Estimation of Head Pose and Focus of Attention

Prof. Dr. Rainer Stiefelhagen

Rainer.stiefelhagen@kit.edu

10.1.2013

Tracking a User's Focus of Attention

- **Focus of Attention tracking**:
  - To detect a person's interest
  - To know what a user is interacting with
  - To understand his actions/intentions
  - To know whether a user is aware of something

- **Human-Human Interaction:**
  - to determine the addressee of a speech act
  - to understand the dynamics of interaction
  - for meeting indexing / retrieval
- **Human-Robot Interaction**
  - Was the robot addressed or not?
- Smart Environments, Cars, …

Attention during Social Interaction

- **“Looking Means Listening!”** (Ruusuvuori 2000)
- **Attentional signals**:
  - Eye gaze
  - Head orientation
  - Body posture
  - Gestures

- **Head Orientation** is a good cue to predict focus of attention !
- **Eye-gaze** is difficult to track …

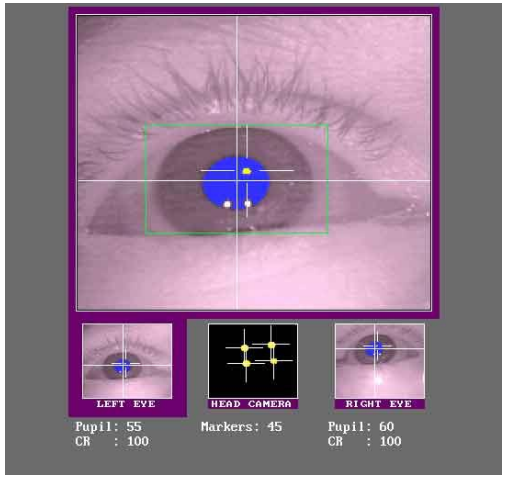
Typical Eye-Gaze Trackers

- Head mounted systems
  - user is allowed to move (limited by wires, etc.)
  - accurate head pose and eye-gaze (< 1 degree error)
  - fast (> 60 Hz possible)

➔ Very intrusive !

Typical Eye-Gaze Trackers (2)

- Remote Trackers:
  - user‘s movement is very limited
  - problems with head rotation



Vision-based Eye Gaze Tracking – Problems

- Requires the user to wear head gear
- or position of the user relative to the camera is rather fixed

- Certainly not acceptabel for everyday use in multimodal rooms

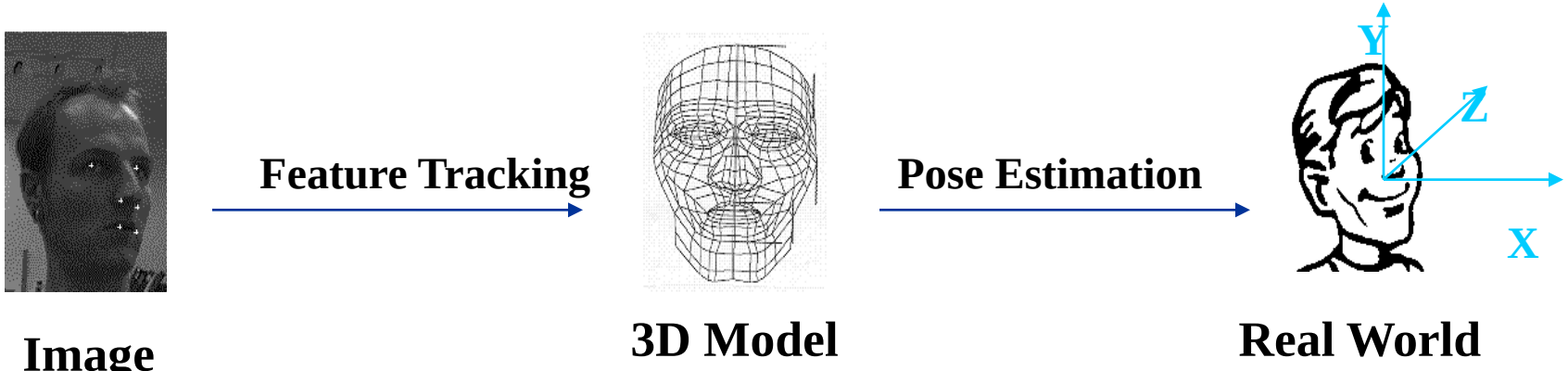
Head Pose Estimation

- Model-based approaches:
  - Locate and track a number of facial features
  - Compute head pose from 2D to 3D correspondences (Gee & Cipolla '94, Stiefelhagen et.al '96, Jebara & Pentland '97,Toyama '98)

- Appearance-based approaches:
  - estimate new pose with function approximator (such as ANN) (Beymer et.al.'94, Schiele & Waibel '95, Rae & Ritter '98)
  - use face database to encode images (Pentland et.al. '94)

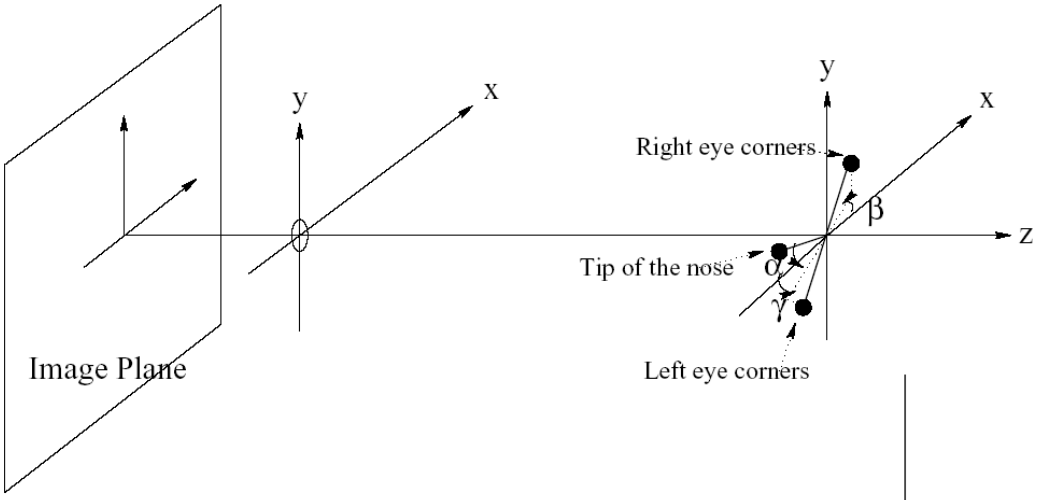
Model-based Head Pose estimation

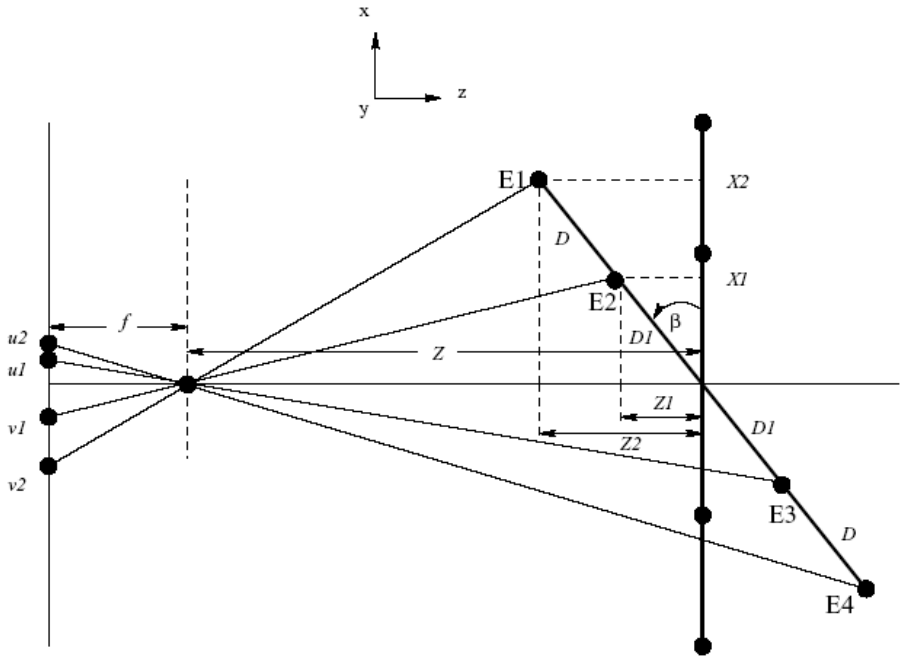
Find correspondences between points in a 3D model and points in the image



Model-based Head Pose estimation (2)

(Iteratively) solve linear equation system to find pose parameters (r_x, r_y, r_z, t_x, t_y, t_z)





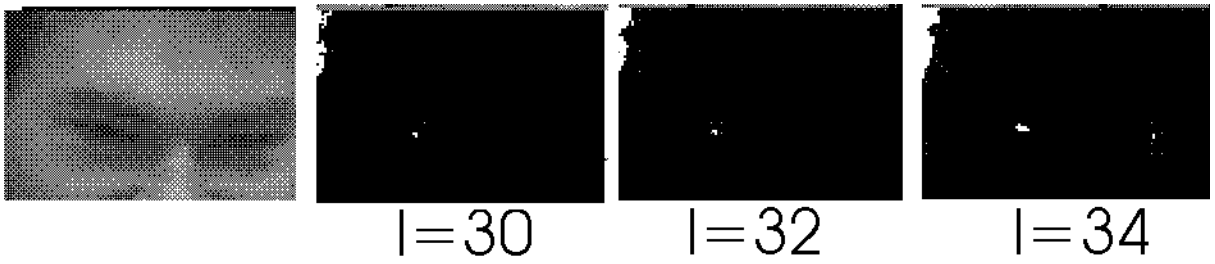
Model-based Head Pose estimation - Real-Time Facial Feature Tracking

Top Down Approach:

- searching the face by the face tracker
- searching and tracking of 6 facial feature points: pupils, nostrils and lip corners

Eye detection using iterative thresholding

- Idea: they eyes (i.e. the pupils) are darker than the surounding area
- Approach:
  - search for 2 dark regions in certain search area
  - Use anthropometric constraints for search
  - iterative approach to adapt to different lighting conditions

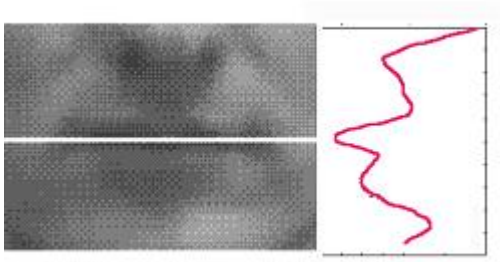


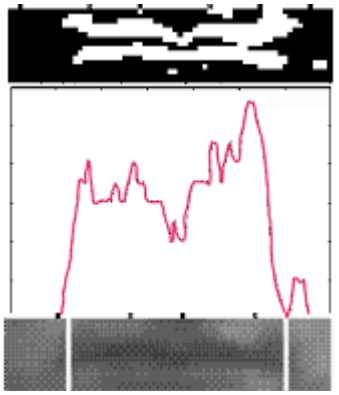
Iterative thresholding on grayscale image,
Until two valid eye candidates are found

