Human Activity and Vision Summer School

Recognition of Visual Focus of Attention in group conversation

Extraction of head and body pose cues in open spaces

jean-marc odobez

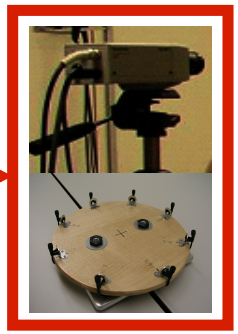
03.10.2012

idiap RESEARCH INSTITUTE

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

the overall goal: to infer relevant information from audio-visual human scenes

audio-visual scenes



- representation (what is a person?)
- detection (are there any people?)
- localization (where are they?)
- tracking (where do they go?)
- identification (who are they?)
- activity recognition & discovery (what do they do? what do they look at?, do they interact? who do they interact with? what do they do together?, …)

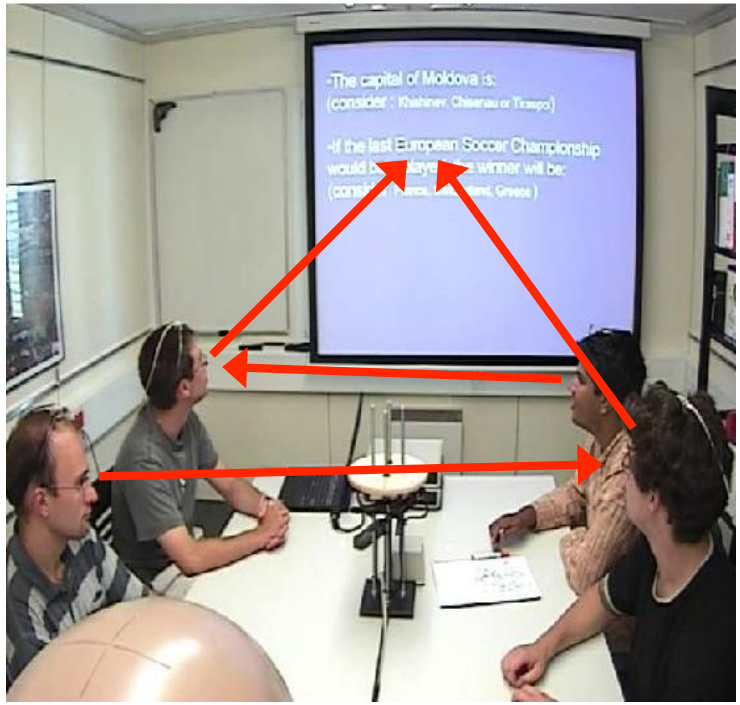
Visual focus of attention (VFOA)

- Focus of attention (**eye gaze)**

  “who is looking at whom (or what)”

- Non-verbal signal which conveys rich information about a person
  - his interests, what is he doing
  - how does he explore a new environment ?
  - reaction to different stimuli

- Gaze is a strong social interaction cue
  - regulate conversation
    turn taking/yelding cue
  - social control
    => dominance, personality traits



Applications of VFOA recognition

- **Group support** [DiMicco, MIT2004,Kaplan EPFL2006]
  - holding the floor too much is perceived as overcontrolling
  - people not looked at feel frustrated/excluded
  - affects group cohesiveness & effectiveness

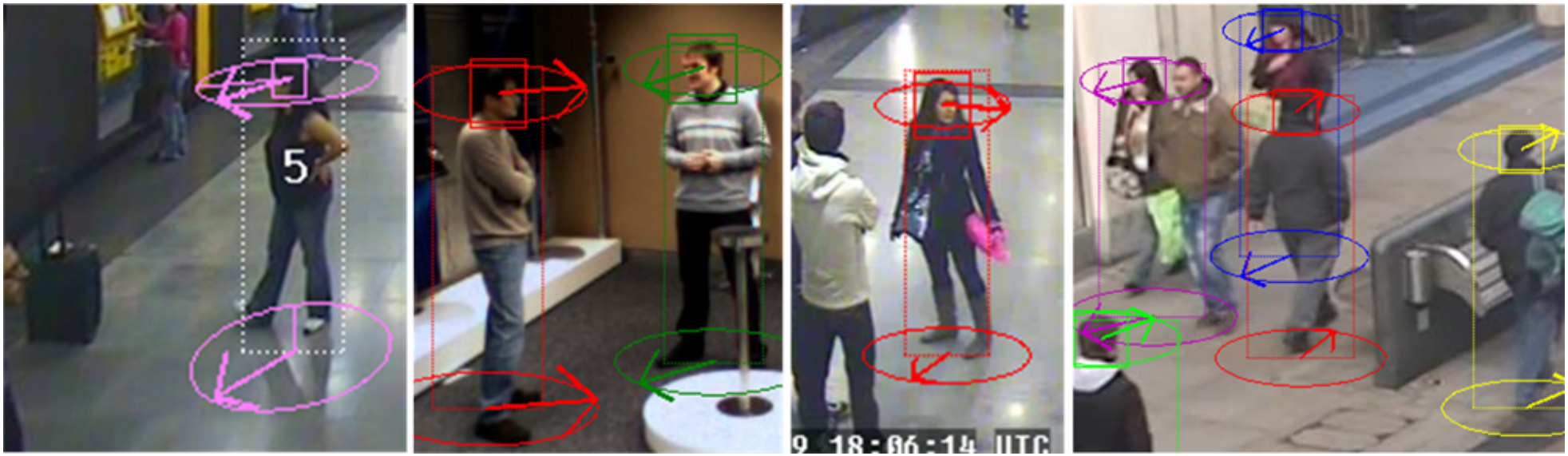
Sturm, Eindhoven, 2007

- **Addressee** recognition
  - Human-computer/robot interaction
    - presence of several people
    - **artificial agent**: needs to know whether it/he is addressed or not

Information kiosk

=> **gaze is a good predictor** of addressee-hood

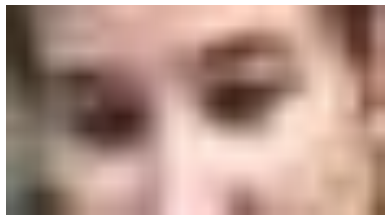
Surveillance and Smart rooms



- Extraction of behavioral cues (head pose, body pose, VFOA)
  - Move beyond location based analysis
  - Better characterize the state/instantaneous activity of individuals
- Application
  - Security (e.g. left luggage attendance detection)
  - Group identification
  - Behavior and interaction analysis
  - Scene or poster attraction statistics

VFOA Recognition challenges

- gaze can (often) not be measured directly
  - HCI gaze estimation approaches
    - invasive, restrict mobility
    - interfere with natural conversation
  - video resolution is not enough

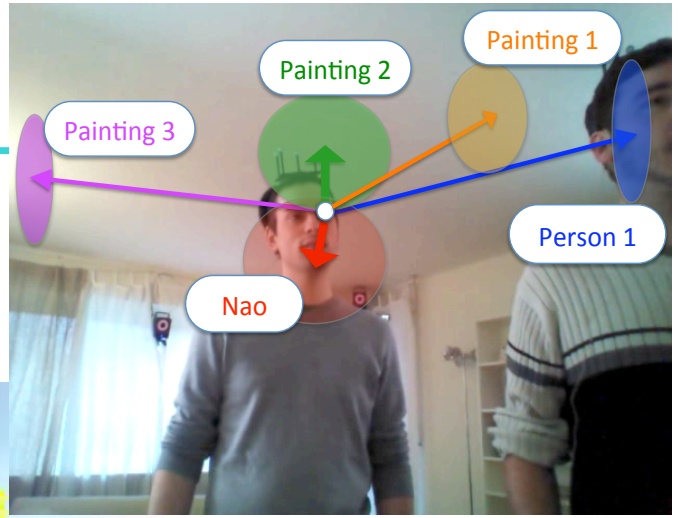


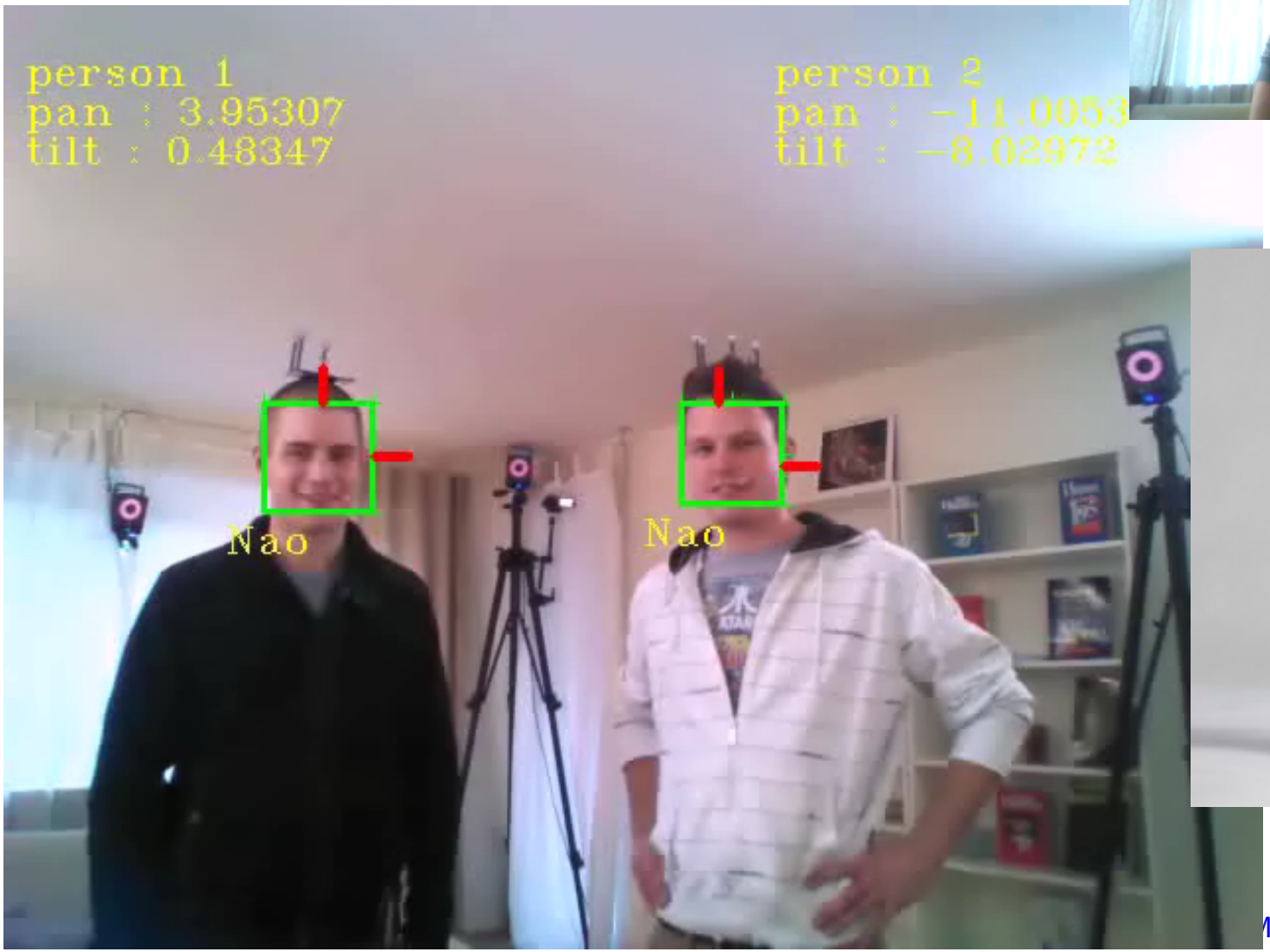
=> use head pose as a surrogate

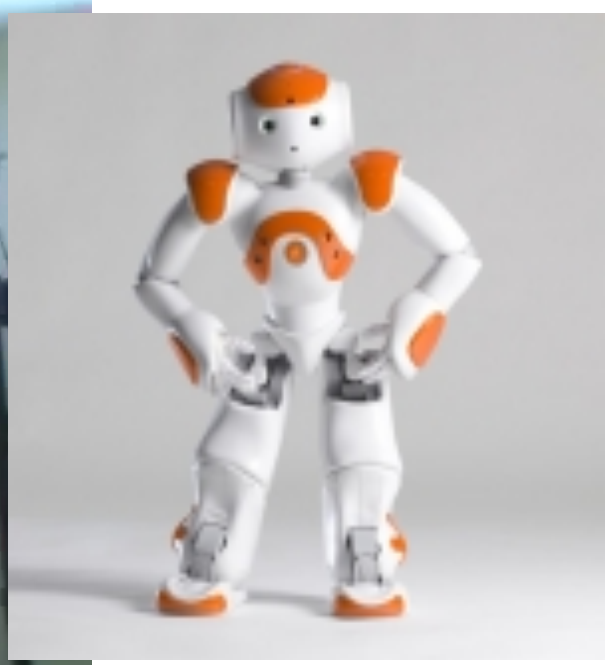
- psychological evidence [Langton et al, 2000]
- empirical evidence [Stiefelhagen et al, 2002, Otsuka 2006]

[Voit et al MLMI 2008]

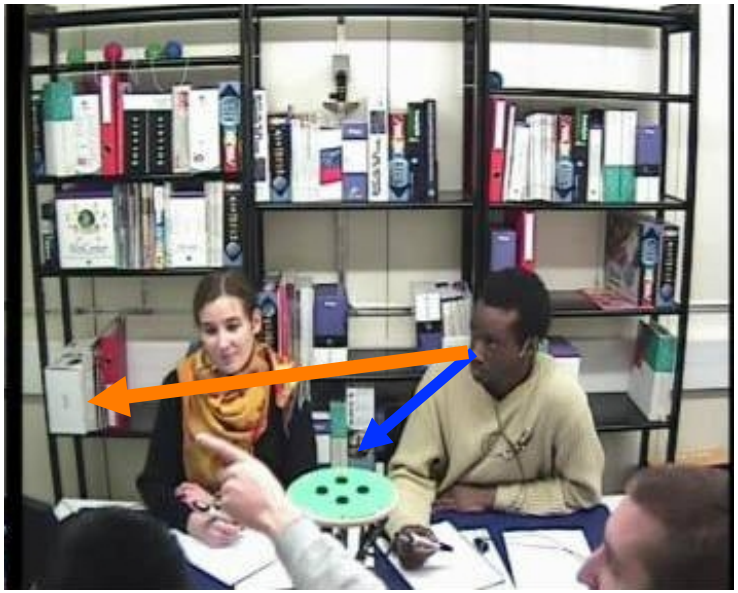
Example in HRI







VFOA Recognition challenges



Interpretation of head pose or gaze

- gaze directions need to be mapped/associated to VFOA targets

  = 3D objects/people in the 3D space

  ⇒ knowledge and monitoring of the environment

- head pose ≠ gaze direction

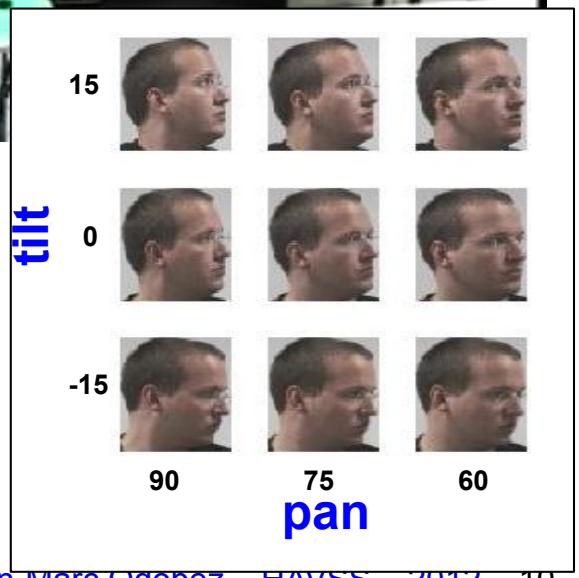
  => pose needs to be mapped to gaze direction

  => mapping is ambiguous, depends on context (activity, social)

Presentation plan

- VFOA analysis in group conversation
  - Head pose is the main cue
    how and how well can we estimate it ?
  - VFOA modeling
- Head and body pose extraction in open settings

Joint Head Location and Pose Tracking



- on 60 minute data: around **10-13 degree** error in pan
- tilt more difficult to estimate
- larger error near profile views
- large accuracy variation across people

