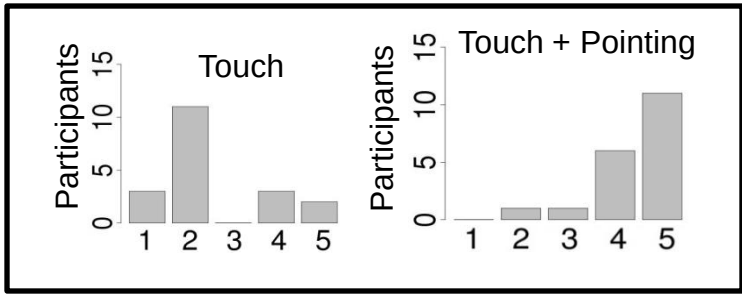
- Fragebögen mit Fünfpunkt Likert-Skala
- Wilcoxon Rank Sum Test zur Analyse statistischer Signifikanz

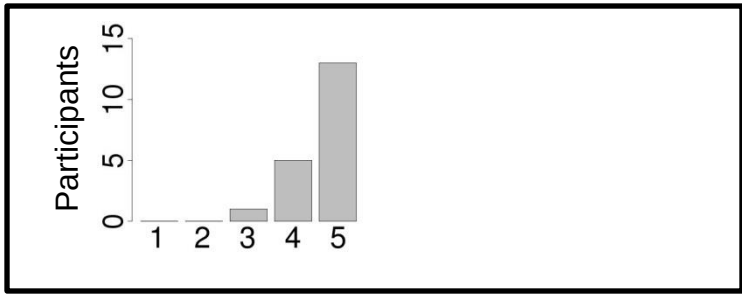
- Hypo 1:“touch + pointing” ist intuitiver und einfacher zu benutzen
  → Nein, beides gleich gut bewertet (p=0.14)

- Hypo 2: Alle Regionen an der Videowand sind per “touch + pointing” einfacher zu erreichen
  → Ja ! (p<<0.01)

- Hypo 3: “„Zeigegesten waren eine hilfreiche Ergänzung“
  → Ja ! (μ>4 with p<<0.01)