- Same approach applied to find nostrils
  - Search area below eyes, above lips

Lip detection using integral projections

- Idea:
  - Line between lips is darker than surrounding
  - There are horizontal edges around the lips
- Approach
  - predict search window using eye position and face-model
  - find vertical mouth position (line between the lips) using horizontal integral projection
  - find horizontal boundaries using edge detection and vertical integral projection



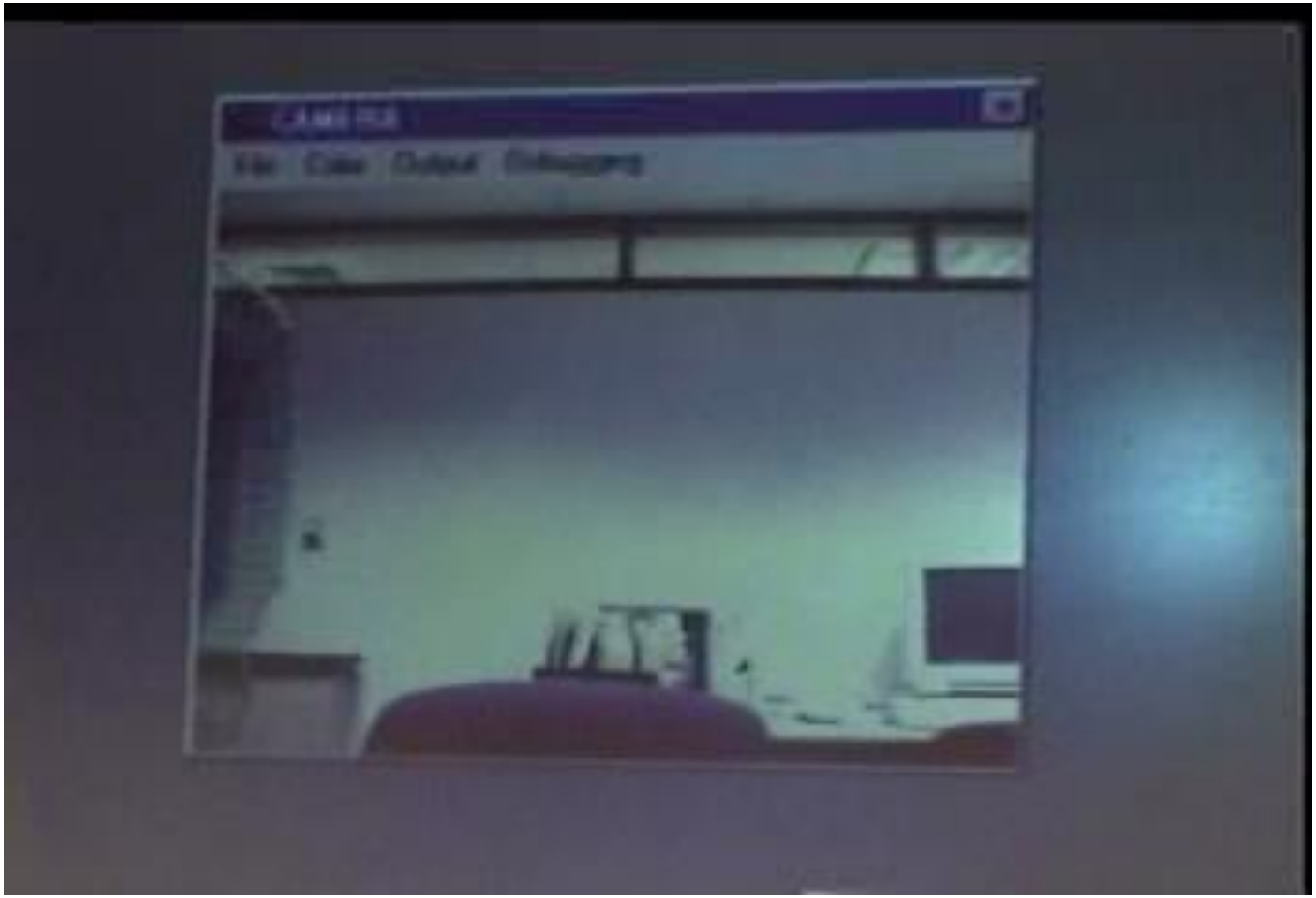


Model-based Head Pose estimation - Real-Time Facial Feature Tracking (2)

- Pose estimation accuracy depends on correct feature localization!

- Problems:
  - Choice of good features
  - Occlusion due to strong head rotation
  - Fast head movement
  - Detection of tracking failure / re-initialization
  - Requires good image resolution


Gaze Tracker



Video

Head Pose Tracking in *Meetings*

Why:

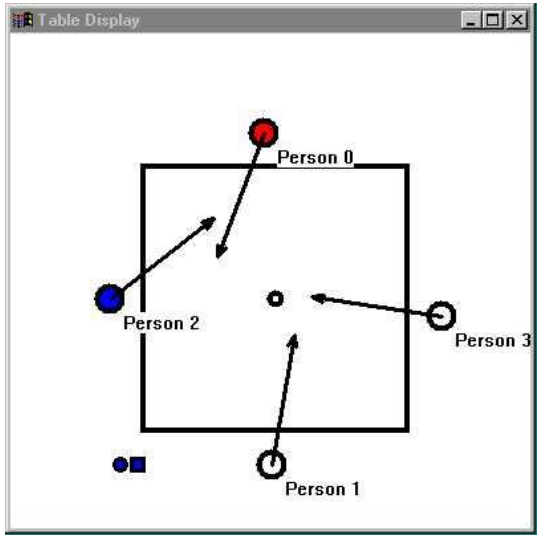
- to determine the addressee of a speech act
- to track the participants focus of attention
- to analyse, who was in the center of focus
- for meeting indexing / retrieval), …

Problems with model-based approach:

- Requires good image resolution
- Would require one camera per person
- Robustness
  - Fast movement, strong head rotation, etc.

Focus of Attention Tracking in Meetings (2)

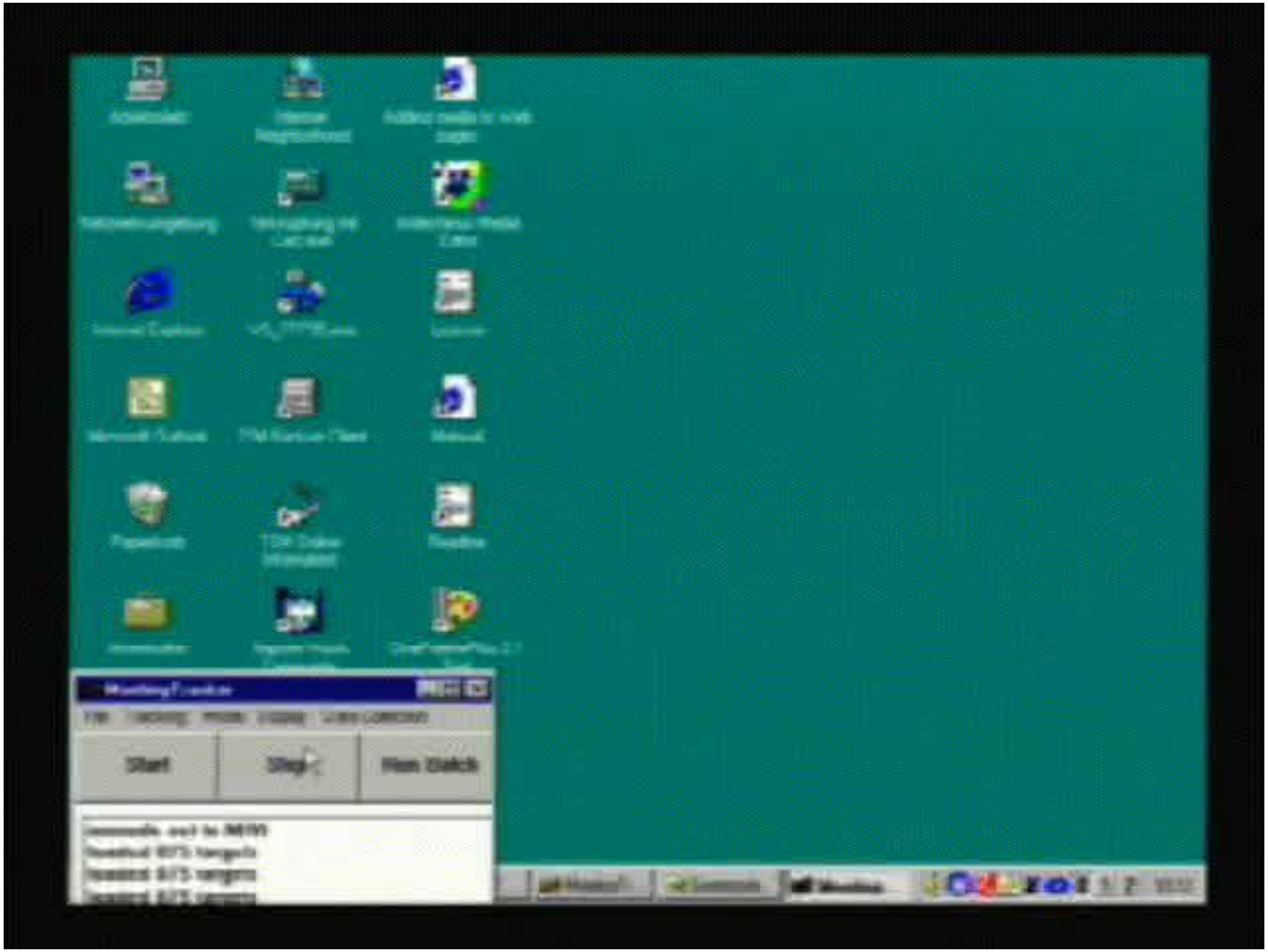
(1)



(2/3)

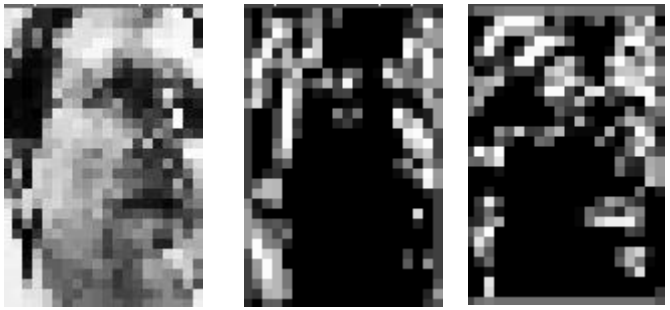
How:

1. Track all participants‘ faces **(color-based) √**
2. Estimate their head orientations (ANNs)
3. Map head orientations onto likely targets (e.g. the other participants)