(depending on appearance; some people easier to track)

Presentation plan

- VFOA analysis in groups
  - Head pose estimation

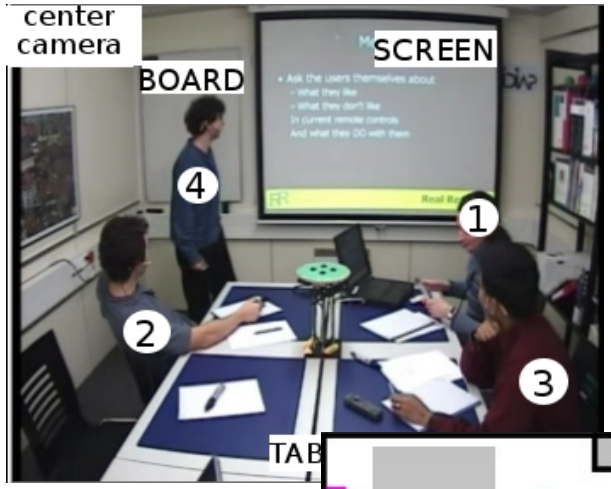
  - VFOA modeling in the context of meetings
    - Task
    - Dynamic Bayesian Networks for contextual multi-party VFOA recognition
    - Important issues

  ⇒ Goals: illustrate one one example
    - How to integrate context in a recognition problem
    - How to introduce context which have temporal
    - How to exploit soft labels from prior knowledge about normal behavior to increase recognition accuracy

- Head and body pose extraction in surveillance scenarios

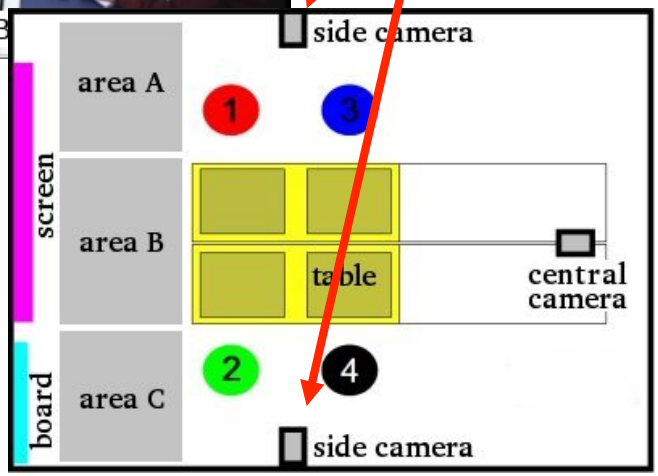
Set-up and task

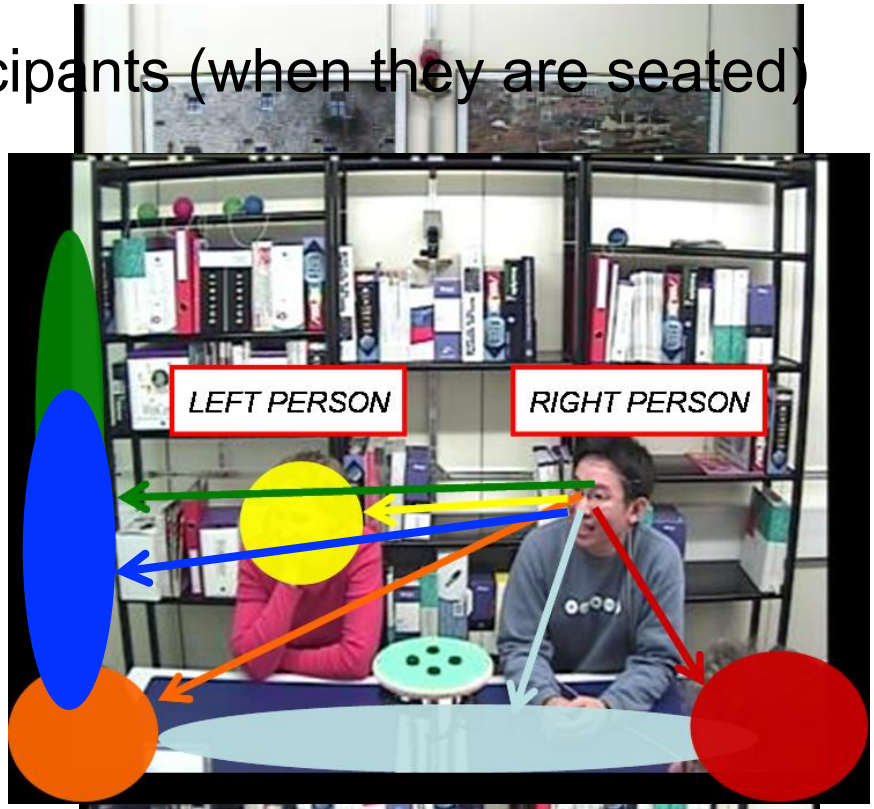
- **task:** recognize the VFOA of all participants (when they are seated)



center camera

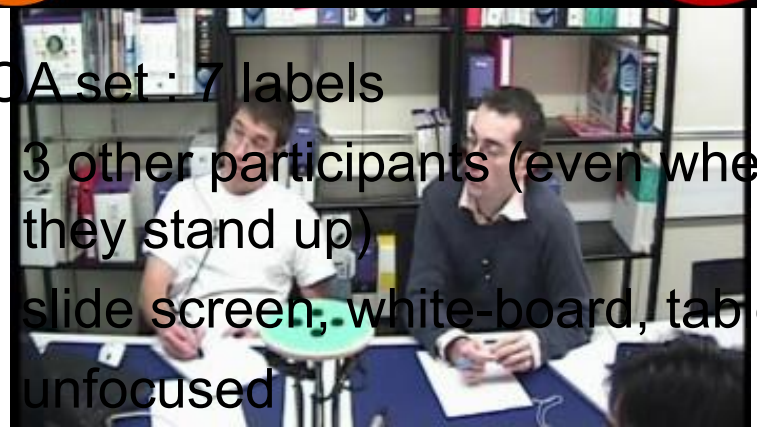
side cameras





- Setup: 4 persons
  - 3 cameras
  - head sets microphones

- FOA set : 7 labels
  - 3 other participants (even when they stand up)
  - slide screen, white-board, table
  - unfocused

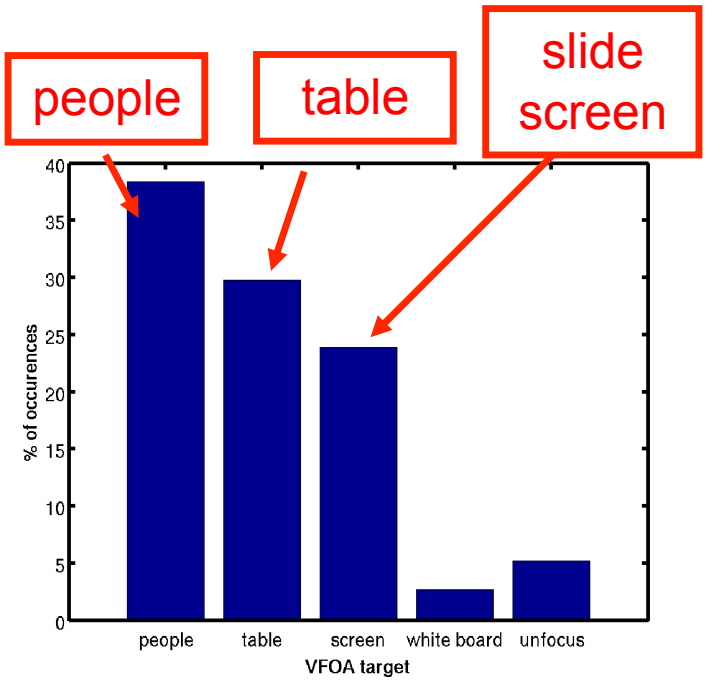


Dataset

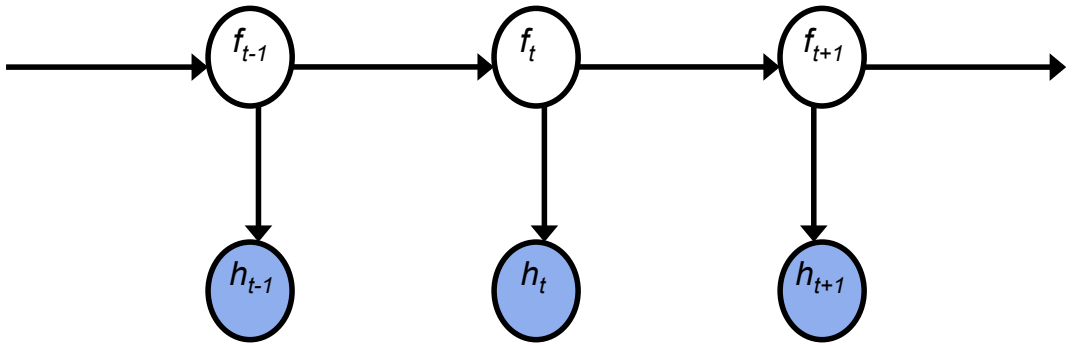
- 12 meeting from AMI public corpus
  - 4 **static** meetings (90min) – only seating people
  - 8 **dynamic** meetings (210 min)- people standing (33% of the time)

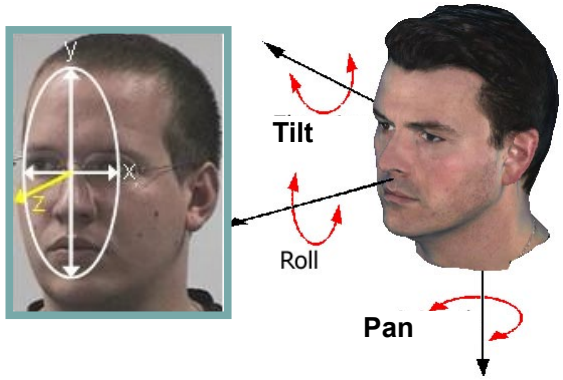
- real behavior
  - body poses, gaze behavior, gestures
  - laptop and object manipulation

- VFOA analysis
  - only 38% looking at people
  - around 30% looking at table
    - looking at laptop
    - 'long-meeting' effect



VFOA modeling using HMM





- Input: head pose features h_t (pan/tilt angles at time step t)
- Ouput: recognized VFOA f_t

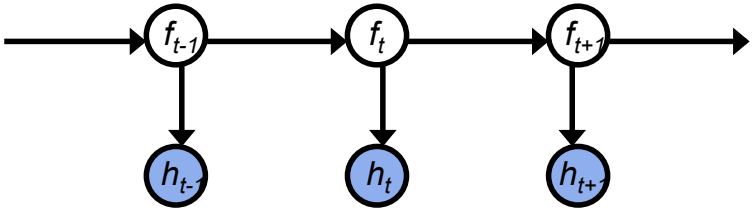
$$p(f_{0:T}, h_{1:T}) = p(f_0) \prod_{t=1}^{T} p(h_t|f_t) p(f_t|f_{t-1})$$

- Method: HMM statistical model
  - Dynamic model
    higher self-loop => smooth decoding
  - Observation model (likelihood)

Observation model

Dynamical model

VFOA modeling using HMM

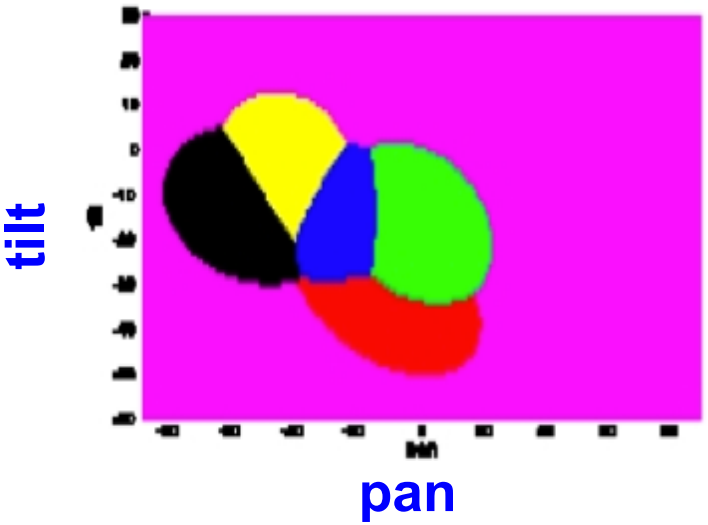


- **observation likelihood $p(h_t \mid f_t)$**
  - Gaussian distribution

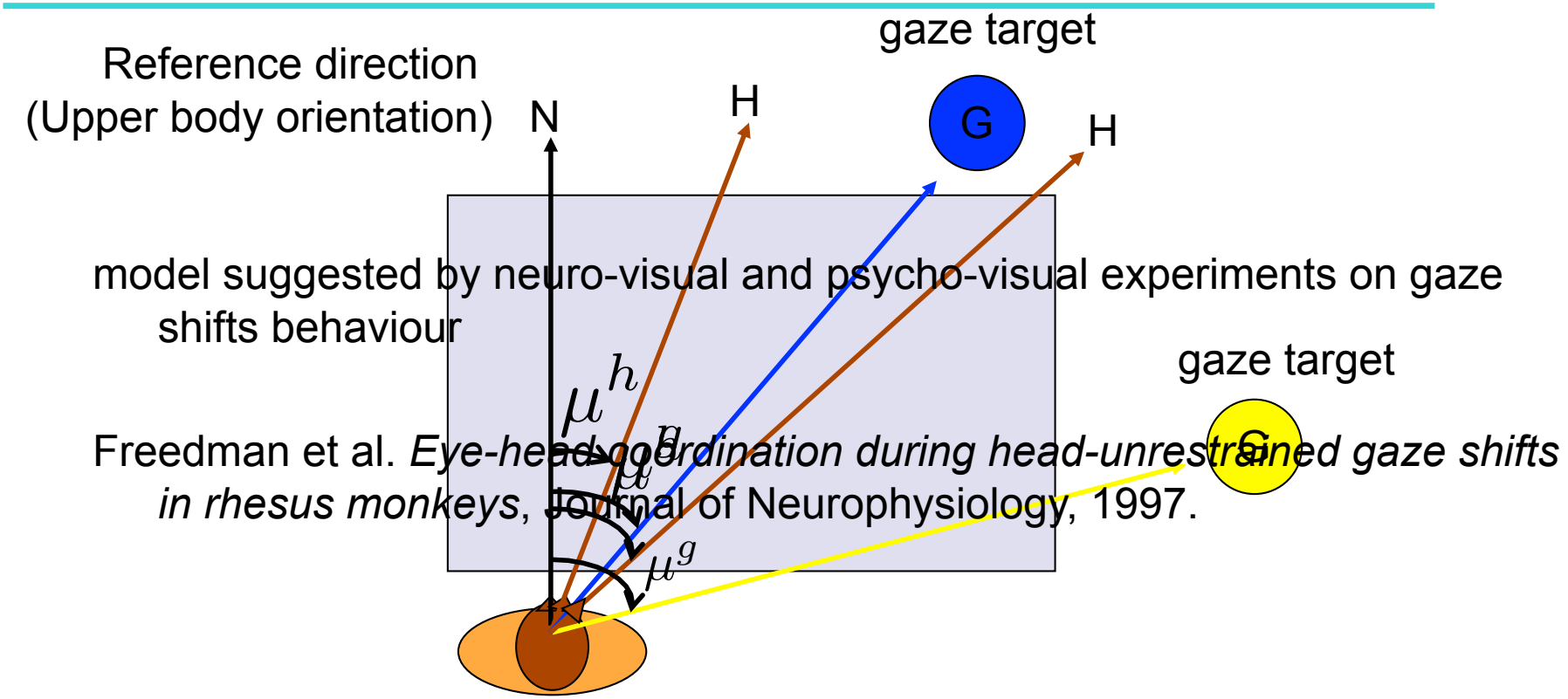
$$p(h_t|f_t = i) = \mathcal{N}(h_t|\mu_i, \Sigma_i)$$

For considered person, mean head pose corresponding to looking at target i

- How to set these mean head pose values ?
  - Supervised learning ?
    labeling time consuming
    training data needed each time we change set-up
  - Use 3D information ?
    head pose orientation ≠ gaze direction
    => need for a gaze-to-head mapping function