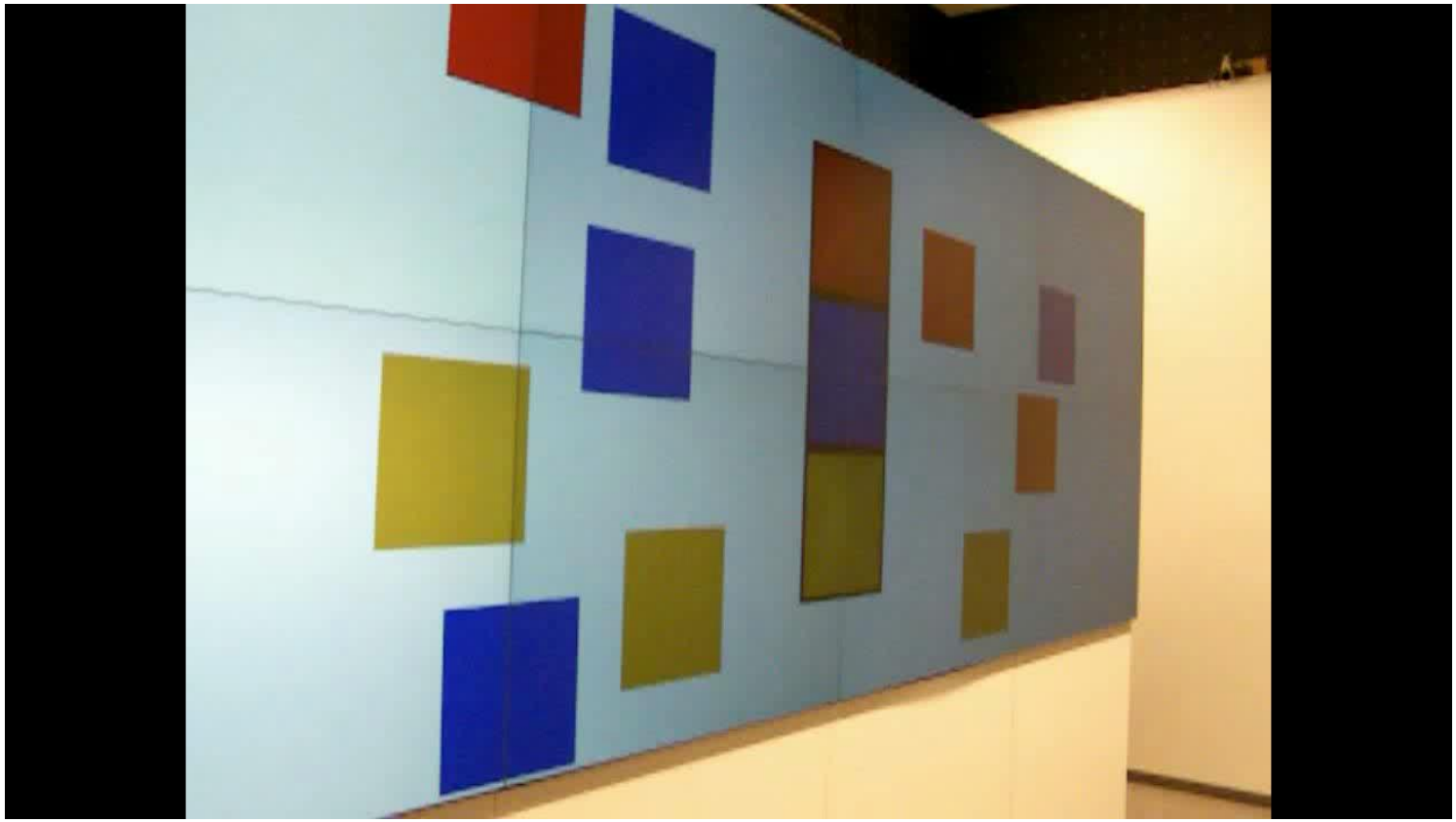






Fortschritte: Multi-Touch

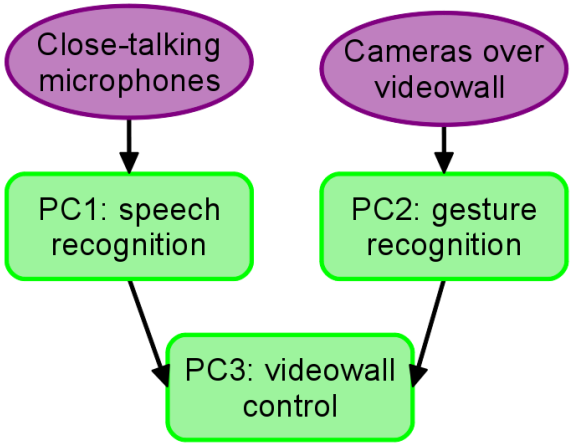
- Multitouch



Multimodale Interaktion

Anwendung:

- Lageraum für Feuerwehr
- Hinzufügen taktischer Symbole zur Lagekarte



Methode:

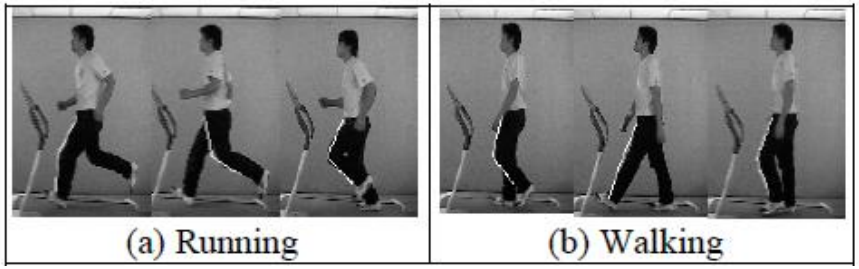
- Sprachkommando zur Auswahl des Symbols
- Berührung/Zeigen zur Positionierung

Geplant:

- Mehr und komplexere Sprachkommandos
- Bessere Speech-Detektion
- Manipulation von Symbolen
- Multimodaler Dialog

Related Topics

- Human motion Analysis
  - Gait: classification and person identification
  - Classification of motion figures in dance, sport, …
- Facial expression
- Synthesizing gesture and facial expressions
  - Talking heads
  - Avatars



[Yam02]

Summary

- Gesture Taxonomy:
  - Manipulative - semaphoric - conversational
  - Static vs. dynamic
- Hidden Markov Models
  - "The 3 problems"
  - Gaussian mixture models
- Examples
  - American Sign Language recognition
  - Pointing gestures recognition
  - Interaction with a video wall, Speech & Gesture

References

*** [Pavlović97]**
V.I. Pavlović, R. Sharma, T.S. Huang: Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997.

*** [Rabiner89]**
Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE, 77 (2), 257–286, 1989.

*** [Starner98]**
T. Starner, J. Weaver, A. Pentland: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12):1371--1375, 1998.

*** [Nickel04]**
Nickel, K., Stiefelhagen, R.: 3D-Tracking of Heads and Hands for Pointing Gesture Recognition in a Human-Robot Interaction Scenario, Sixth Int. Conf. On Face and Gesture Recognition, May 2004, Seoul, Korea.

Additional:

[Becker97]
Becker, D.A.: Sensei: A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures. M.I.T. Media Lab Perceptual Computing Group Technical Report No. 426, 1997.

[Cassell98]
Cassell, J.: A Framework For Gesture Generation And Interpretation. In Cipolla, R. and Pentland, A. (eds.), Computer Vision in Human-Machine Interaction, pp. 191-215. New York: Cambridge University Press. 1998.

[Poddar98]
Poddar, I., Sethi, Y., Ozyildiz, E. and Sharma, R.: Toward Natural Gesture/Speech HCI: A Case Study of Weather Narration. Proc. Workshops on Perceptual User Interfaces, pages 1-6, November, 1998.

[Xiong03]
Y. Xioung, F. Quek, D. McNeill: Hand Motion Gestural Oscillations and Multimodal Discourse. ICMI 2003.