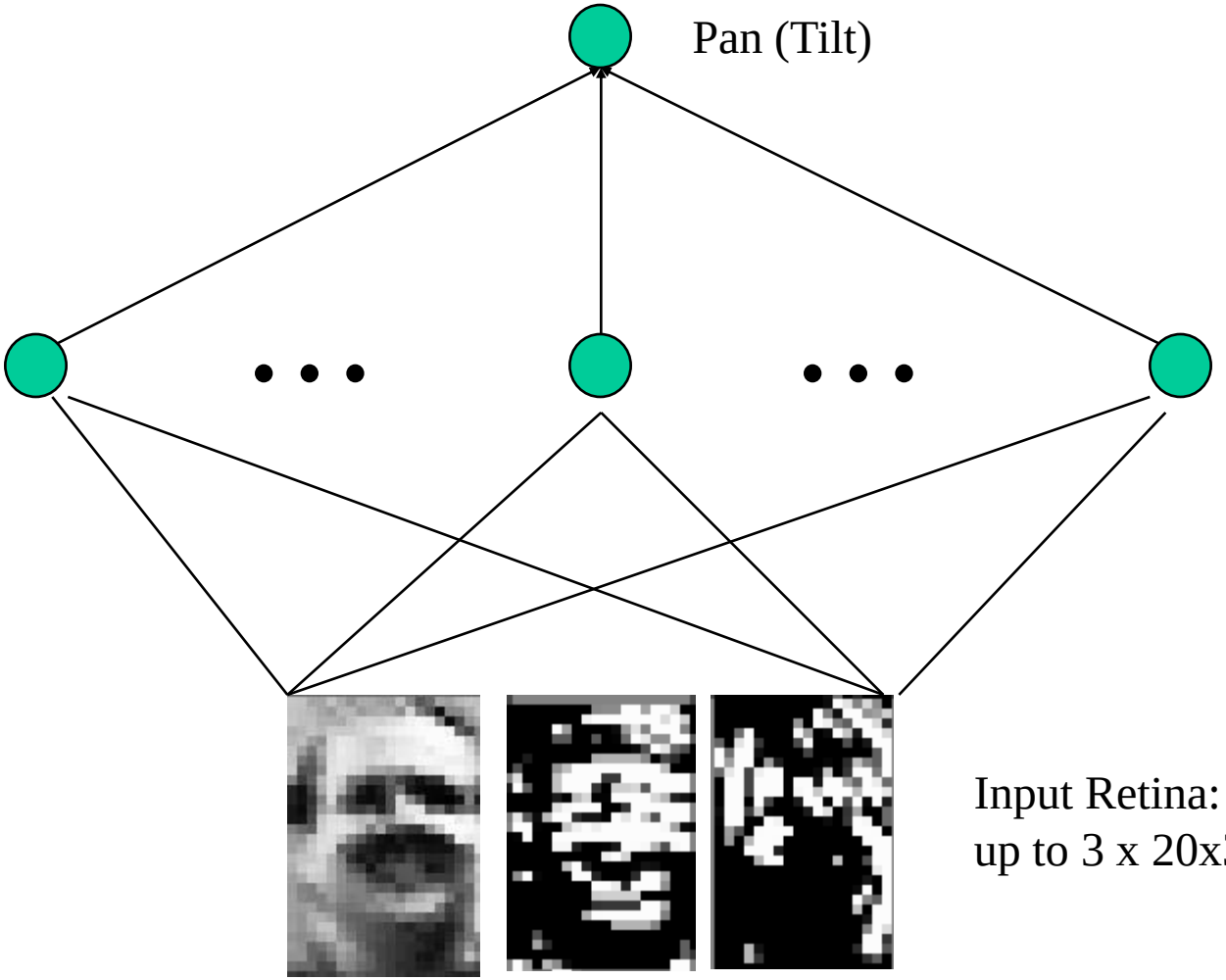

Start

Estimating Head Pose with ANNs

- Train neural network to estimate head orientation
- Preprocessed image of the face used as input



Estimating Head Pose with ANNs – Network Architecture

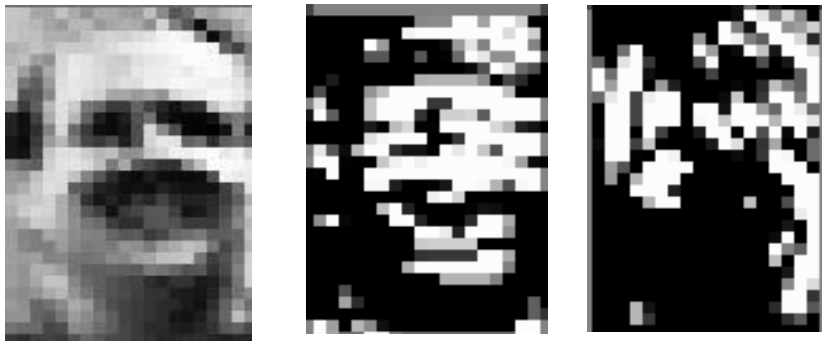


Estimating Head Pose with ANNs – Image Preprocessing

- Automatic extraction of faces
  - using a skin-color model
- Preprocessing

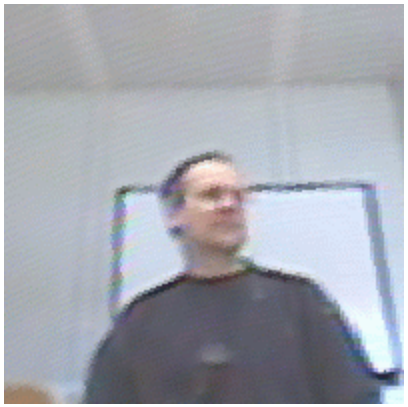
  a) Histogram normalization of grayscale images

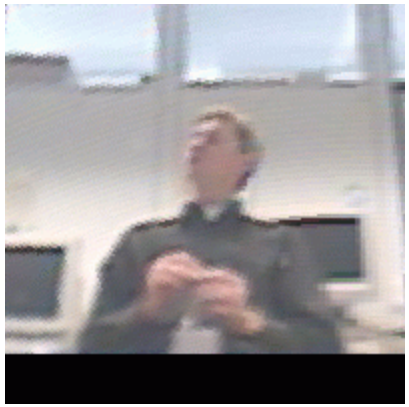
  b) Extracting horizontal- and vertical edges
  - Down-sampling to 20x30 pixel

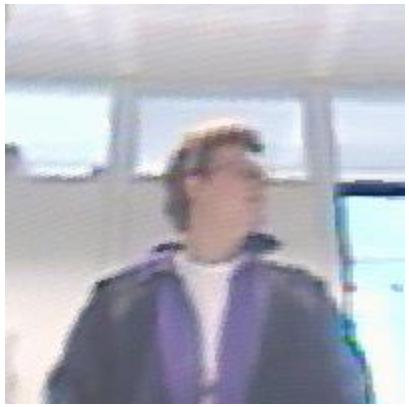


Estimating Head Pose with ANNs – Data Collection

- Estimate head pan and tilt from the facial image
- Data Collection:
  - perspective views of users are collected with pano-cam
  - labels from a magnetic field pose tracker
  - data from fourteen users, at four positions around table collected







Estimating Head Pose with ANNs – Training / Results

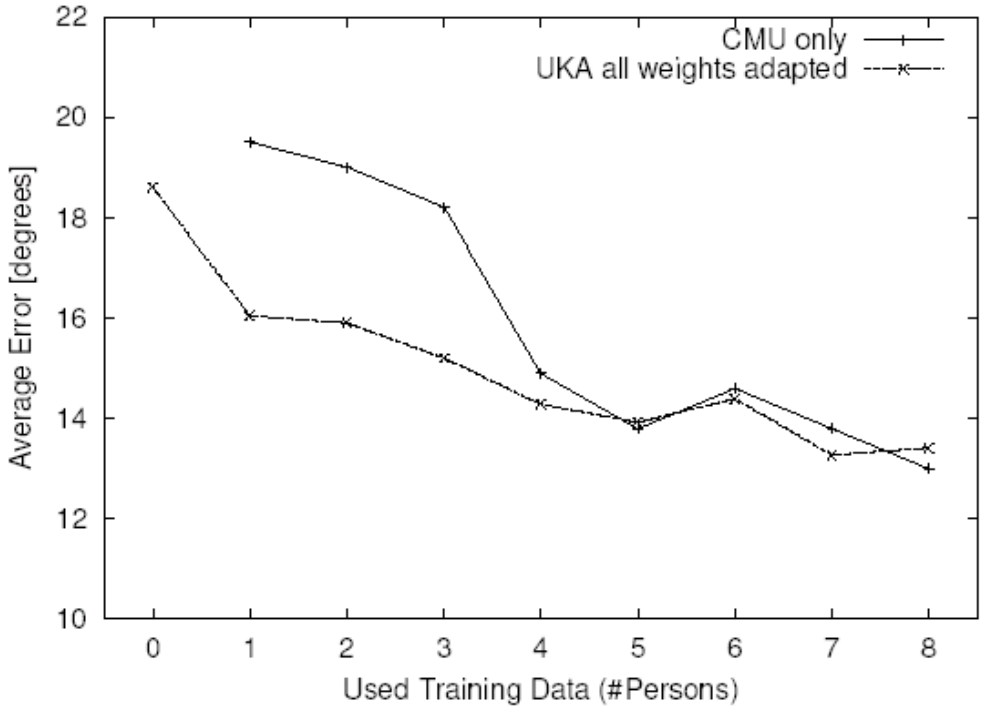
- Training on 6100 images from 12 users (+mirrored images)
- Crossevaluation on 750 images from same 12 users
- Tested on 750 images from these users
- Additional User Independent Testset: 1500 images from two new users

	training set	test set	new users
histo	4.8 / 3.6	5.5 / 4.1	10.4 / **9.3**
edges	4.3 / 2.9	5.6 / 3.5	12.2 / 10.3
both	2.1 / 2.1	**3.1 / 2.5**	**9.5** / 9.8

Average Error in degrees for pan / tilt

Estimating Head Pose with ANNs – Problems

- Illumination Changes are problematic
- Retraining / Adaptation is necessary

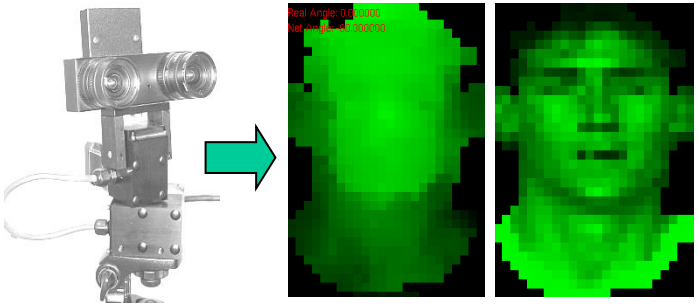


Head Pose Estimation Using Depth from Stereo

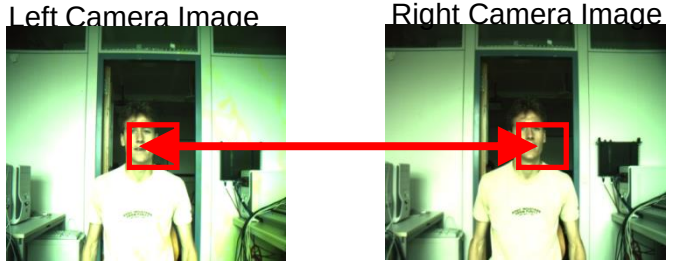
- Using stereo cameras, the distance of a point on the object from the cameras can be computed
  - Distance *r* of a pixel is inversely related to the difference in projections of the point in the two camera views („*disparity*“)