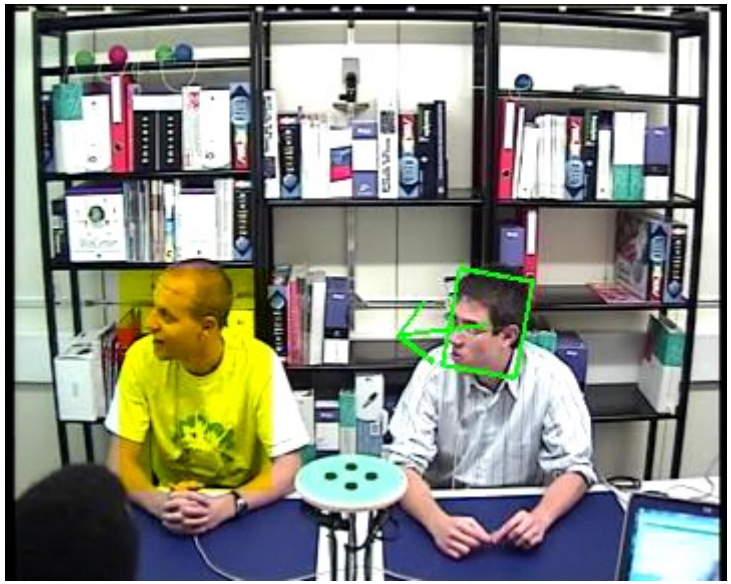


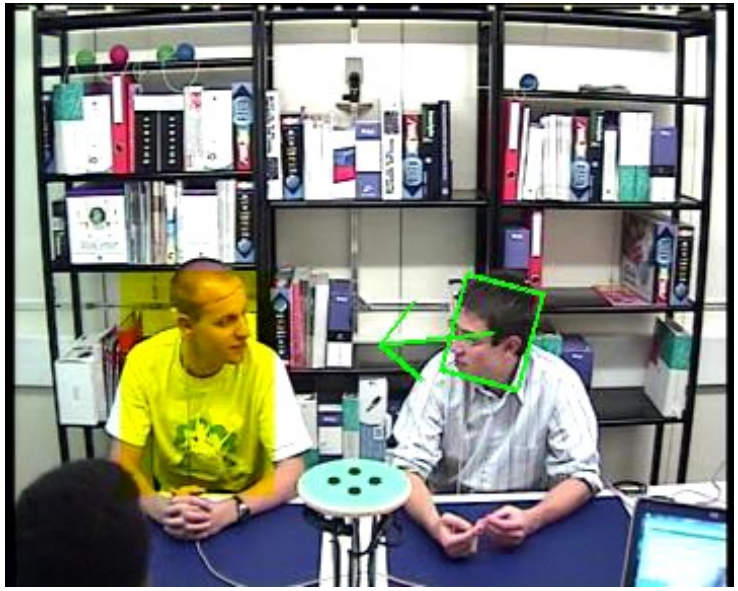
Gaze mapping function



- gaze direction - head orientation relationship
  - linear relation $\mu^h = \alpha\mu^g$
- VFOA model parameters
  - Gaussian head pose mean parameters predicted by cognitive model
  - Requires approx. camera calibration, position of people and VFOA target

Ambiguities + context

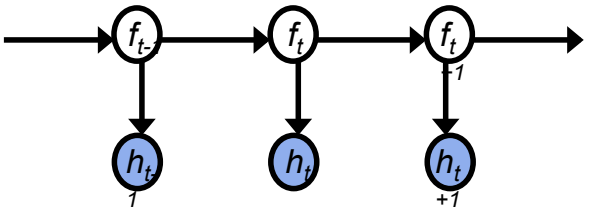


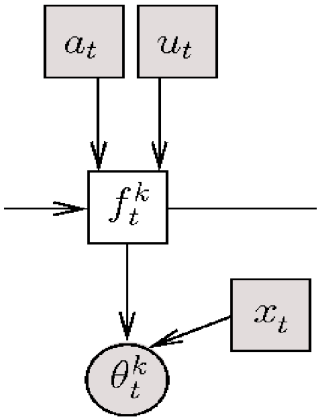


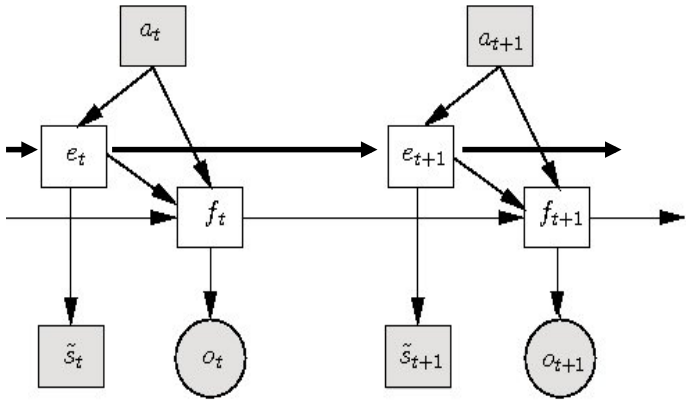
- **People are not alone in meetings !**
  interaction and social conventions provides context
  - we often share the same VFOA
  - when a person speaks, we tend to look at her/him
  - exceptions:
    - when a new slide is displayed, we tend to look at it
    - people look at their laptop (…) ; people are bored…

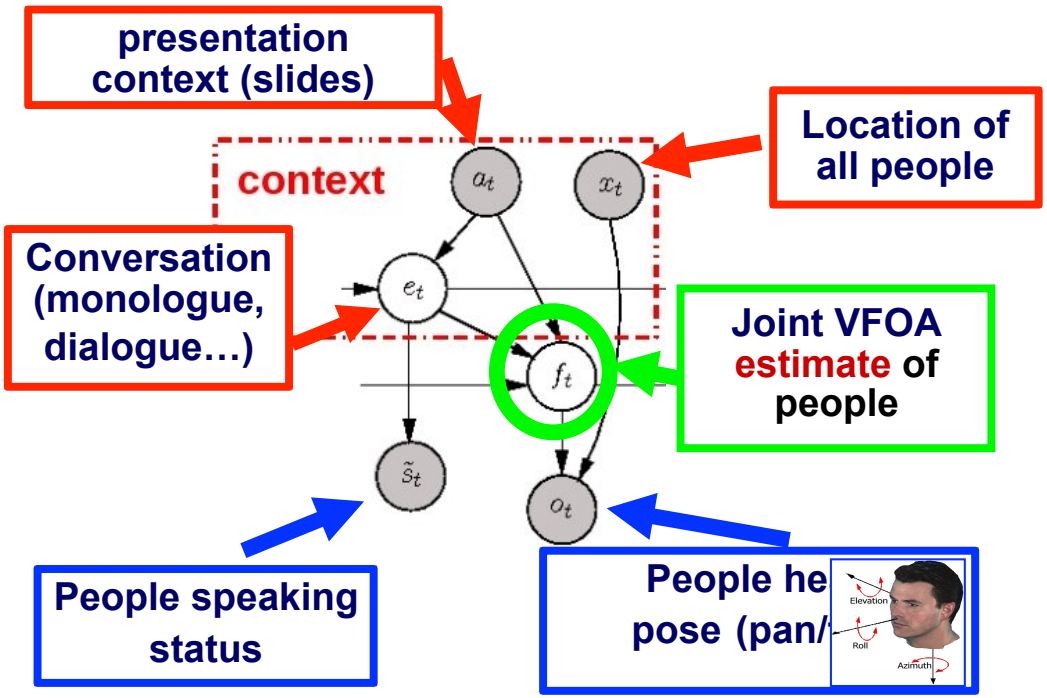
Goal: integrate this knowledge into a principled model

Multi-party Dynamic Bayesian Network Model



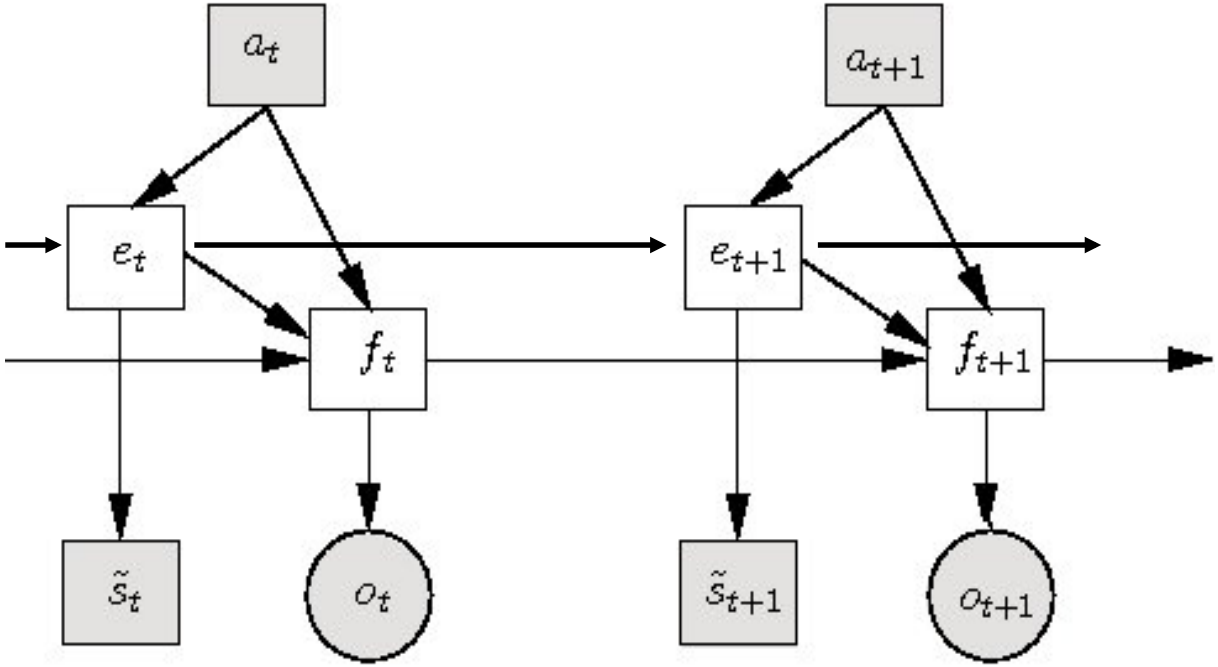




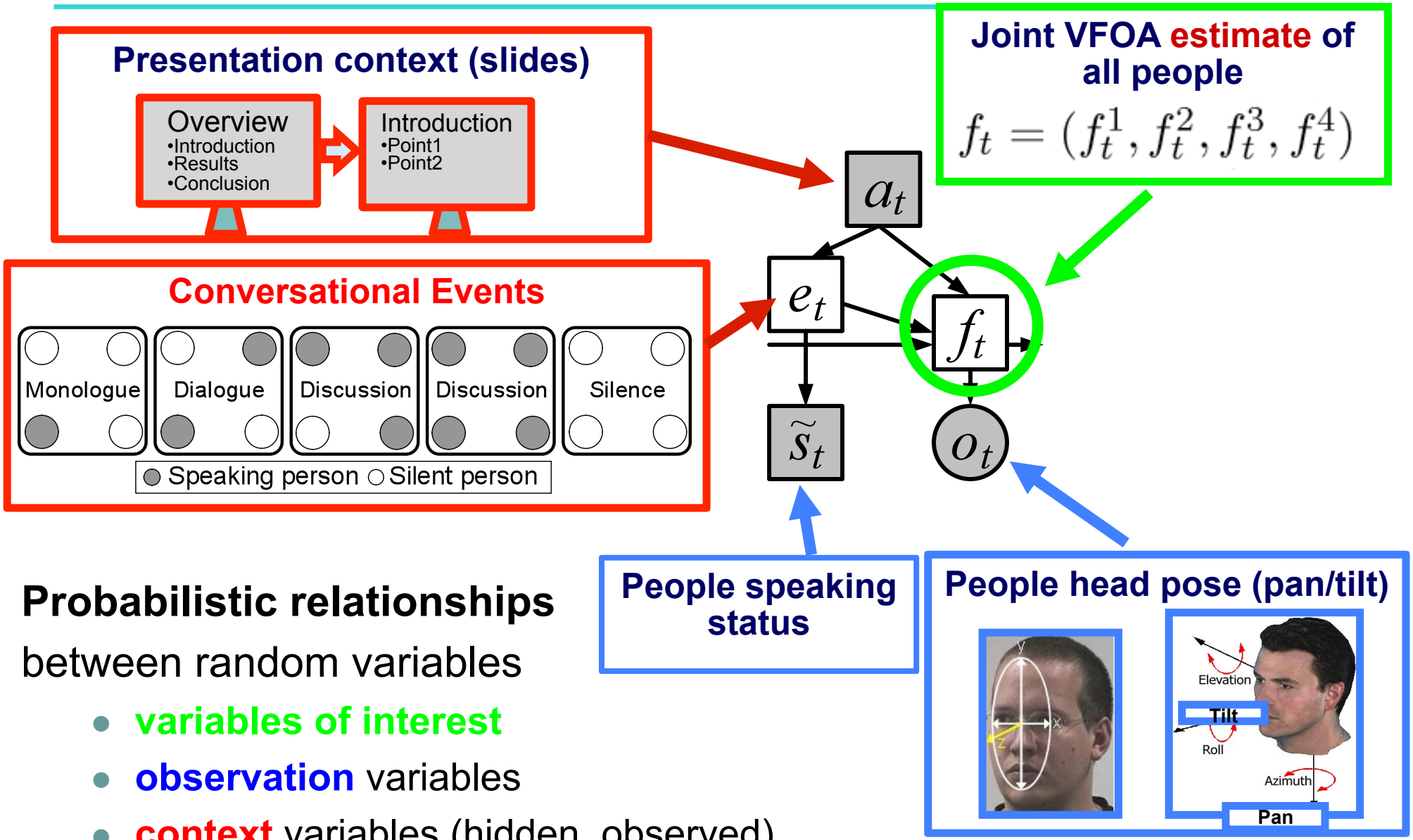


DBN: models the
probabilistic relationships
between random variables

Multi-party Dynamic Bayesian Network Model



Multi-party Dynamic Bayesian Network Model



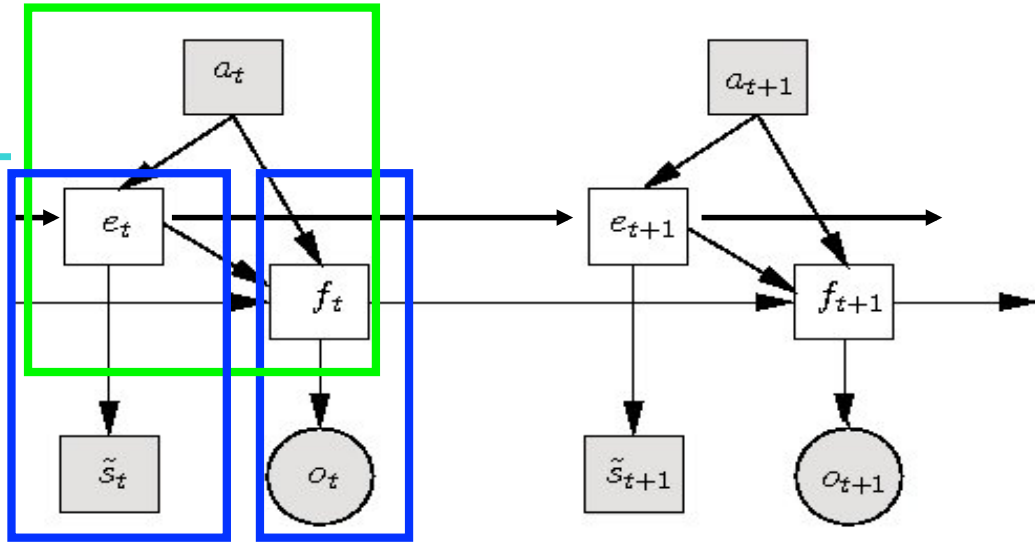
Probabilistic relationships

between random variables

- **variables of interest**
- **observation** variables
- **context** variables (hidden, observed)