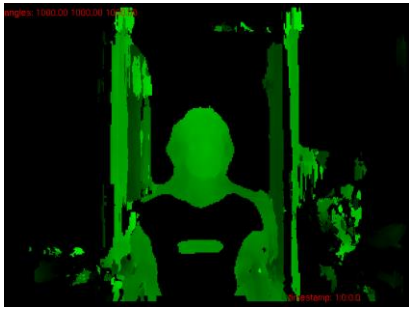
$$r = (b \cdot f) / (d_L - d_R)$$

  - Disparities are computed for *each* pixel by finding correspondences in the two views
  - (→ See Tracking II / Multi-Camera Topologies)

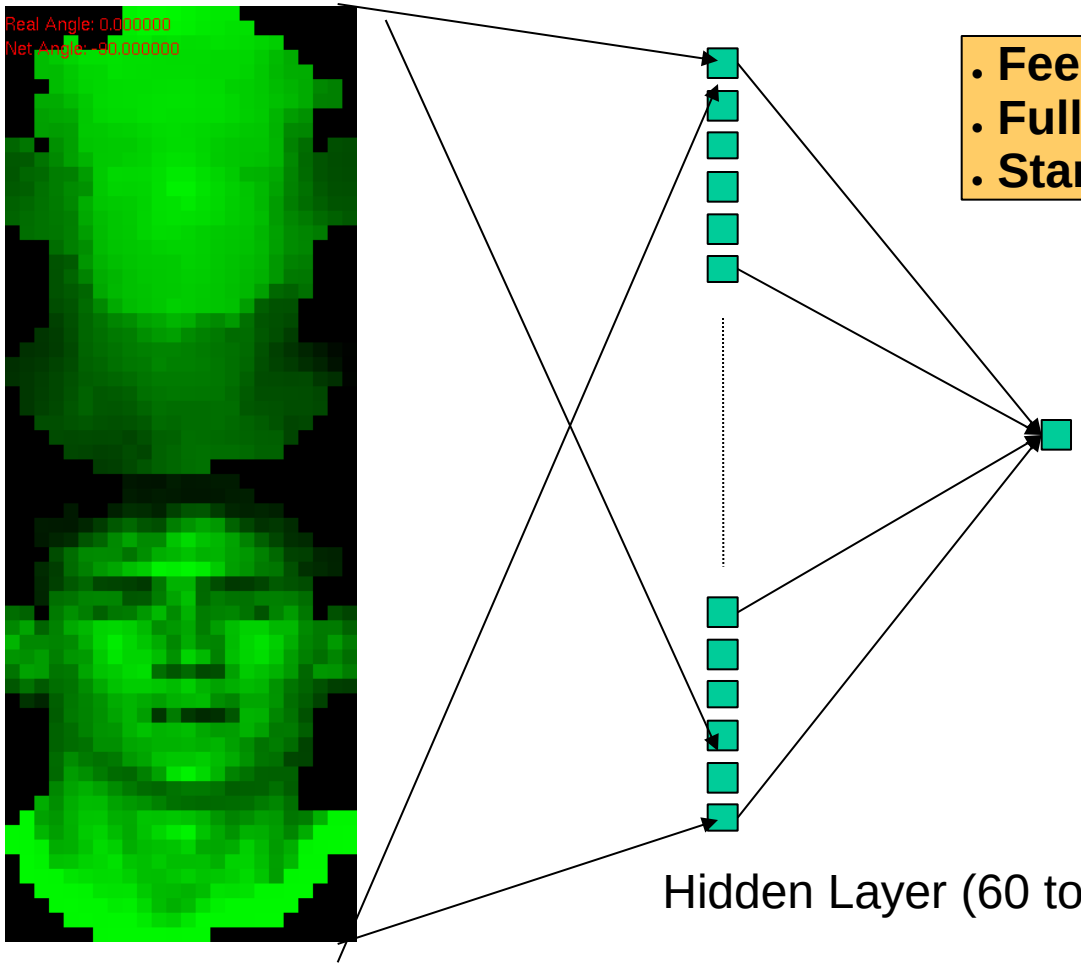


- Idea
  - Disparity (depth) images should be less affected by illumination changes than monocular greyscale images
  - Since both camera views share the illumination change, correspondence finding should still be robust
  - → Use disparity images for pose estimation





Estimating Head Pose with ANNs – Network Architecture

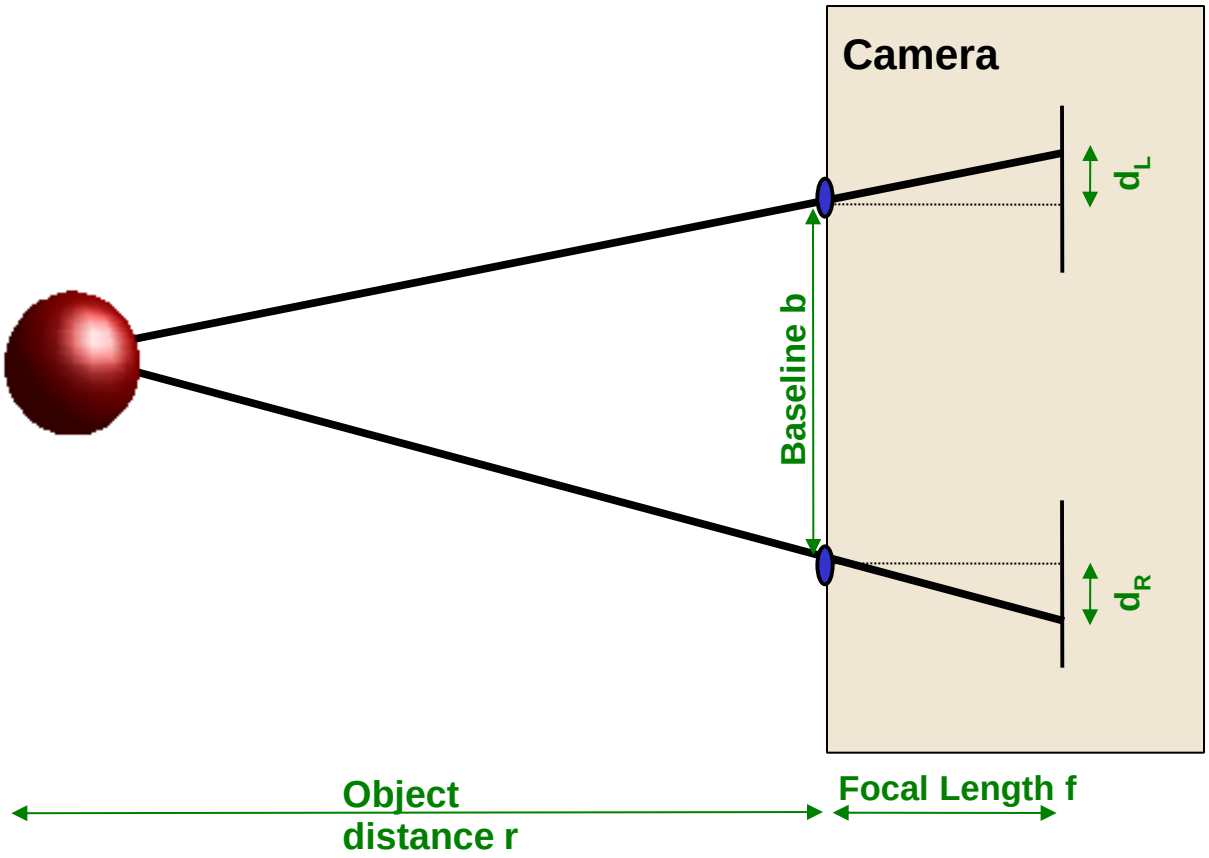


- **Feed-forward network**
- **Fully connected**
- **Standard back-propagation**

Network Output:
Rotation angle (pan, tilt or roll)

Hidden Layer (60 to 80 units)

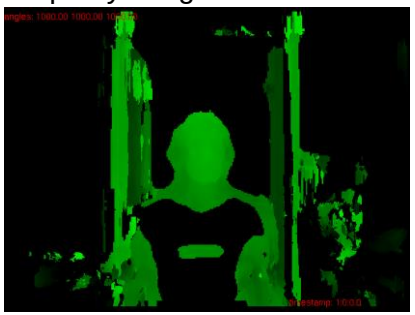
Stereo Image Geometry



Distance:

$$r = \frac{b \bullet f}{(d_L - d_R)}$$

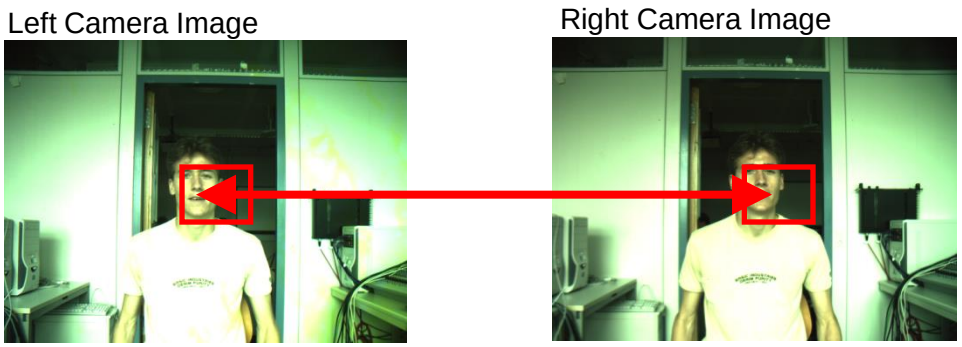
Disparity Image



Close pixel Far pixel

Finding Corresponding Pixels

- Summing the absolute value of differences over a small window
  (area correlation)
- Post-Filtering:
  - Confidence measure based on edge energy, and a left/right match consistency check