Ba, Odobez, PAMI 2011

Multi-party DBN : details



Maximize posterior

$$p(f_{1:t}, e_{1:t}, \lambda | a_{1:t}, o_{1:t}, \tilde{s}_{1:t})$$

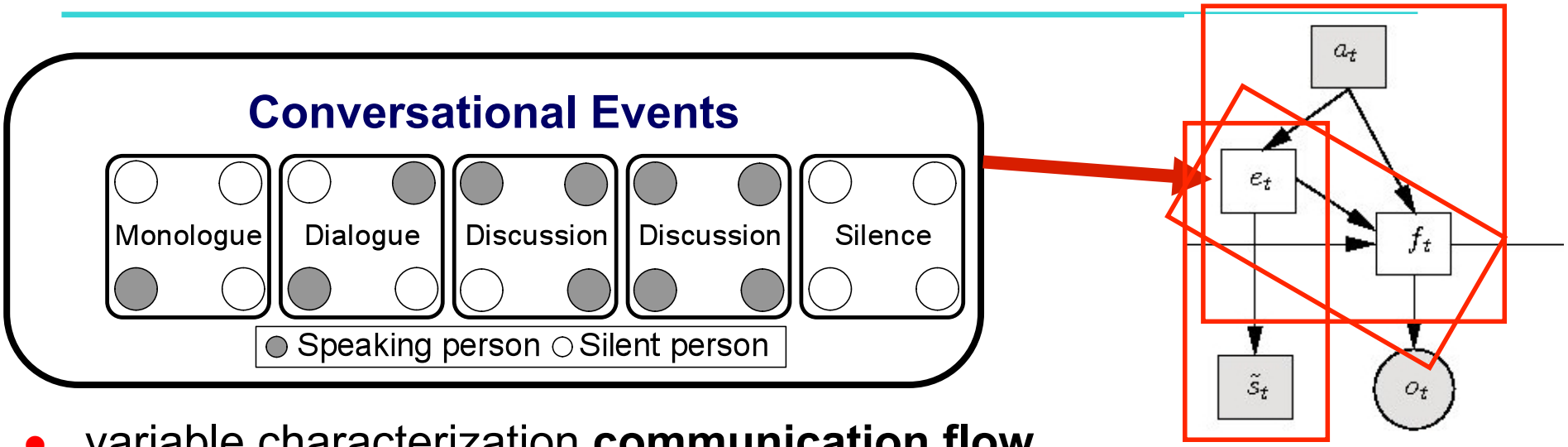
$$\propto p(\lambda) \prod_t p(o_t|f_t, \lambda) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t)$$

Prior distribution on model parameters

Observation model

Dynamical model

Interaction modeling : conversational events



- variable characterization **communication flow** specified by
  - event type (monologue, dialog,,..)
  - who is involved

- **assumptions**
  conversational event controls
  - **speaking** activity
  - dynamics of **gaze**
  - this control is modulated by the slide activity variable

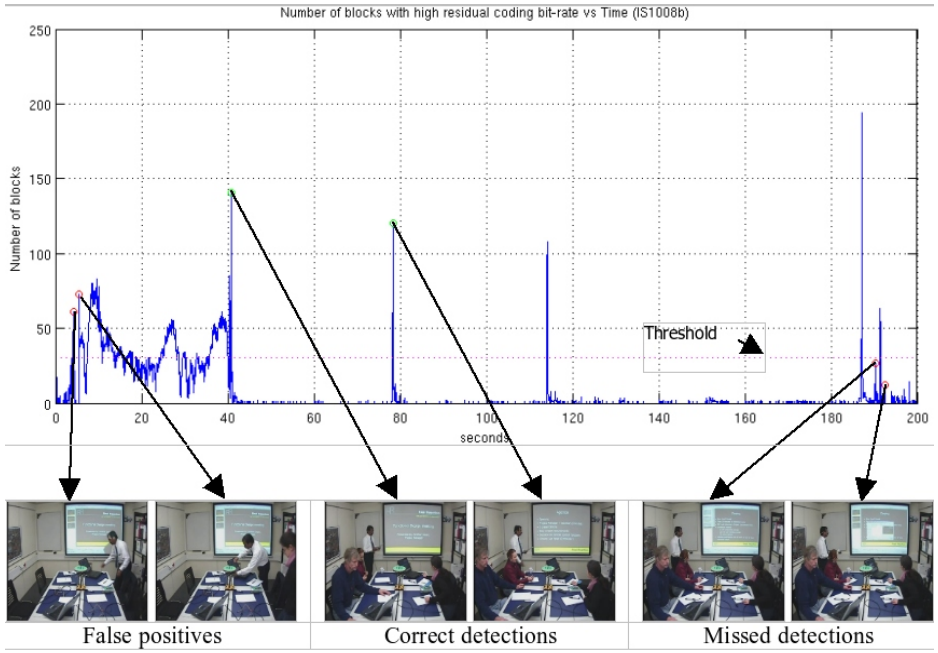
Contextual cue: presentation activity modeling a_t

- intuition: a new slide?
  - people turn their attention to it
  - then, attention **progressively** shifts back to the discussion

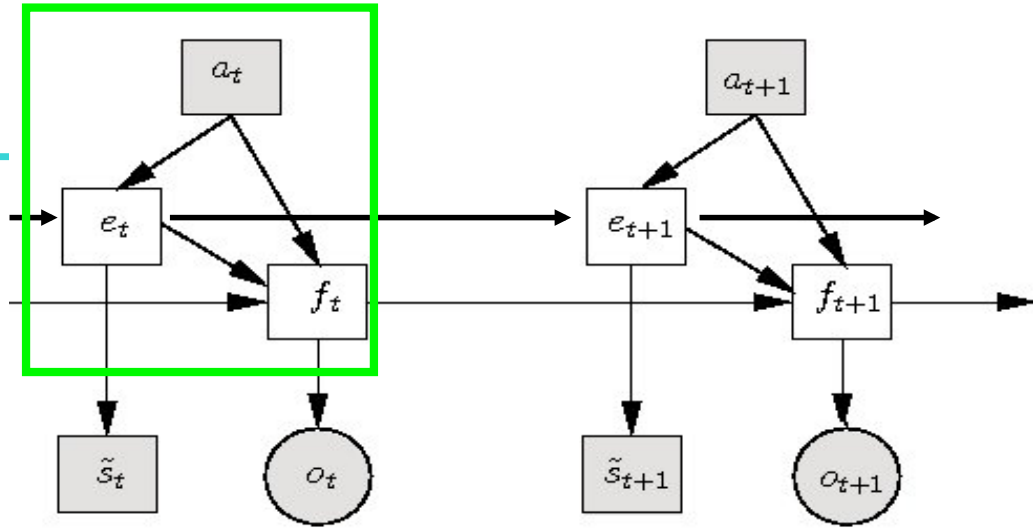
  => timing information is important

- presentation cue a_t **"Elapsed time since the last slide change"**

- automatic detection of slide changes from motion energy features



Multi-party DBN : details



Maximize posterior

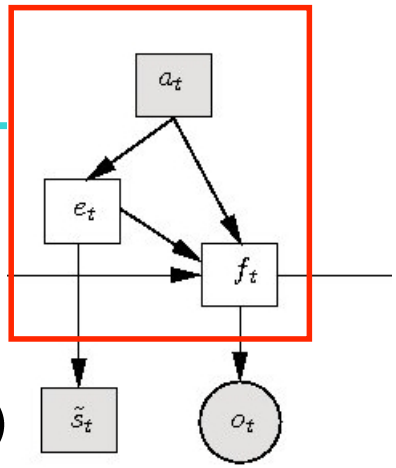
$$p(f_{1:t}, e_{1:t}, \lambda | a_{1:t}, o_{1:t}, \tilde{s}_{1:t})$$

$$\propto p(\lambda) \prod_t p(o_t|f_t, \lambda) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t)$$

Dynamical model

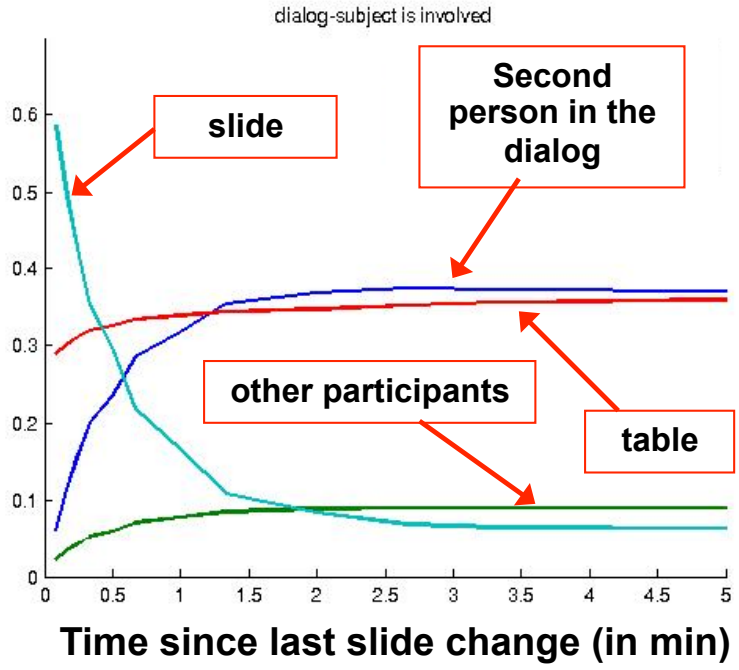
- Dynamical model $p(f_t|f_{t-1}, e_t, a_t) \propto \prod_{all\ person\ k} p(f_t^k|f_{t-1}^k) p(f_t^k|e_t, a_t)$
  - VFOA **temporal smoothness**
  - **Contextual prior on VFOA label**

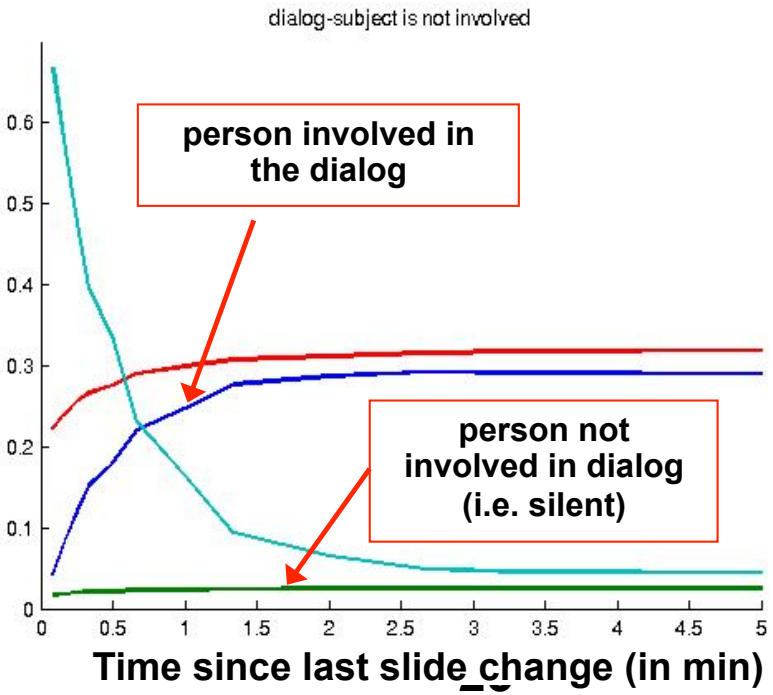
Context



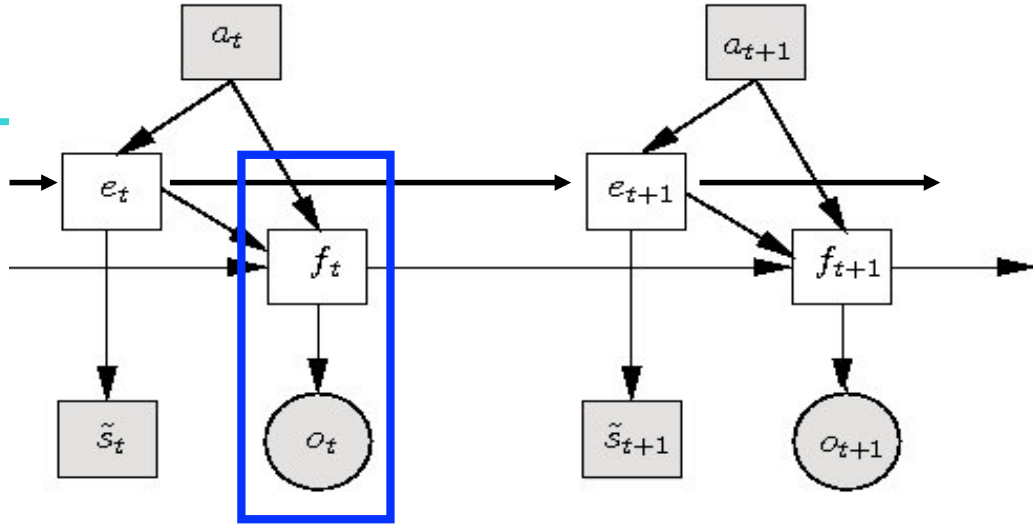
$$p(f_t^k|e_t, a_t) = p_{e_t}(f_t^k|a_t)$$

- **joint influence** of conversational event and slide activity on focus
- e.g. **dialog**
  learn prior probability of focus
  - person involved in the dialog
    - looks at slide when new slide displayed
    - after, looks mainly at dialog partner
    - looking at table important
  - person not involved
    - same focus behaviour w.r.t slide/table
    - looks almost exclusively at people involved in the dialog, not at the 4th participant





Multi-party DBN : details



Maximize posterior

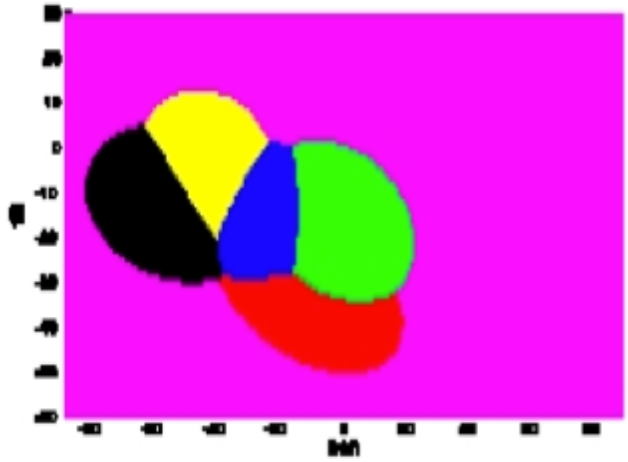
$$p(f_{1:t}, e_{1:t}, \lambda | a_{1:t}, o_{1:t}, \tilde{s}_{1:t})$$

$$\propto p(\lambda) \prod_t p(o_t|f_t, \lambda) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t)$$

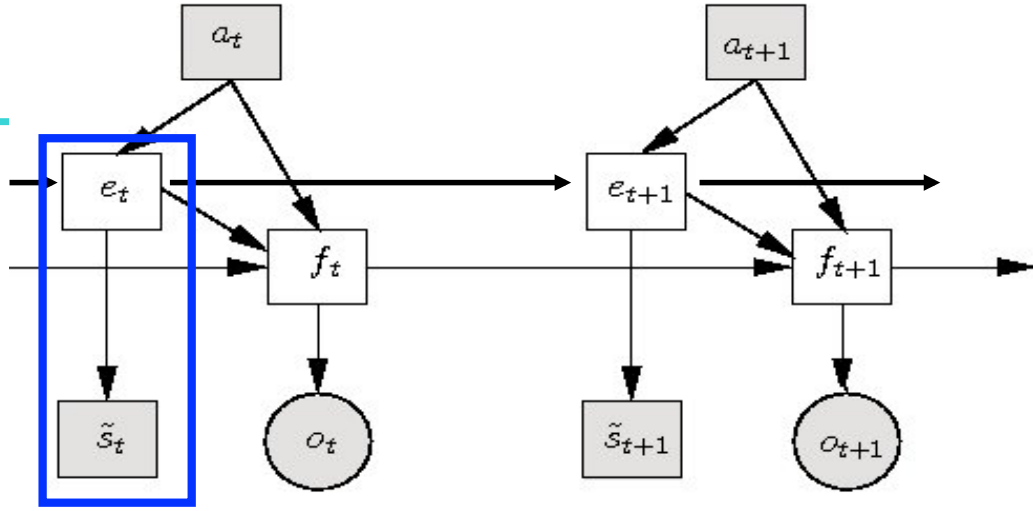
Observation model

$$p(o_t|f_t, \lambda) = \prod p(o_t^k|f_t^k, \lambda)$$

- head poses
  - same model as with the independent case



Multi-party DBN : details



Maximize posterior

$$p(f_{1:t}, e_{1:t}, \lambda | a_{1:t}, o_{1:t}, \tilde{s}_{1:t})$$

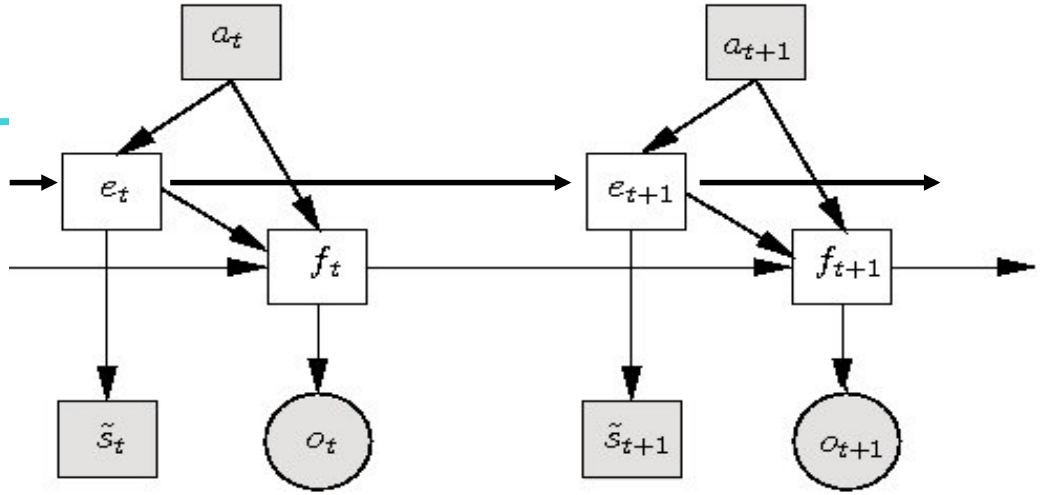
$$\propto p(\lambda) \prod_t p(o_t|f_t, \lambda) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t)$$

Observation model

- speaking status

  probability (high, low), depending on who is expected to speak given the conversational event

Bayesian inference



- Maximization of joint posterior distribution of hidden variables (including parameters) given observations

$$p(f_{1:t}, e_{1:t}, \lambda | a_{1:t}, o_{1:t}, \tilde{s}_{1:t})$$

- Inference more complex than with normal HMM
  - several interdependent hidden variables
  - however, we can exploit hierarchical structure

  - estimate the event assuming known VFOA
  - estimate VFOA and parameters assuming known events

Multimodal multiparty VFOA recognition

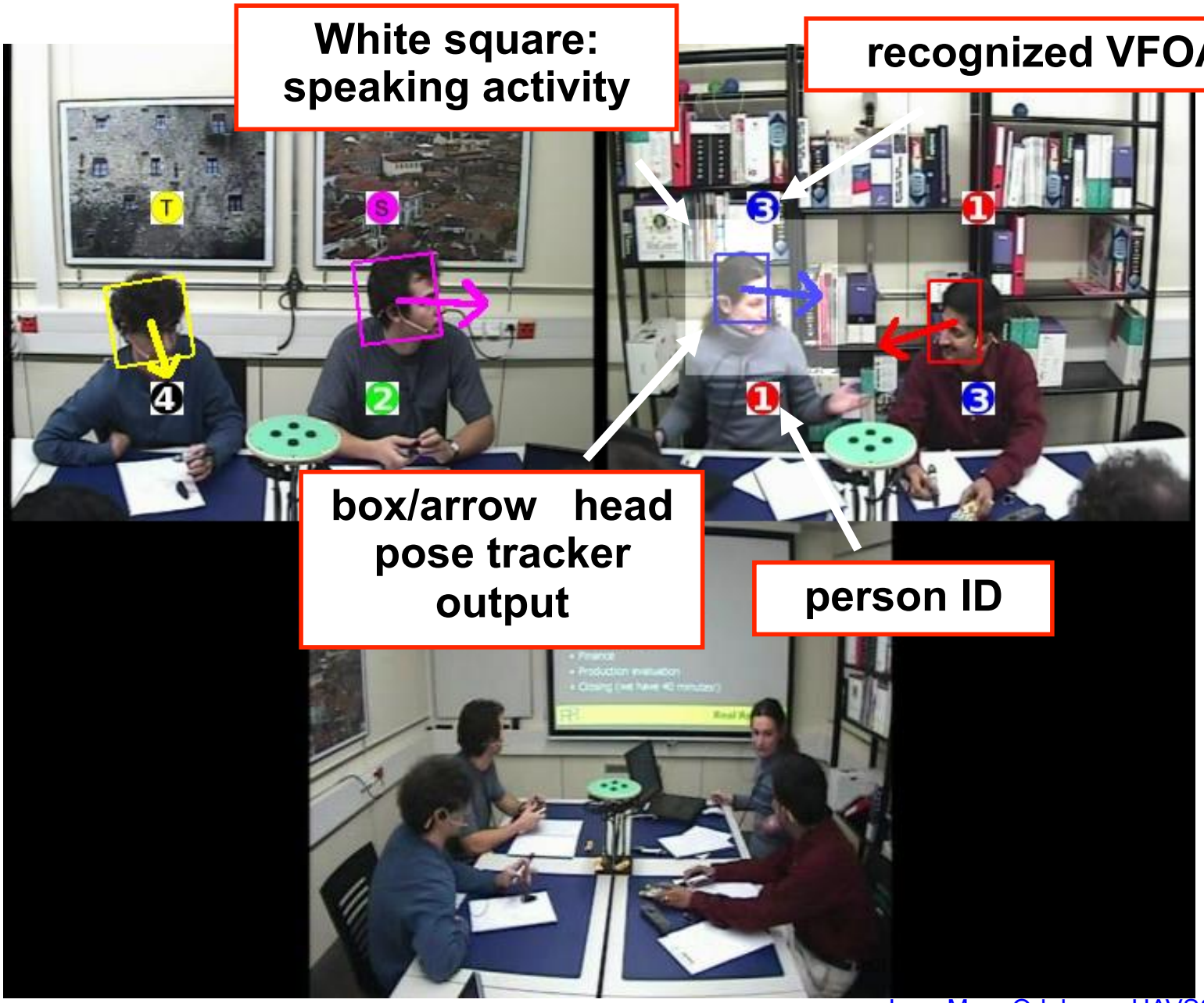
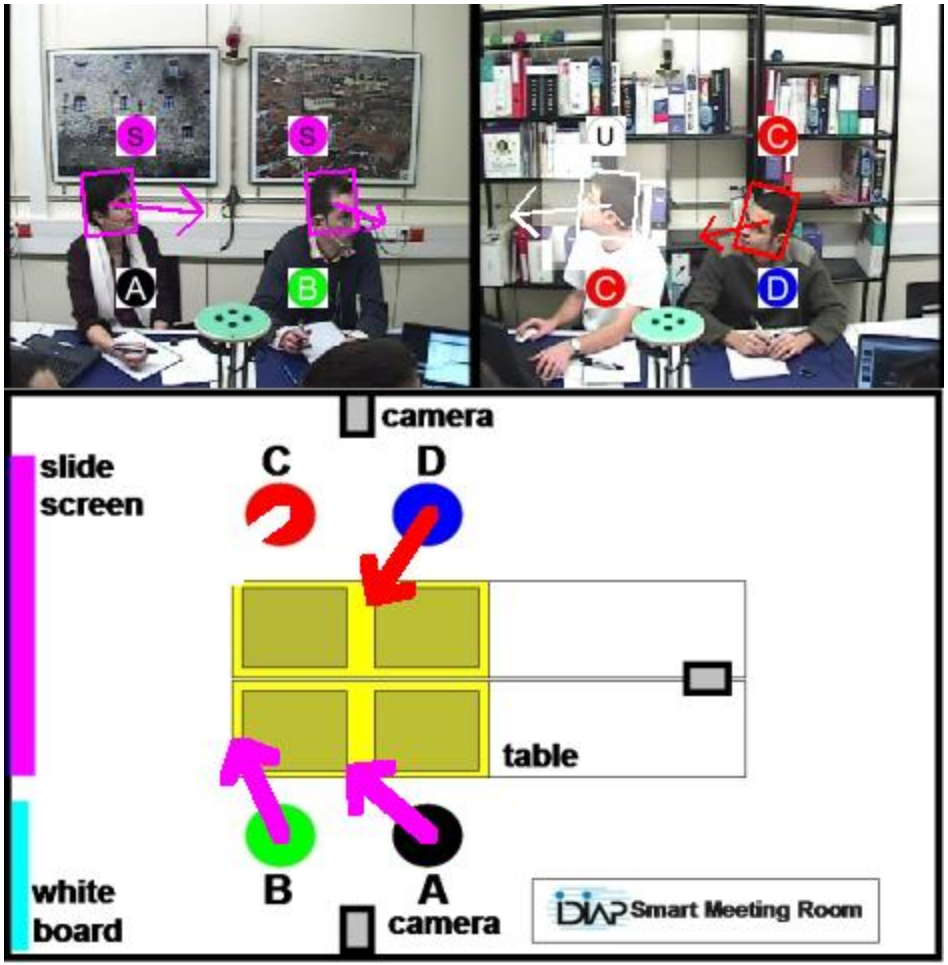
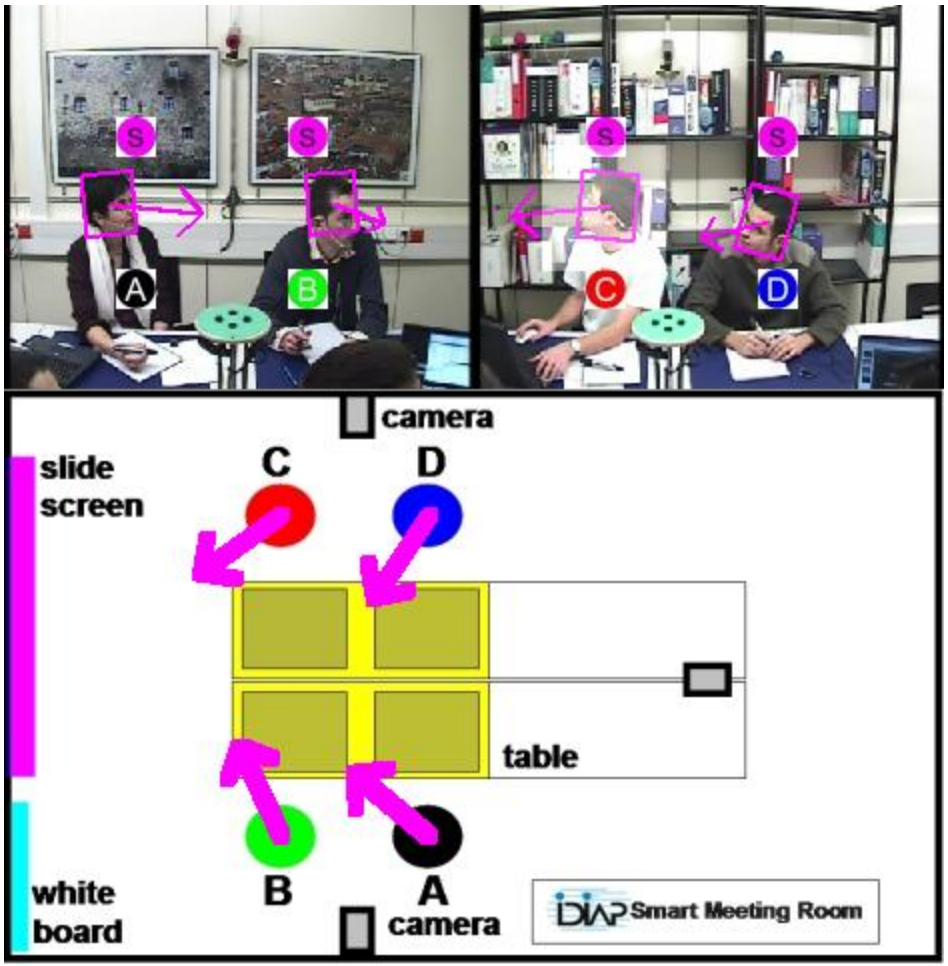


Illustration: group and slide activity

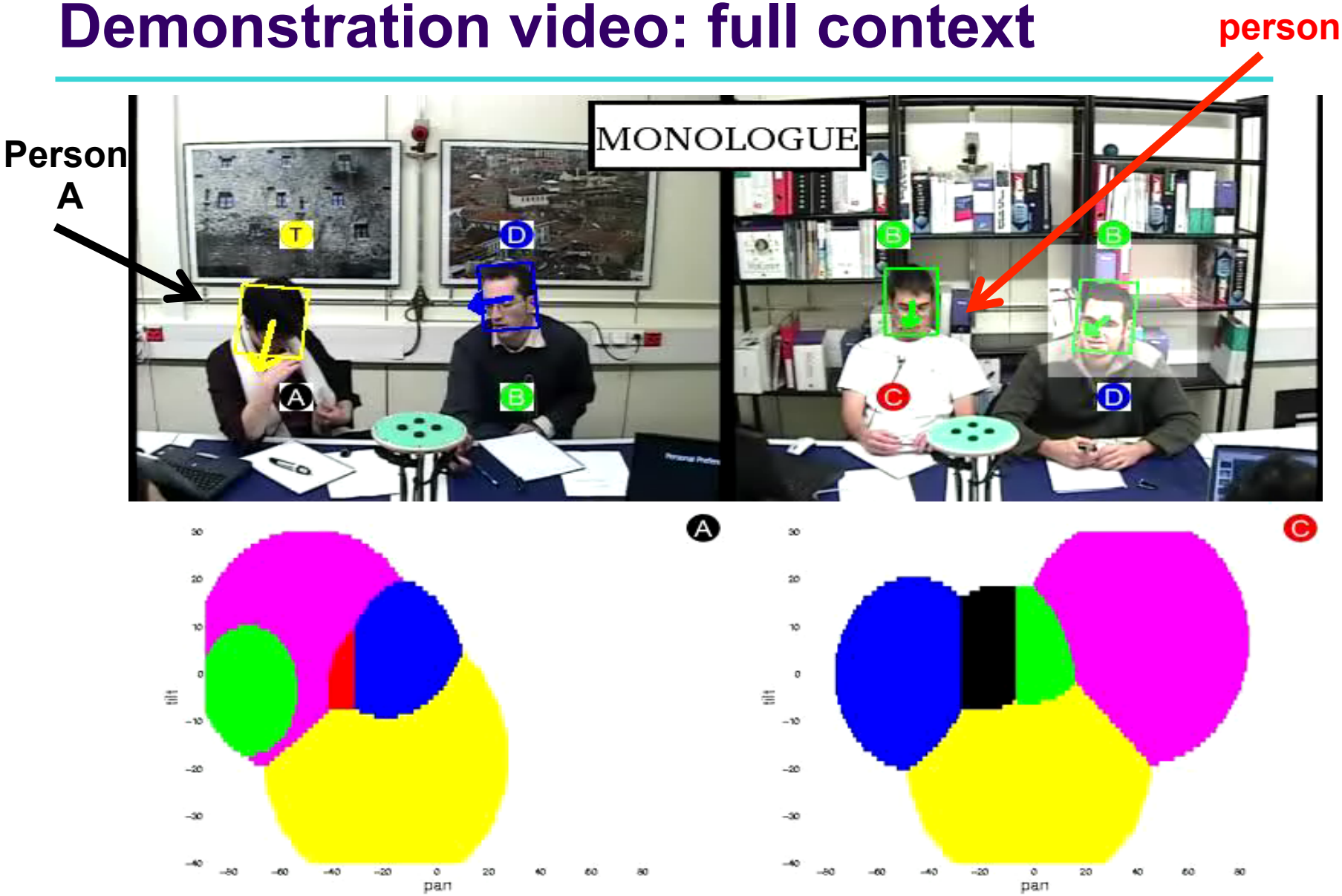


independent recognition
(head pose only)