Estimating Head Pose with ANNs – Data Collection

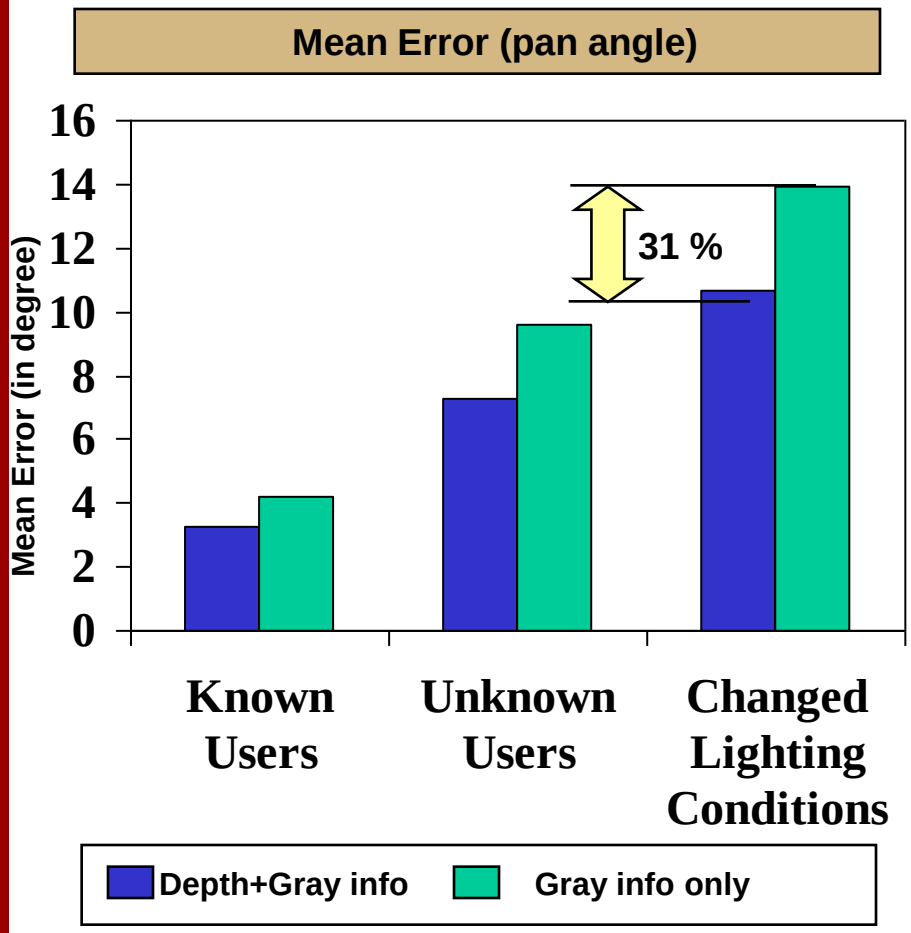
- 10 users
- 2 different lighting conditions (day light, artificial light)
- 250-500 images per person and lighting condition
- Reference angles captured with magnetic sensor (FoB)
- Resolution 640x480

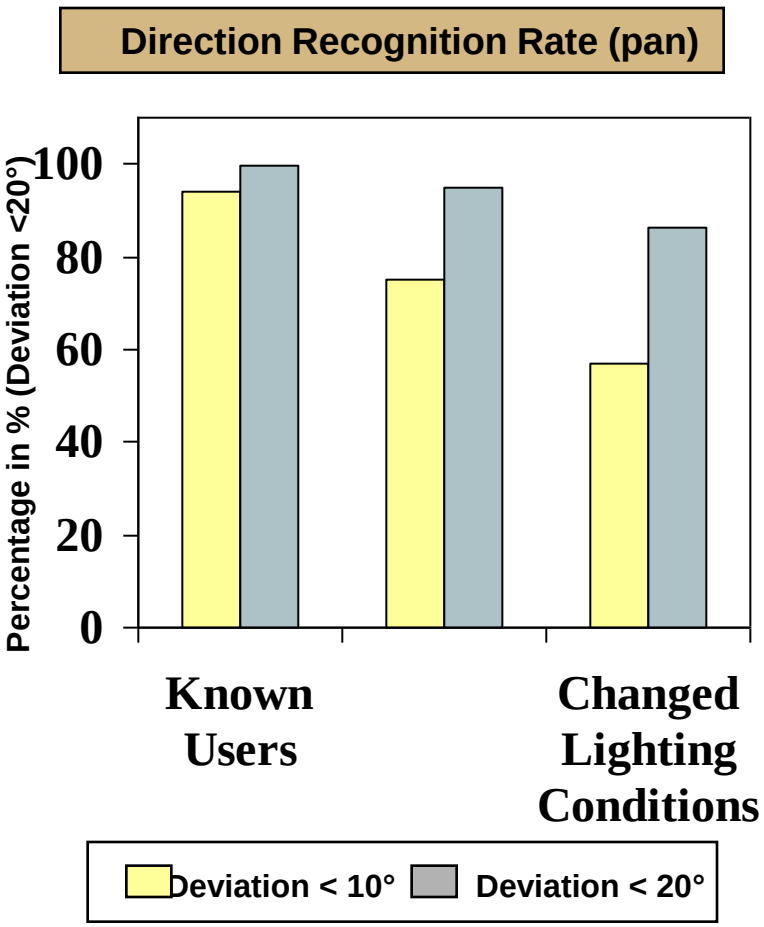



Test persons could move freely in pan, tilt and roll direction

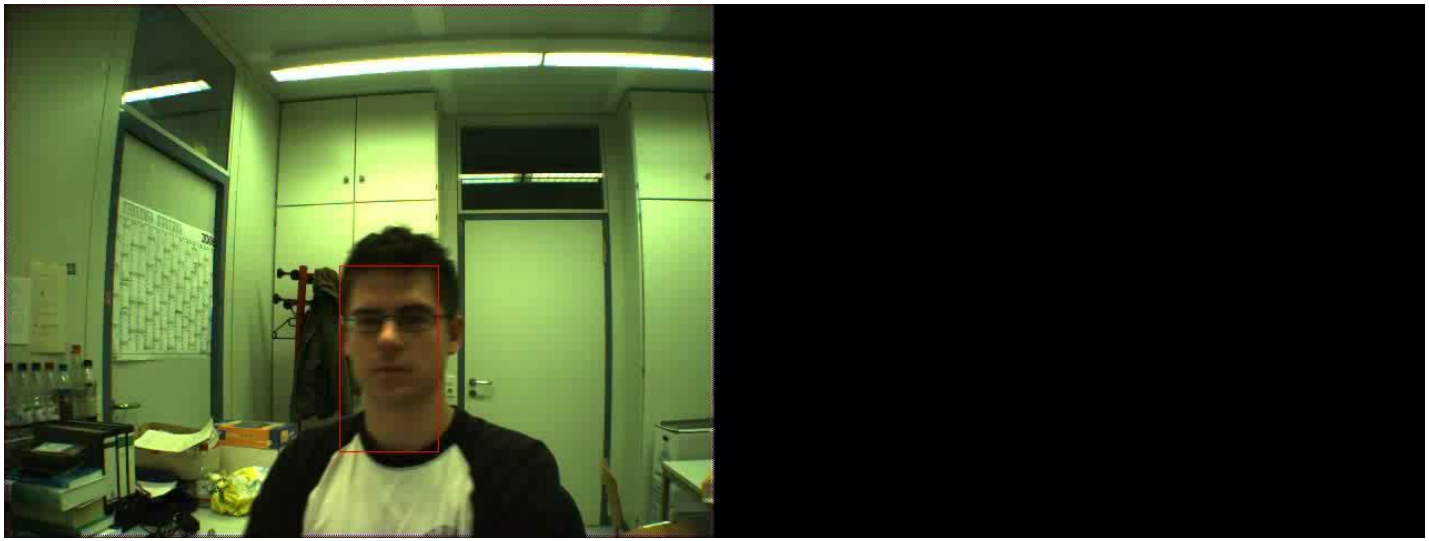
Estimating Head Pose with ANNs – Results under Changed Illumination

Experimental Results





Video



System details:

- depth from stereo + Haar-Cascades to detect face
- color to Track Face
- one neural network to estimate head pan
- input images:
  - Normalized greyscale + edges
- trained on data from ca. 15-20 people

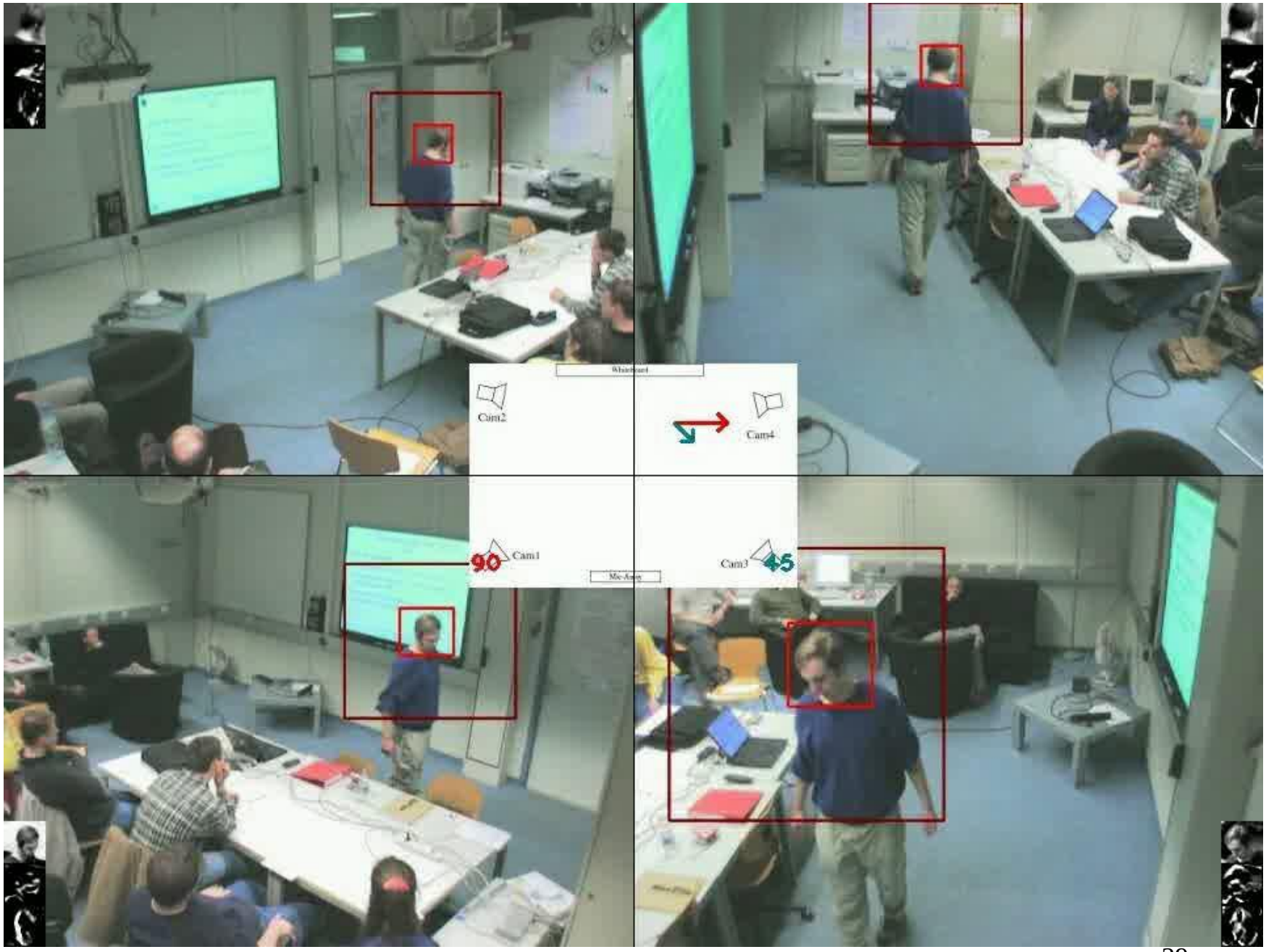
Kopfdrehung eines Sprechers im Smart Room