multi-party recognition
using contextual cues

Demonstration video: full context



VFOA decision map of person A

VFOA decision map of person C

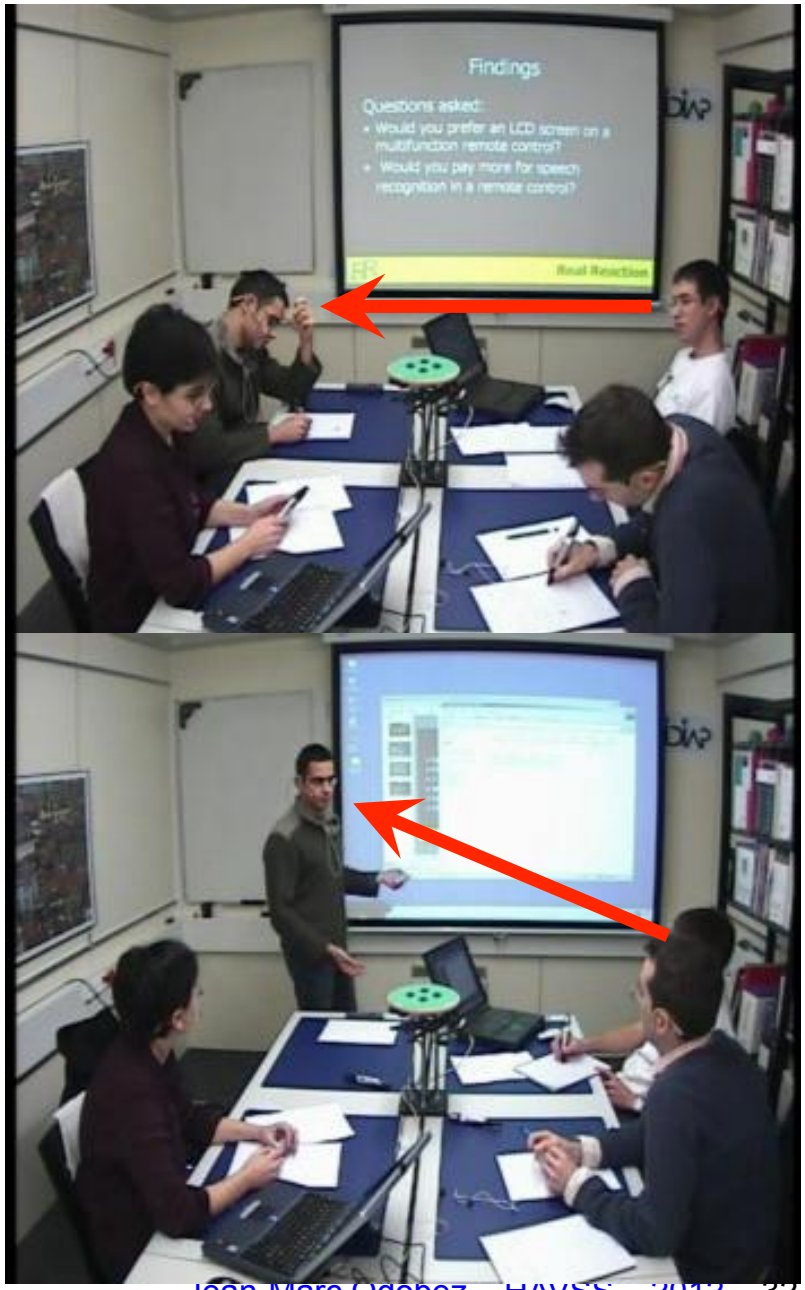
Lowest row: influence of conversation context on VFOA decision maps

handling moving people

- moving people
  standing for presentation

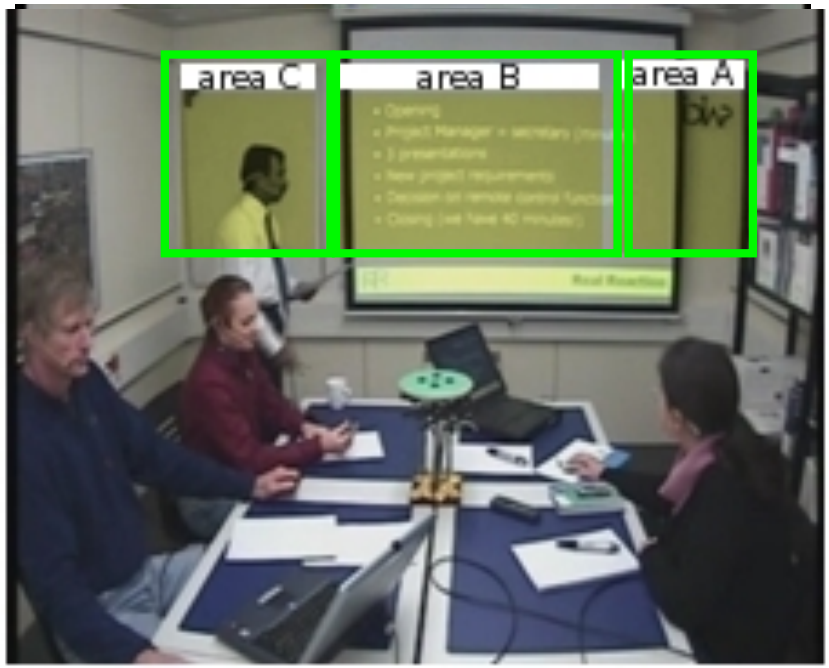
- direct impact on gaze
  - same semantic target
    different gaze directions

- two main implications
  - track people when they leave their seat
  - adapt gaze model of sitting participants dynamically



Contextual cue: people location x_t

- track people in 3D space
  - + precise location
  - -- precise camera calibration
  - -- difficult
  - -- precision might not be exploitable by gaze model

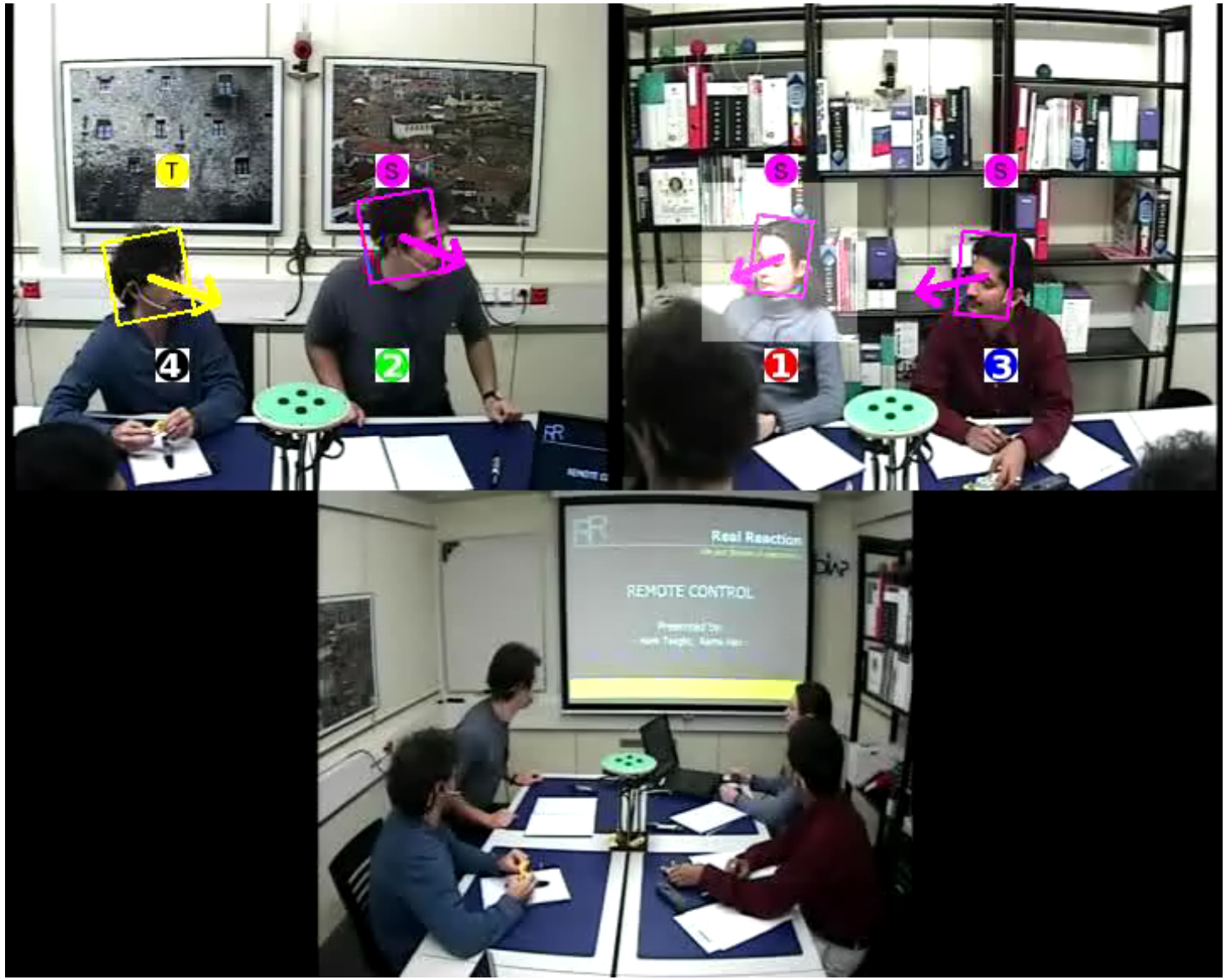


- alternative : use discrete locations
  - for person k
    - seat k
    - center of one of the 3 presentation areas A/B/C
  - tracking
    - side camera when people are seated (cf head pose tracking)
    - central camera: maximum of motion energy features in area A, B, C

Multimodal multiparty VFOA recognition

Notice

- Liveliness, difficulty of data
- person 3 focus changes according to context
  (between looking at person1, slide, standing person)
- slide changes favor looking at slides
- person 4 erroneous VFOA estimation (mainly due to head pose estimation problems)

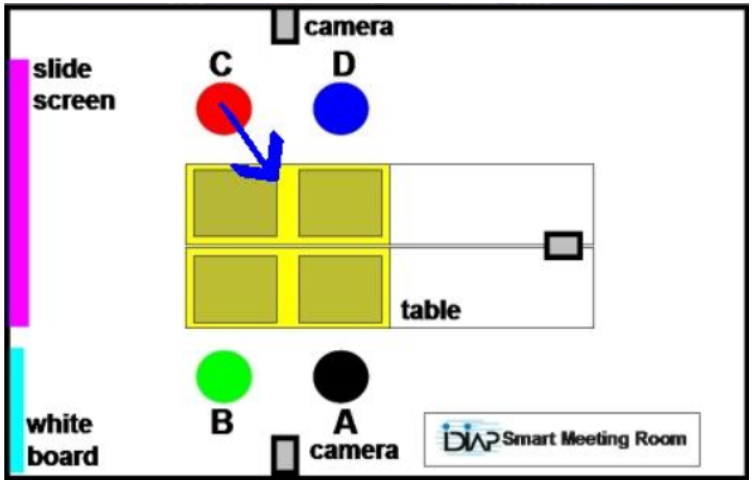


Results

- 12 full meetings: 5 hours of data
- performance measure : percentage of correctly recognized VFOA

position	A	B	C	D	**mean**
Baseline (head pose only), **independent**	36.4	45.5	41.3	29.7	**38.2**
Multi-party, **full context**	56.5	56.2	62.3	46.2	**55.3**

- baseline: 38.2% => challenging problem



- seats A and D: more VFOA ambiguities

- multi party
  - context helps: **+ 17%** absolute improvement
  - higher improvement on seats with larger ambiguities

Presentation plan

- VFOA analysis in groups
  - Head estimation accuracy
  - VFOA modeling
    - Task
    - Contextual multi-party DBN VFOA recognition
    - Remarks

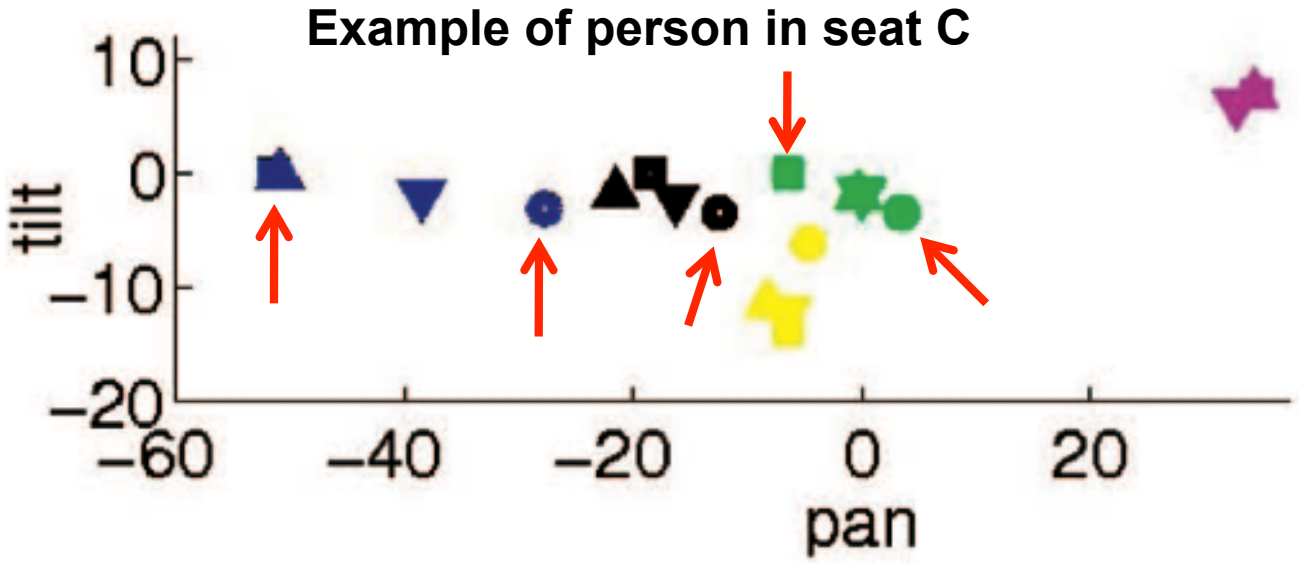
      parameter adaptation

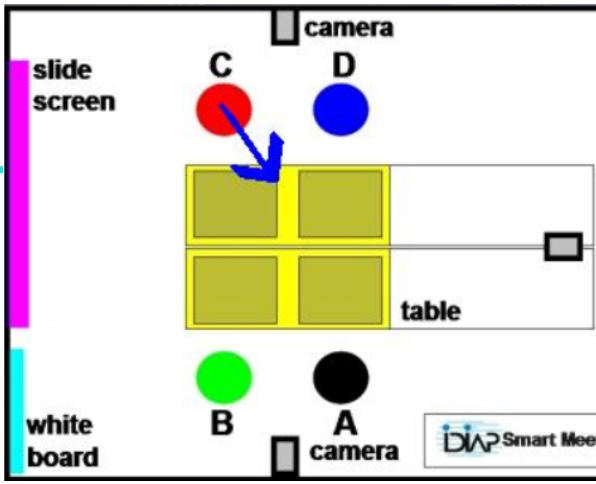
      head pose estimation

      gesturing

- Head and body pose extraction in surveillance scenarios

Contextual parameter adaptation (1)