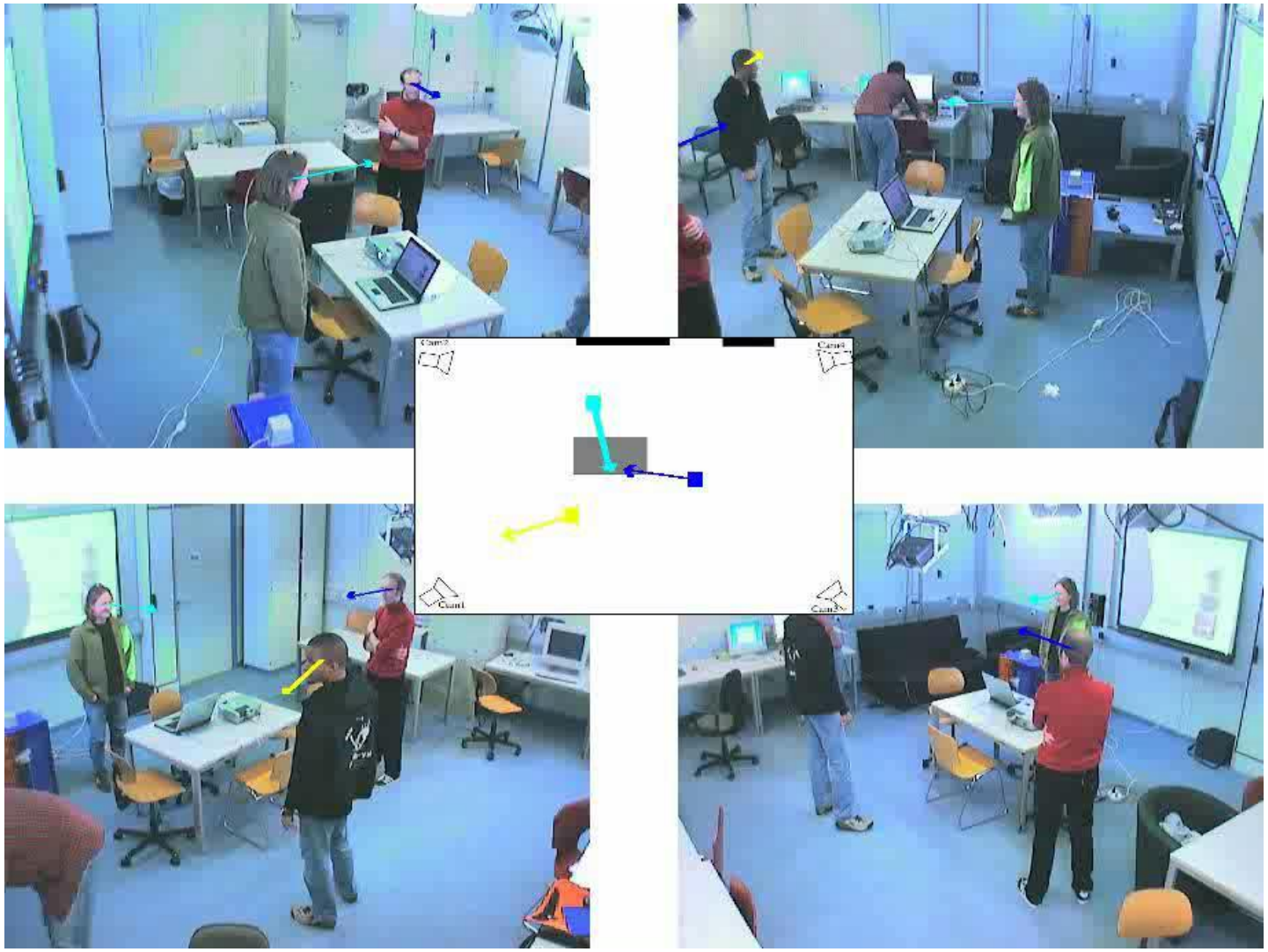
- Idee:
  - Benutze mehrere Kameras zur Schätzung von Kopfdrehungen
  - Fusioniere Ergebnisse

- Vorteile
  - Größerer Arbeitsbereich
  - Robusteres Ergebnis

- Ansatz:

1) *Lokalisierung Tracking der Person(en)* √
2) Detektion des Gesichtes und Bestimmung der Kopfdrehung pro Bild
3) Fusion der Ergebnisse


Cam2
Cam4
Cam1
90
Cam3
45



Other apperance based approaches

- Train detectors for various fixed poses
  - E.g. using Viola&Jones approach
  - Angular resolution is limited (typical resolution: profile / half profile / frontal)
- Match input images to sample templates
  - E.g. view-based Eigenspaces (Pentland '94, see Lecture "Face Recognition I")
- 3D morphable model (Blanz & Vetter)
  - See lecture "Face-recognition II"
  - Quite slow!
- Register a facial image to a 3D surface model
  - E.g Cascia, Isidor & Sclaroff 1998, Head tracking via robust registration in texture map images
  - Problem is initialization (frontal image needed) & error accumulation

Head pose estimation techniques – (Dis-)Advantages

- **Model-Based**
  - Tracked features can be used for other applications, such as lip-reading
  - Features are difficult to find and track
  - Range is limited due to occlusions
  - Good resolution is needed
- **Appearance-based (ANN)**
  - Only the head has to be detected / tracked
  - No limitation of rotations
  - Works with low image resolution
  - No initialization required
  - No error accumulation („drift“)
  - Illumination changes are problematic
    - Stereo/Depth helps
  - A lot of training data is needed

Tracking Focus of Attention (FoA)

- **Focus of Attention tracking**:
  - To detect a person's interest
  - To know what a user is interacting with
  - To understand his actions/intentions
  - To know whether a user is aware of something

- **Human-Human Interaction:**
  - to determine the addressee of a speech act
  - to understand the dynamics of interaction
  - for meeting indexing / retrieval
- **Human-Robot Interaction**
  - Was the robot addressed or not?
- Smart Environments, Cars, …

FoA Tracking in Meetings

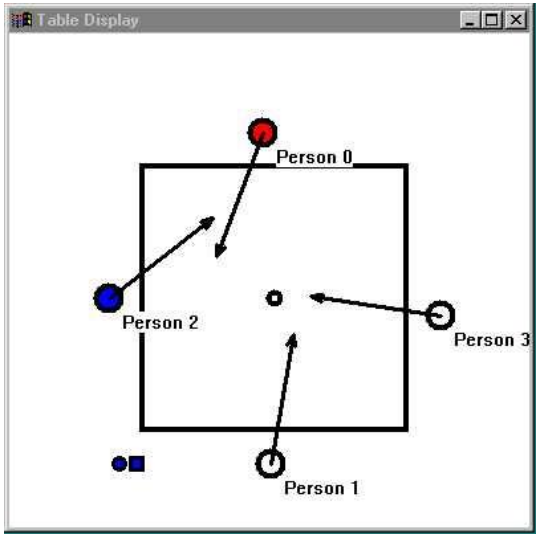
Why:

- to determine the addressee of a speech act
- to track the participants attention
- to analyse, who was in the center of focus
- for meeting indexing / retrieval), …

How:

1. Track all participants‘ faces **(color)** √
2. Estimate their head orientations **(ANNs)** √
3. Map head orientations onto likely targets (e.g. the other participants)

FoA: Detection of likely Targets



Idea: For each person, find the most likely target person ***T***, given the observed head orientation ***x***

Modeling Focus from Head Orientation

- Head Orientation is a strong indicator of social attention
- Eye-Gaze is difficult to measure

Estimate a persons focus of attention based on his head orientation

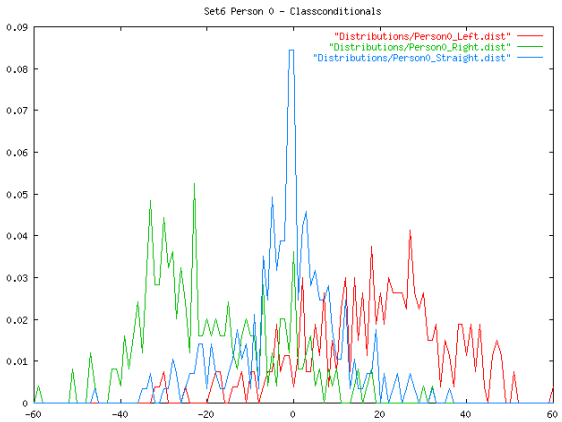
- can be formulated with Bayes rule:

$$P(Focus_S = T \mid x) = \frac{p(x \mid Focus_S = T) \cdot P(Focus = T)}{p(x)}$$

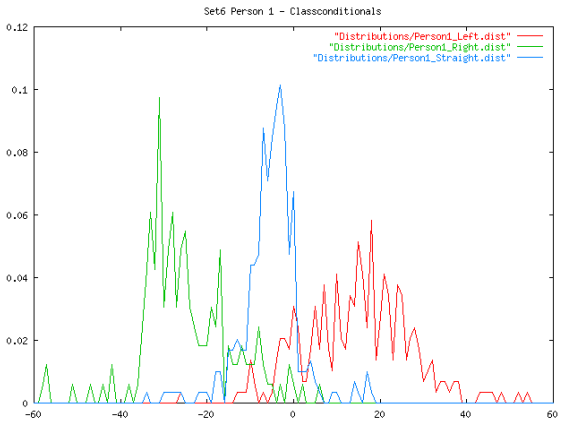
x: head pan in degrees,

$T \in \{Person\ 1, Person\ 2, \ldots, Person\ M\}$

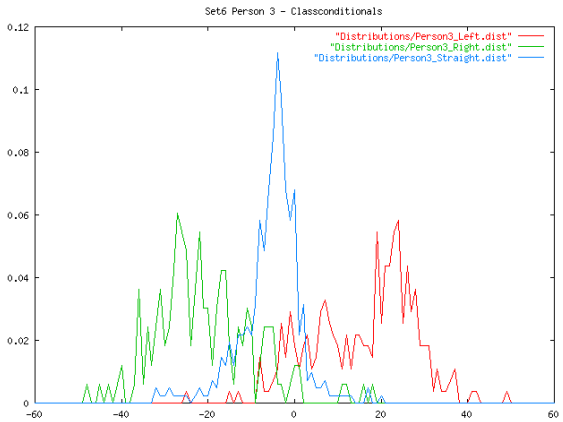
Head Pan Distributions $p(x|F_i)$



Person 1



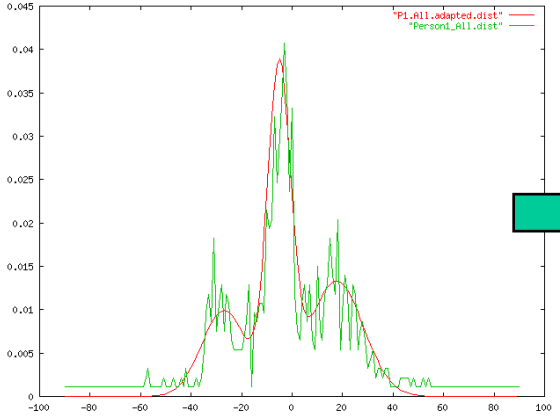
Person 2



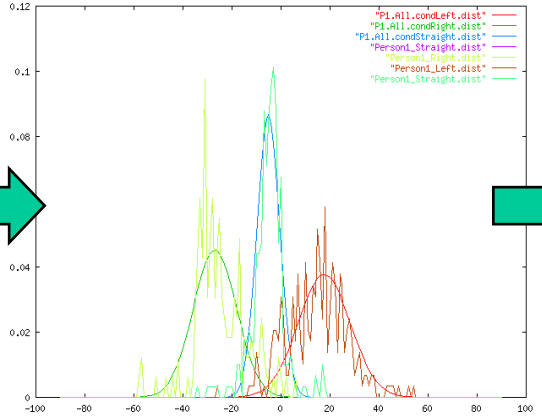
Person 3

- Head pan distributions are dependent on
  - - personal head-turning "styles"
  - location of targets
- Distributions should be adapted for
  - - each person
  - - each meeting