□ prediction

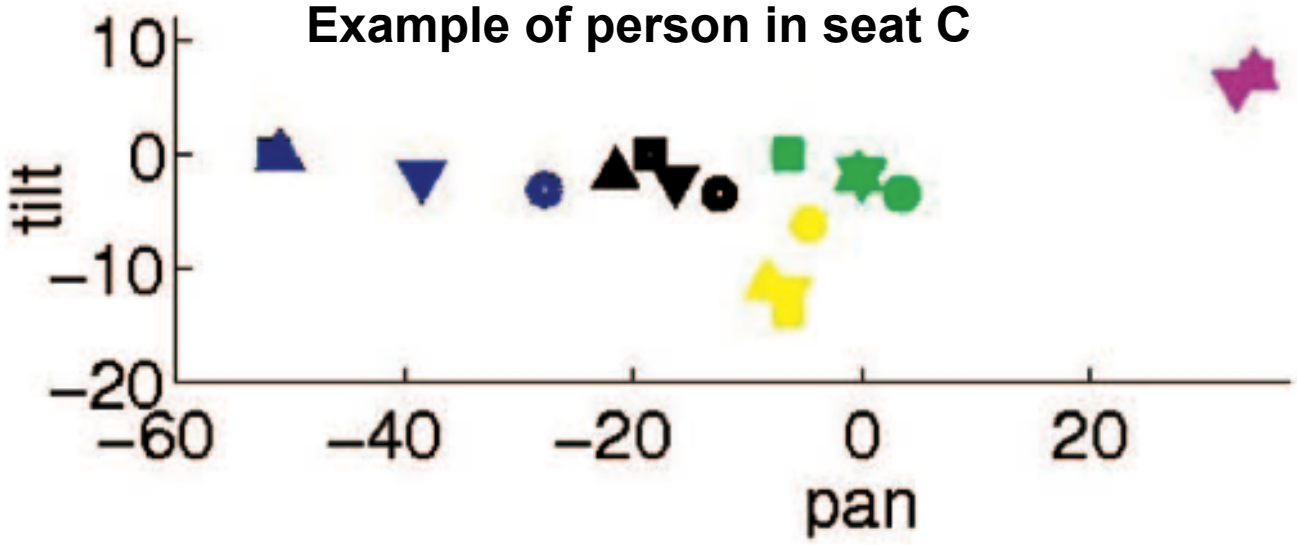
● empirical mean

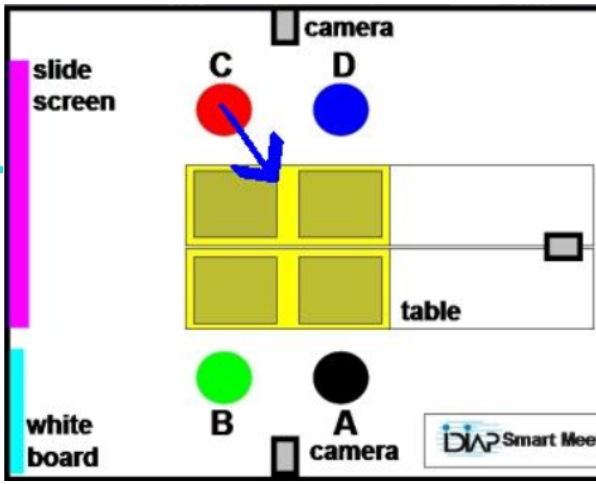
△ no-context

▽ context

- Head pose parameters predicted by cognitive model are crude
  - Biased head pose estimated, different individual behaviors
- MAP Bayesian adaptation mean for target i : combination of
  - prior value (prediction model)
  - average head poses assigned to target i

Contextual parameter adaptation





□ prediction

● empirical mean

△ no-context

▽ context

- Head pose parameters predicted by cognitive model are crude
  - Biased head pose estimated, different individual behaviors

  => requires unsupervised adaptation

- Bayesian adaptation

  mean for target i : combination of
  - prior
  - average head poses assigned to target i

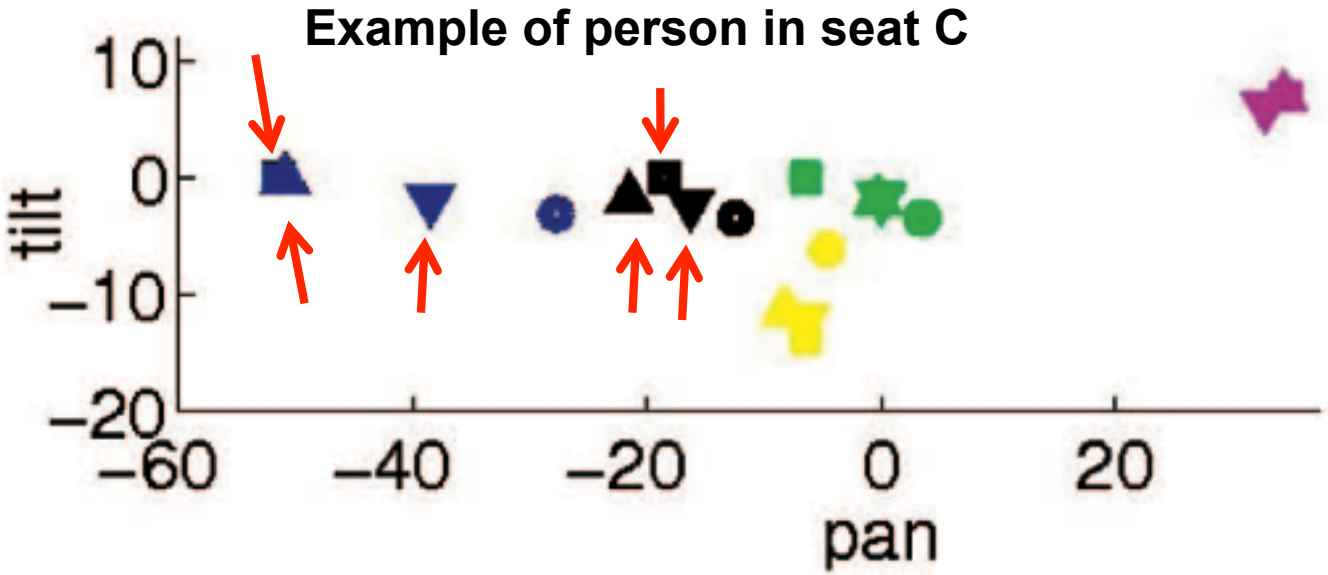
Assignement to target i

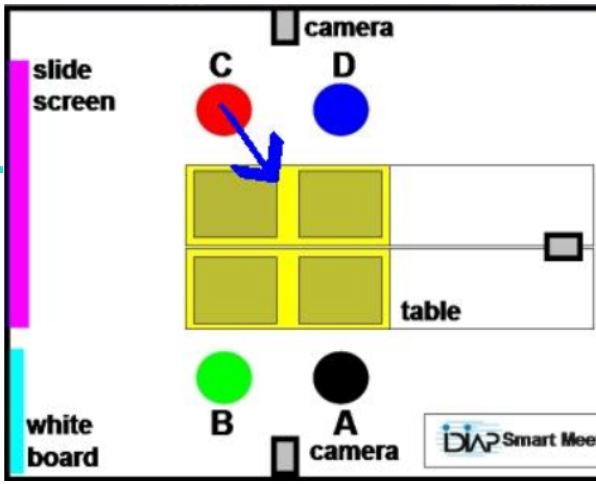
prior

$$\mu_{k,i} = \frac{\tau m_{k,i} + \sum_{t=1}^{T} \gamma_{i,t}^{k} o_t^k}{\tau + \sum_{t=1}^{T} \gamma_{i,t}^{k}}$$

Head pose at time t

Contextual parameter adaptation (2)





□ prediction

● empirical mean

△ no-contex

▽ contex

- Probabilistic assignement
  - No-context

$$\gamma_{i,t}^k = p(f_t^k = i | o_{1:T}^k, \hat{\lambda}_k)$$

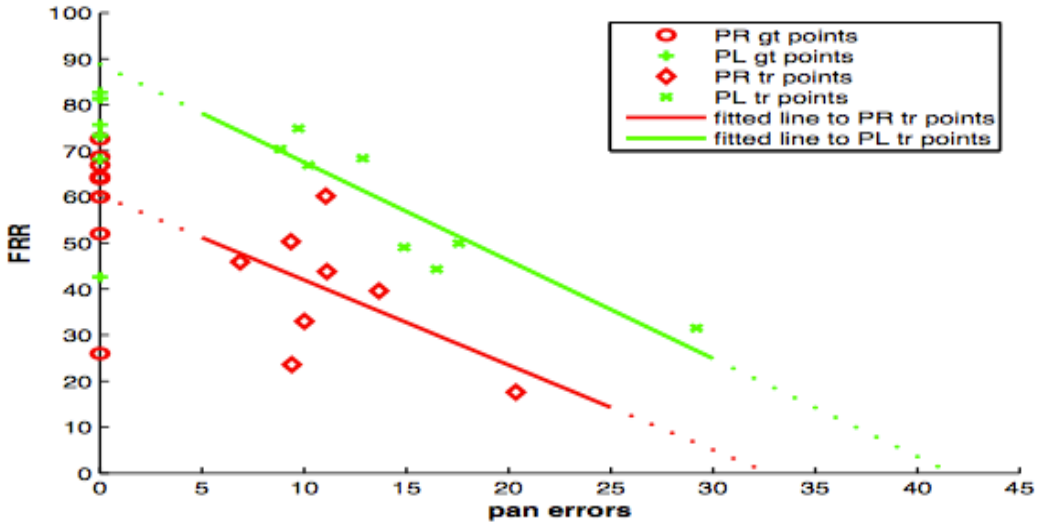
  - Context

$$\gamma_{i,t}^k = p(f_t^k = i | o_{1:T}^k, \hat{e}_{1:T}, a_{1:T}, x_{1:T}, \hat{\lambda}_k)$$

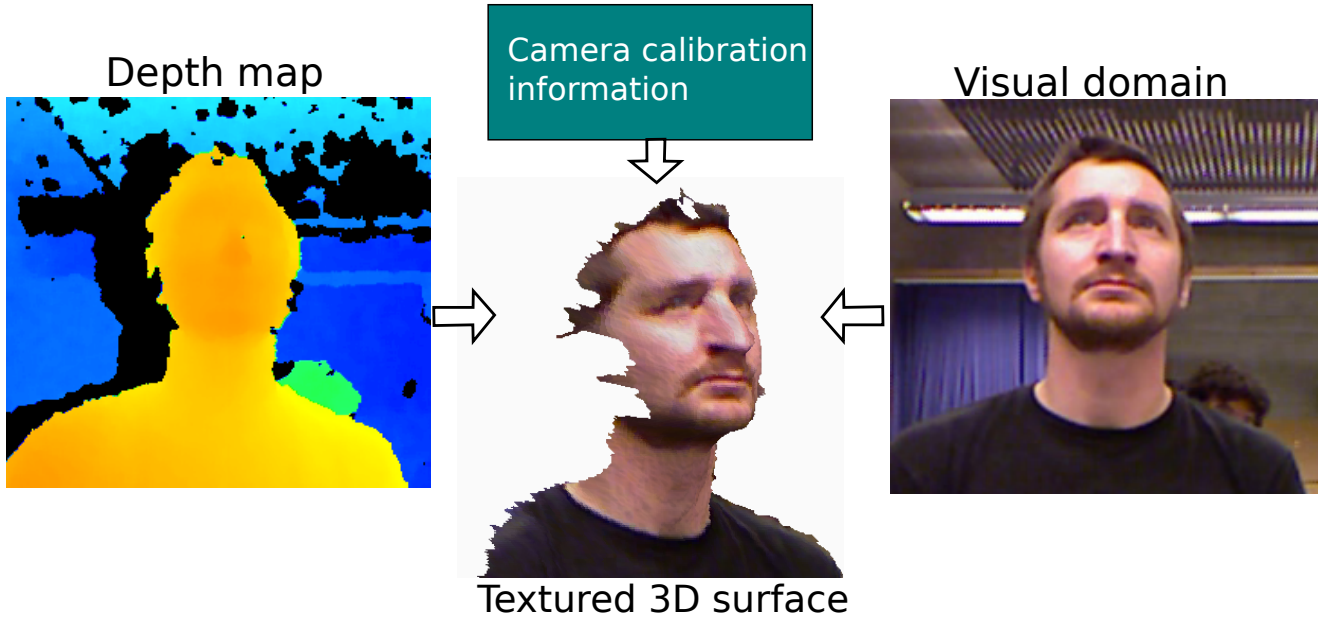
    - Increases reliability of right pose to target assignement
    - more accurate parameter estimates
- Get 5% increase in recognition rate

Head pose estimation

- Better head pose, better results !
  - measured in our dataset



- Gaze tracking – strategy: use RGB

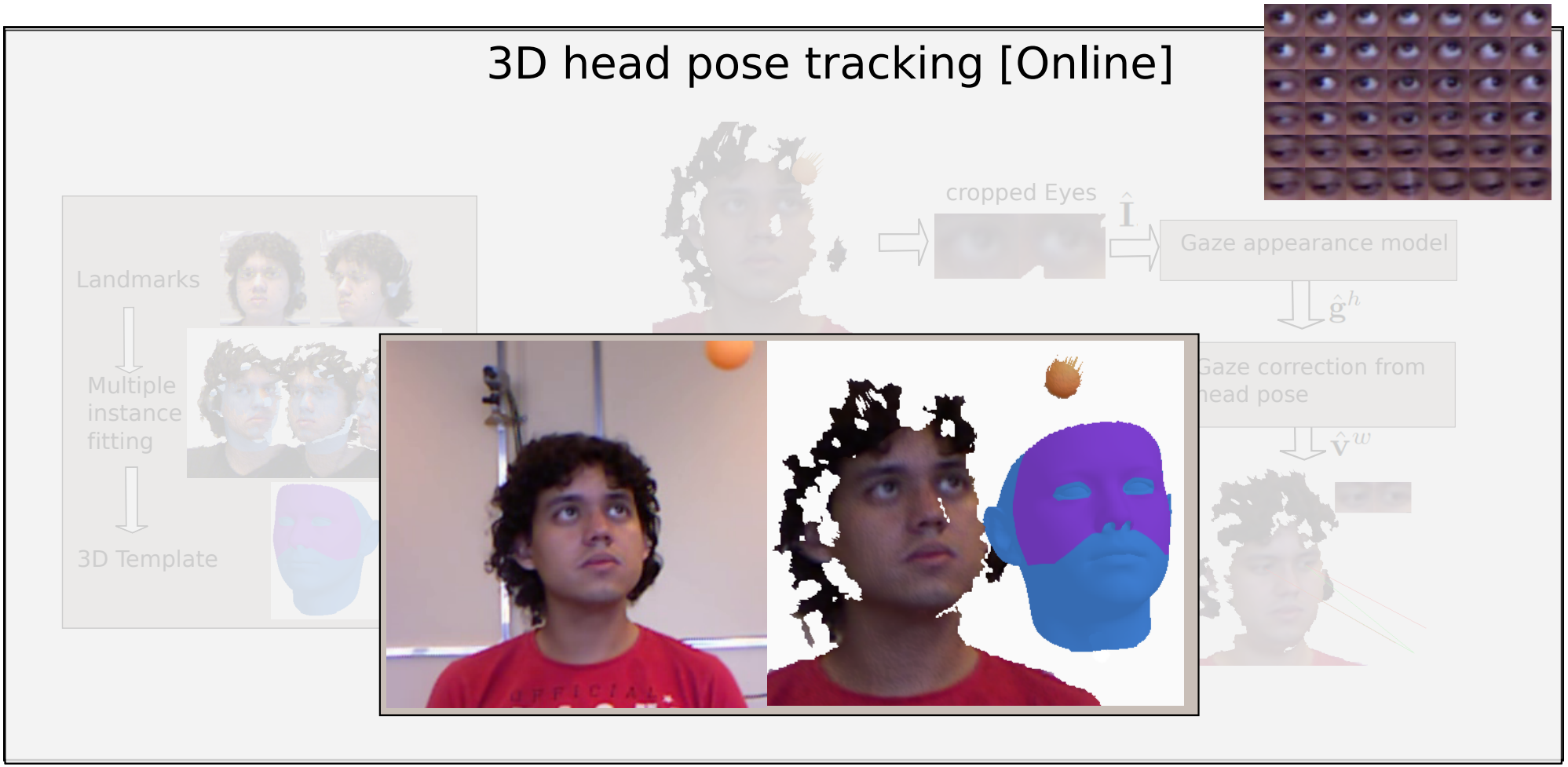


Multimodal information

- Depth: strong cue for head pose estimation
- Vision: necessary for eye appearance

Funes & Odobez, Gesture workshop, CVPR 2012

Gaze tracking: method overview



Results

- Accuracy between 5 and 10 degreees, depending on constraints (frontal vs person free to move head)

Modeling conversation context with visual activity (1)

Audio unavailable

=> conversation context ?

- visual activity encodes body langage ?