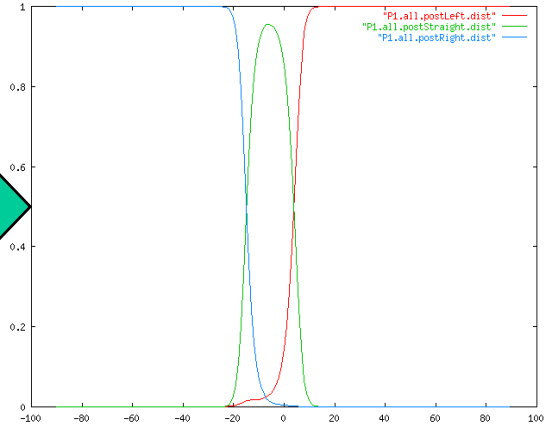
Adaptation of the Model



All observations: p(x)



Class-cond. Distr. p(x|F)



Posteriors P(F|x)

P(x) is modeled as mixture of Gaussians: $p(x) \approx \sum_{j=1}^{M} p(x \mid j) P(j)$

- Model parameters are found using EM-algorithm*
- Individual Gaussian components are used as class-conditionals $p(x|F)$
- Priors of the mixture model $P(j)$ are use as focus prior $P(F = T)$

(* Expectiation-Maximization)

Focus based on Head Orientation

- Results on four meetings
  - 4 participants in each meeting
  - Participants automatically detected and tracked
  - unsupervised adaptation of model parameters

	P(Focus | Gaze)
Meeting A (4 participants)	68.8 %
Meeting B (4 participants)	73.4 %
Meeting C (4 participants)	79.5 %
Meeting D (4 participants)	69.8 %
Avg.	**72.9 %**

Percentage of correctly assigned focus targets
based on computing P(Focus | Head pan)

Head Orientation & Eye Gaze in Meetings

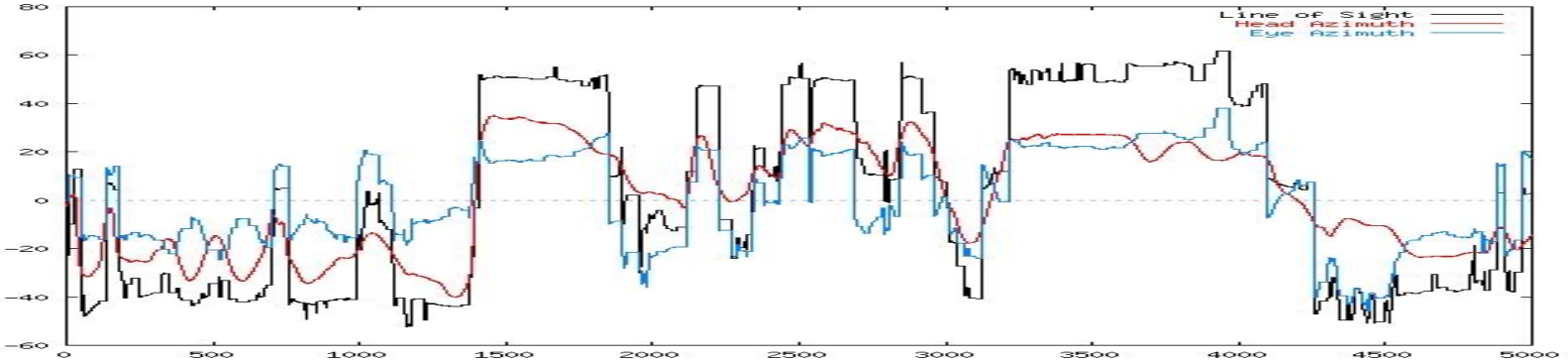
Experiment:

Four people in a meeting, head orientation and eye-gaze measured using ISCAN head-eye tracker

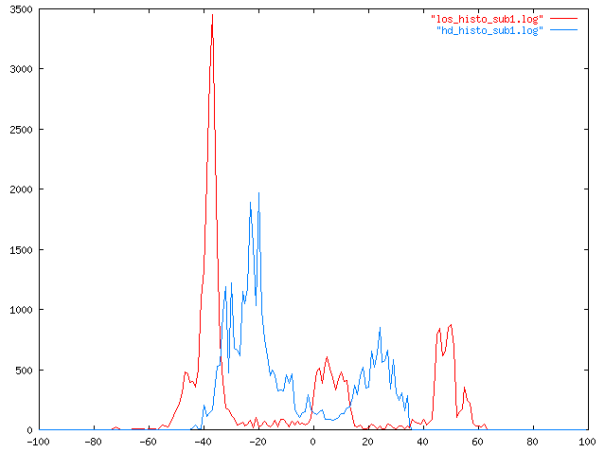
- Head orientation and eye gaze point to the same direction 87% of the time
- Head orientation accounted for 69 % of the overall horizontal gaze direction
- Focus can be predicted based on head orientation in 88.7 % of the frames (using our model ...)

Head Orientation & Eye Gaze in Meetings (2)

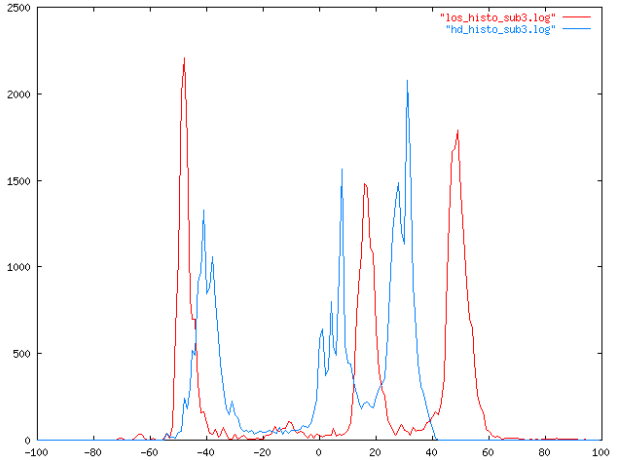
Plot of horizontal orientation of head, eye and line-of sight:



Histograms of line of sight and head orientation:



Person 1



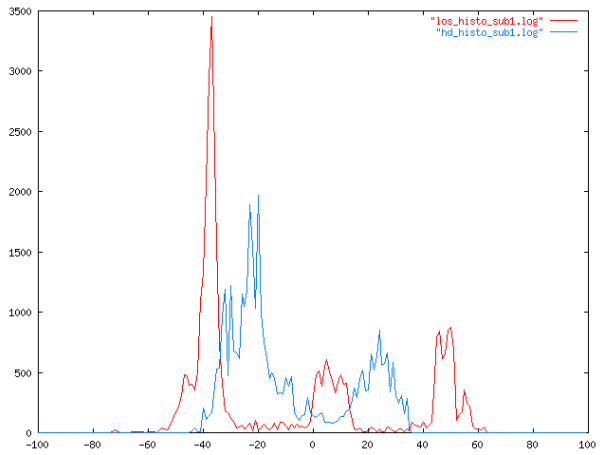
Person 2

Head Orientation & Eye Gaze in Meetings (3)

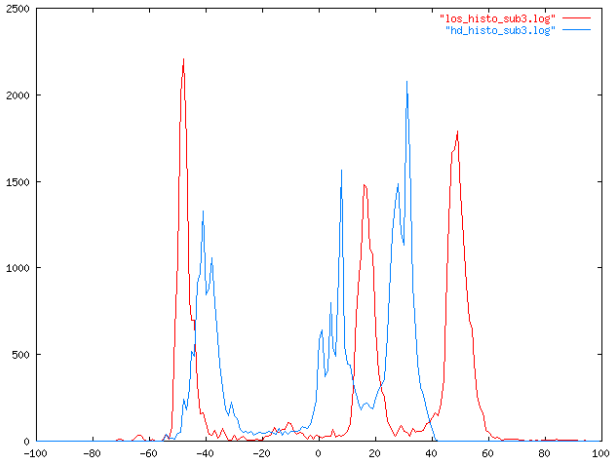
- Four people in a meeting, head orientation and eye-gaze measured using ISCAN head-eye tracker

Head Orientation & Eye Gaze in Meetings (4)

Histograms of line of sight and head orientation:



Person 1



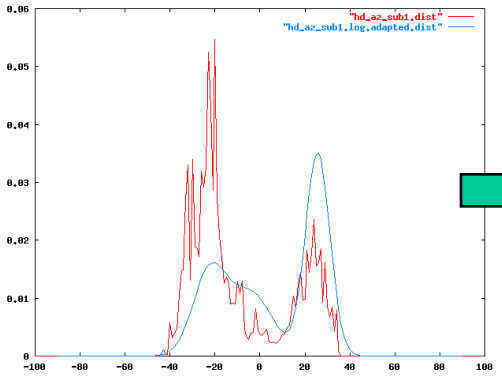
Person 2

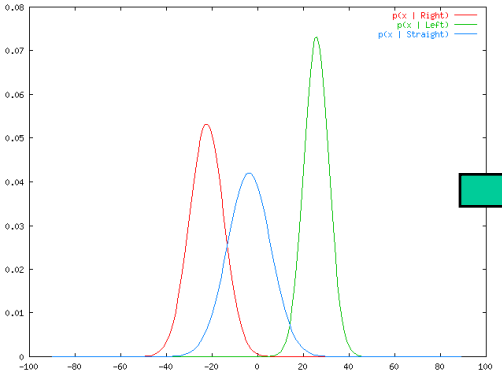
➔ Focus can be predicted based on head orientation in 88.7 % of the frames (using our model ...)

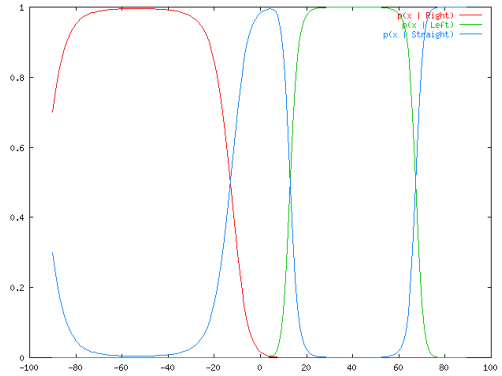
Contribution of Head Orientation

Subject	#Frames	Eye Blinks	Same direction	Head Contribution
1	36003	25 %	83 %	62 %
2	35994	23 %	80 %	53 %
3	38071	19 %	92 %	64 %
4	35991	20 %	93 %	97 %
Average		22 %	87 %	69 %

Focus from Head Orientation



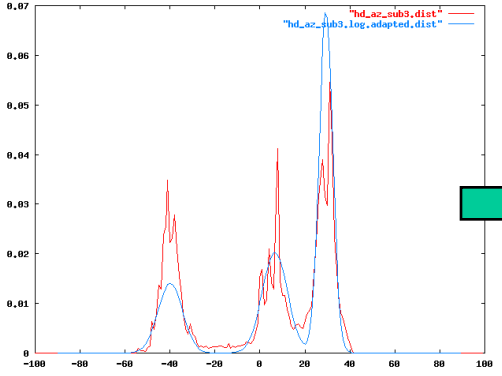


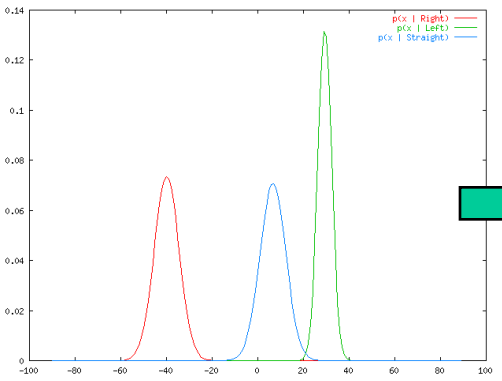


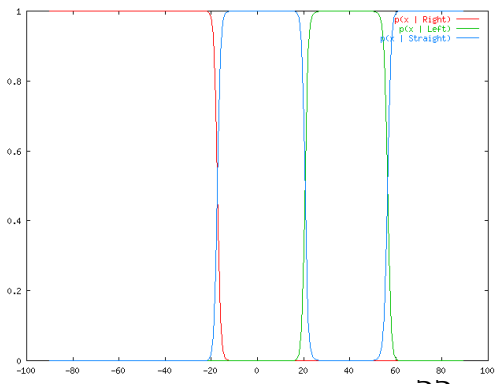
Adapt Mixture Model
to Head-Orientation Data.

Use Components
as Class-Conditionals

P(Target | Head Pan)







Focus Detection Results

Subject	Accuracy
1	86 %
2	83 %
3	93 %
4	93 %
Average	**89 %**

FoA for Indexing and Analysis of Meetings

- Who talked to whom ?
- Activity / Attention Analysis of Meetings
  - temporally filtering the focus / speaking activity
- Meeting statistics:
  - how often were persons paying attention to the speaker
  - how often did people talk ?
  - how often have people been paid attention to ?
- In a Meeting Browser
  - to get additional tags for retrieval
  - for focus-oriented replay of (parts of) meetings

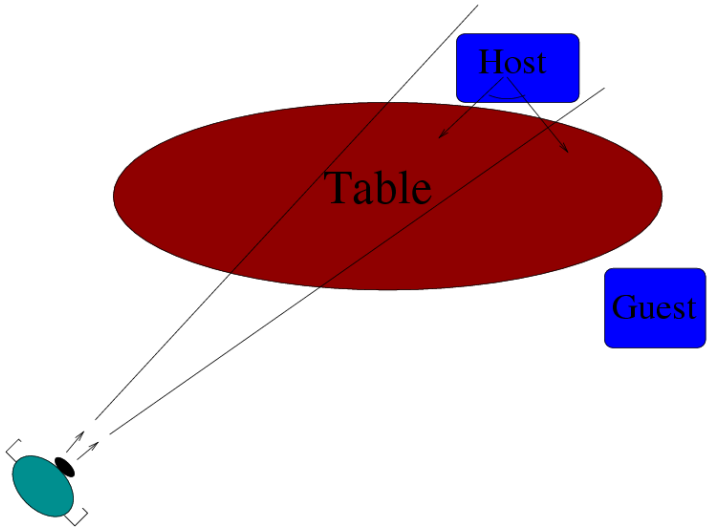
Applications: Human-Robot Interaction

See: www.sfb588.uni-karlsruhe.de

- **Focus of Attention tracking**:
  - to determine the addressee
    - was the robot addressed ?
  - to detect areas and objects of interest
  - to establish joint/shared attention

Determing Addressee in Human-Human Robot Interaction

- Interaction between two humans and a „robot“
  - One host, one guest
  - „Robot“: One PTZ-camera, distance microphone (also a close-talking microphone was used)
- Task: demonstrating the new toy robot to a friend
  - Includes talking about the robot
  - Commands towards the robot
  - Discussing pros and cons
- Recorded data / Transcripts:
  - 18 sessions, roughly 10 min each
  - Includes around 500 commands
  - Speech fully transcribed
  - Addressee manually labeled



From the intructions

- what the robot can do
  - get beer, soda, coffee, tea, candies, cake, newspaper, shoes, ...
  - clean up the table, clean the floor (using a wet rag and a vacuum cleaner), clean the windows, ...
  - turn on/off the lights, dim/brighten the lights,
  - increase/decrease temperature, change air humidity
  - turn on/off music, change settings of equalizer, change loudness ('turn down the volume', ...)
- the robot is good at:
  - monitoring your house
  - helping disabled people
  - helping the elderly
  - helping with cooking

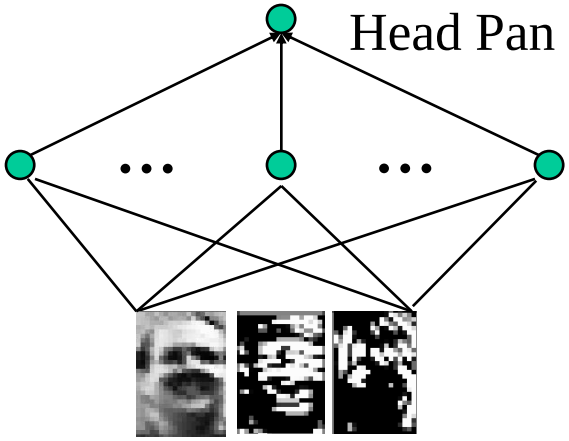
Head Pose and Addressee

- How well is the host's visual target related with the addressee of his speech acts ?

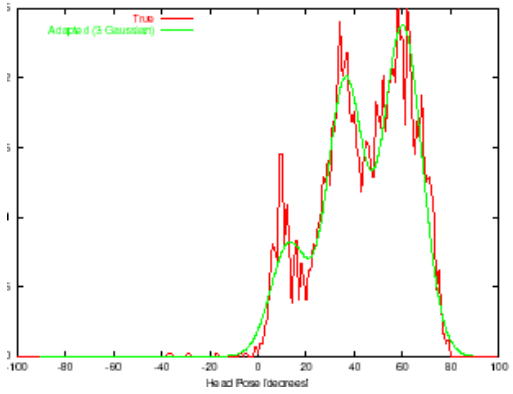
Audio \ Video	$T_V = Guest$	$T_V = Robot$	$T_V = Other$
session 1 $T_A = Guest$	462	44	202
$T_A = Robot$	3	43	2
session 2 $T_A = Guest$	463	69	136
$T_A = Robot$	0	94	0
session 3 $T_A = Guest$	289	34	221
$T_A = Robot$	0	46	3
session 4 $T_A = Guest$	575	2	5
$T_A = Robot$	6	93	2
Sum $T_A = Guest$	**1969** (73%)	**149** (6%)	**564** (21%)
$T_A = Robot$	**9** (3%)	**276** (95%)	**7** (2%)

- when the host looked at the ***other human***, the other human was addressed in **99.5 %** of these cases
- when the host looked at the ***robot***, the robot was addressed in **65 %** of these cases

Visual Estimation of Adressee



- Based on Head Pose Estimation
  - Neural Network based
- Finding the most likely target
  - choose max $P(T_V=\text{Target}| x)$ as addressee
  - Head pose x, Target $\in$ {Robot, Human}
  - Need to find p(x | Target), modelled by GMM



Result:

Distributions	Precis.	Recall	F-Measure	Accuracy
True	0.89	0.77	0.82	0.96
Learned	0.74	0.85	0.79	0.93

$$f-measure = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

Ergebnisse

In 90% der Fälle konnte der Adressat nur aufgrund der automatisch geschätzten Kopfdrehung ermittelt werden

- „Wird der Roboter angesprochen?“: Precision: 0.57, Recall: 0.81 (f-measure: 0.67)

Kopfdrehung ist ein „asymmetrisches“ Merkmal:

- Wenn eine Person angeschaut wurde, war sie auch immer Adressat (99%)
- Wenn der Roboter angeschaut wurde, wurde er in 65% der Fälle auch angesprochen

Sprach-basierte Merkmale:

- Perplexitäten auf versch. Sprachmodellen, Parsebarkeit, Satzlänge, ...
- Bestes Ergebnis mit MLP: Precision.: 0.19, Recall 0.91 (f-measure: 0.31)

Kombination von Audio + Video führt zu Verbesserungen:

- In 92% der Fälle korrekter Adressat ermittelt
- „Wird der Roboter angesprochen?“: F-Measure = 0.72

Summary

- Head pose is a good indicator of attention
  - Meeting Analysis
  - Human-Robot Interaction
- Head pose estimation approaches
  - Model based
  - Appearance based: Neural networks
- Goal is to find most likely *focus targets* based on head orientation
  - Since eye gaze is lost
  - Can be combined with other cues (e.g. audio- and speech-based cues)

References

- Rainer Stiefelhagen, Jie Yang, Alex Waibel, ***A Model based Gaze Tracking System***, Proc. of IEEE International Joint Symposia on Intelligence and Systems, pp. 304-310, Rockville Maryland, November 1996

- Rainer Stiefelhagen, Jie Yang and Alex Waibel, ***Modeling Focus of Attention for Meeting Indexing based on Multiple Cues***, IEEE Transactions on Neural Networks, July 2002, Vol. 13, Number 4, pp. 928-938.

- M. Katzenmaier, R. Stiefelhagen, T. Schultz, I. Rogina, A. Waibel, ***Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech***, International Conference on Multimodal Interfaces ICMI 2004, State College, PA, USA, October 2004

Combination of Gaze and Sound

Combination: $(1- \alpha)\ P(F_i|Gaze) + \alpha\ P(F_i|Sound)$

	Video only	Audio only	Combined (α=0.6)
Set A	68.8	63.0	71.4
Set B	73.4	67.2	77.1
Set C	79.5	60.2	80.5
Set D	69.8	72.1	75.7
Avg.	72.9	65.6	76.2

- Accuracy increases from 72.9 % to 76.2 %
- 12 % relative error reduction using a fixed α