- common sense + studies
  - speaking is accompanied by visual activity
    - face/head, e.g
      mouth/speaking; rotation/addressing….
    - hand gestures
      rhythm (beat), deictic gestures (pointing)
  - visual activity when not speaking
    - head (focus change, backchannel)
    - hands (fidgeting, rubbing the chin, taking notes…)

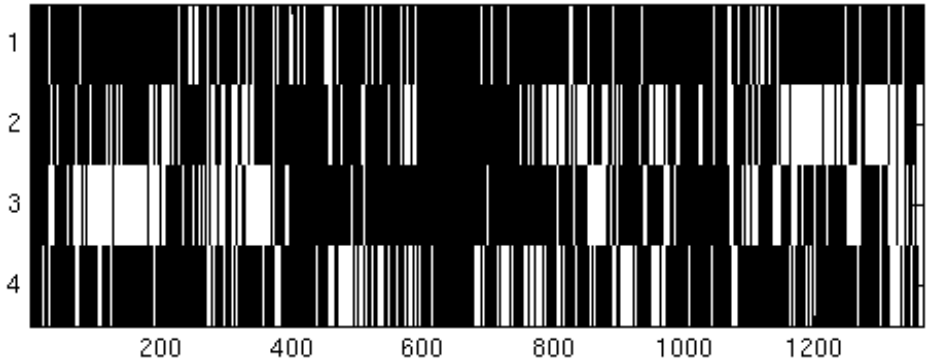
Ba, Hung, Odobez, ICME 2009

Modeling conversation context with visual activity (2)

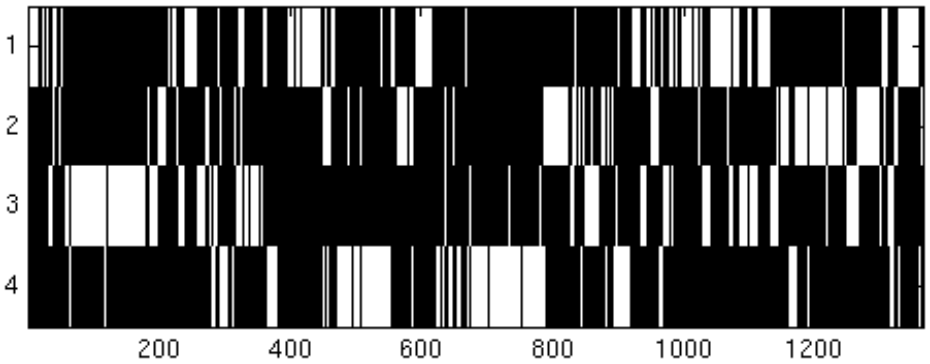
- Chances of speaking
  - average 25%
  - If visually active 47%
- Chances of being visually active
  - average 35%
  - If speaking 66%

=> correlation between speaking and visual activities

Speaking activity – 4 people



visual activity – 4 people



(visual activity obtained by thresholding motion energy)

speaking activity ?

visual activity ?

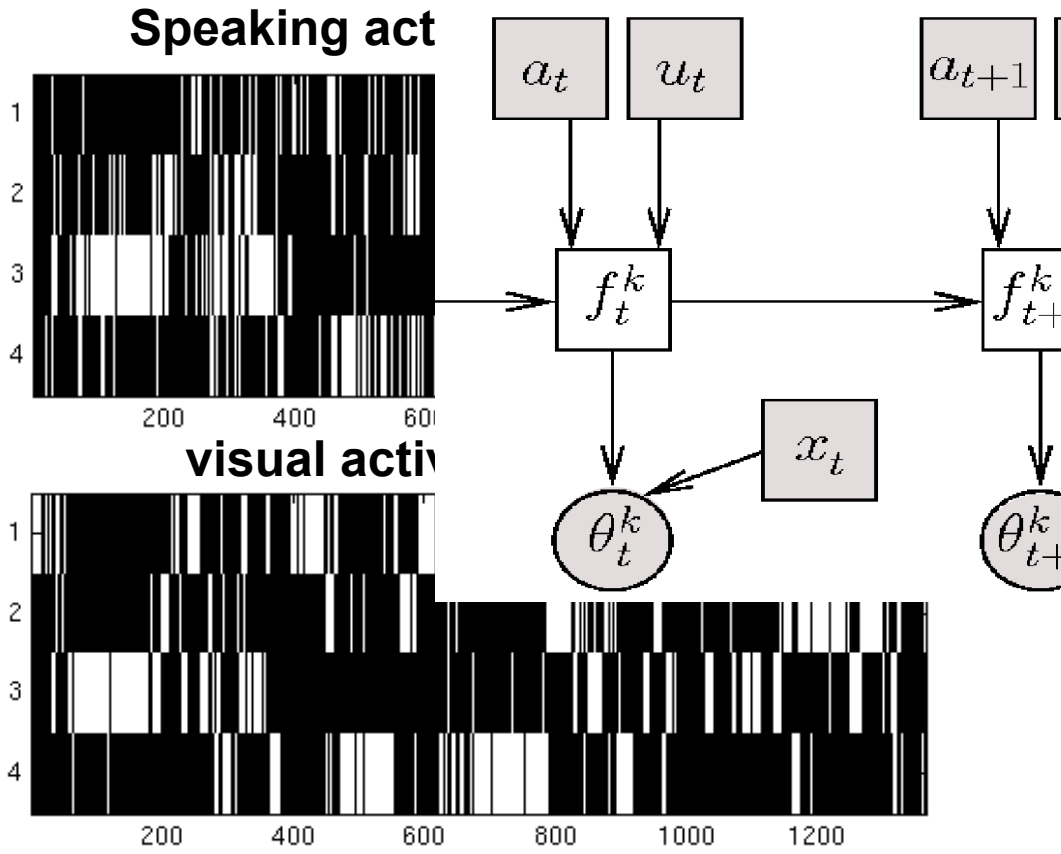
Visual attention

Ba, Hung, Odobez, ICME 2009

Modeling conversation context with visual activity (3)

- Chances of speaking
  - average 25%
  - If visually active 47%
- Chances of being visually active
  - average 35%
  - If speaking 66%

=> correlation between speaking and visual activities

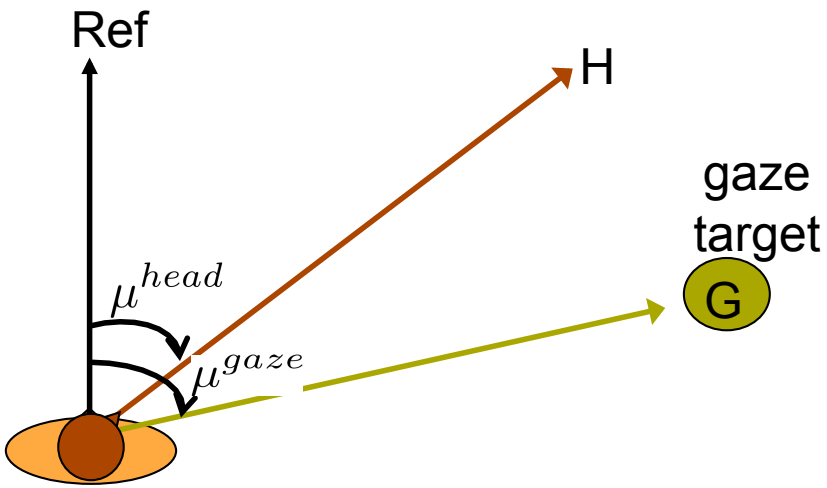


- Experiments
  - Replacing speaking status by visual activity status in model
  - Results: 53.2% (visual) vs 52.7% (speaking)

  => Visual activity as effective as speech to improve VFOA

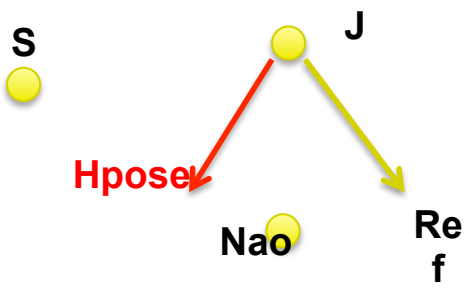
Ba, Hung, Odobez, ICME 2009

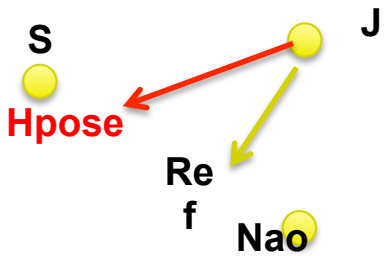
About the gaze mapping



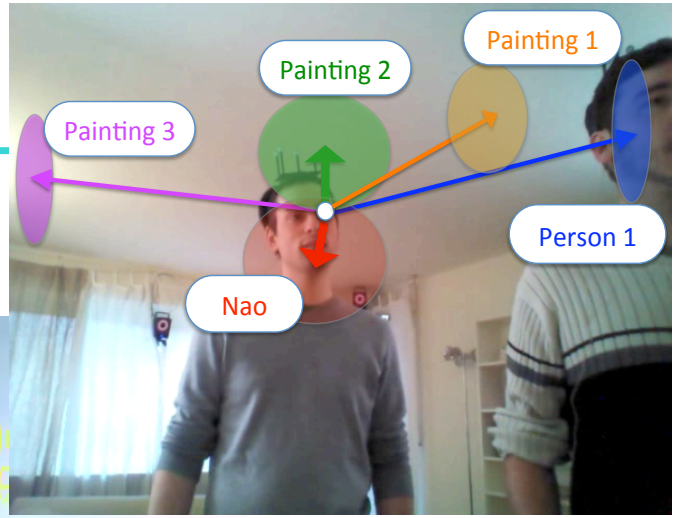
- What is the reference ?

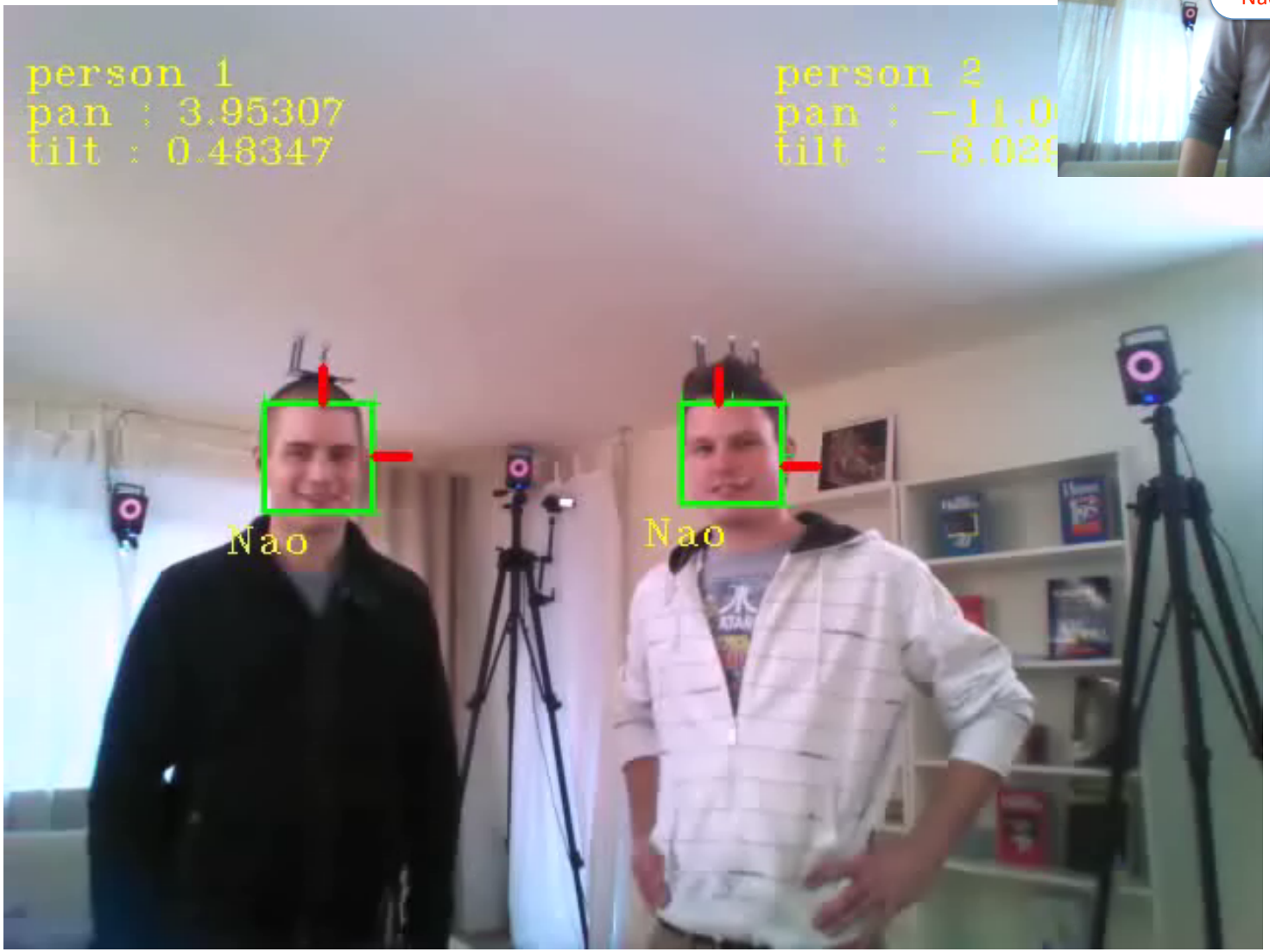
About the gaze mapping





About the gaze mapping





Presentation plan

- VFOA analysis in groups
  - Head estimation accuracy
  - VFOA modeling

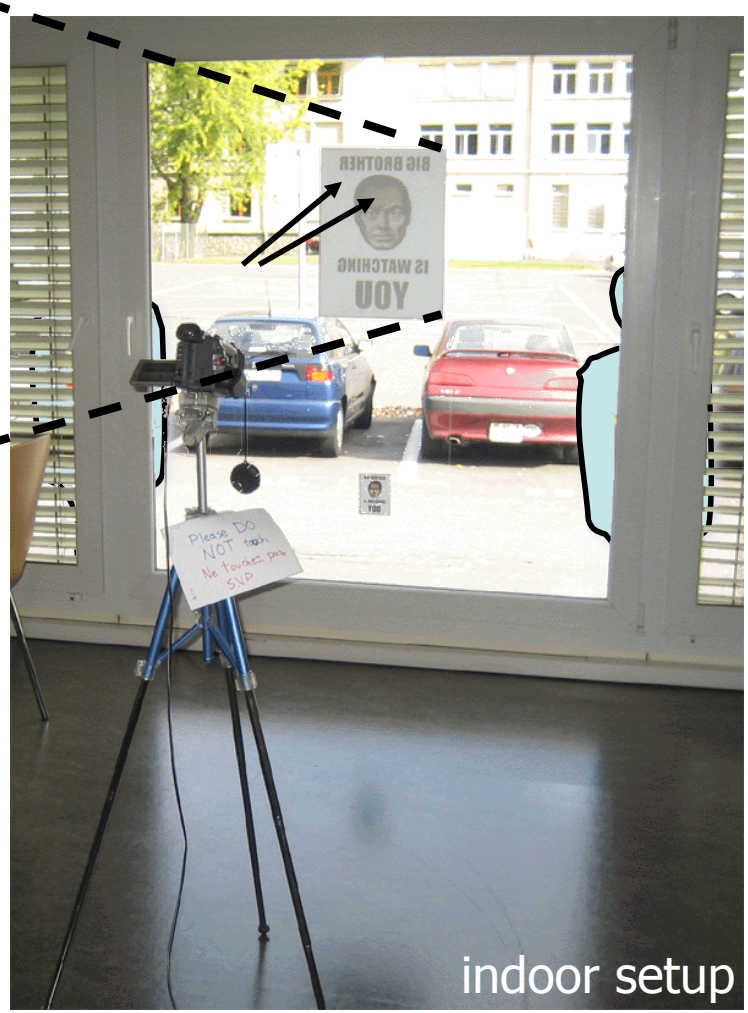
- Head and body pose extraction in surveillance scenarios

Wandering Focus of Attention of people
[Smith et al, PAMI 2008]

- Automatically determine:
  - # people exposed to the ad
  - # people viewed the ad
  - demographics (gender/age)

⇒ Multiple object tracking

⇒ One single focus **but person is moving**
we need to model this

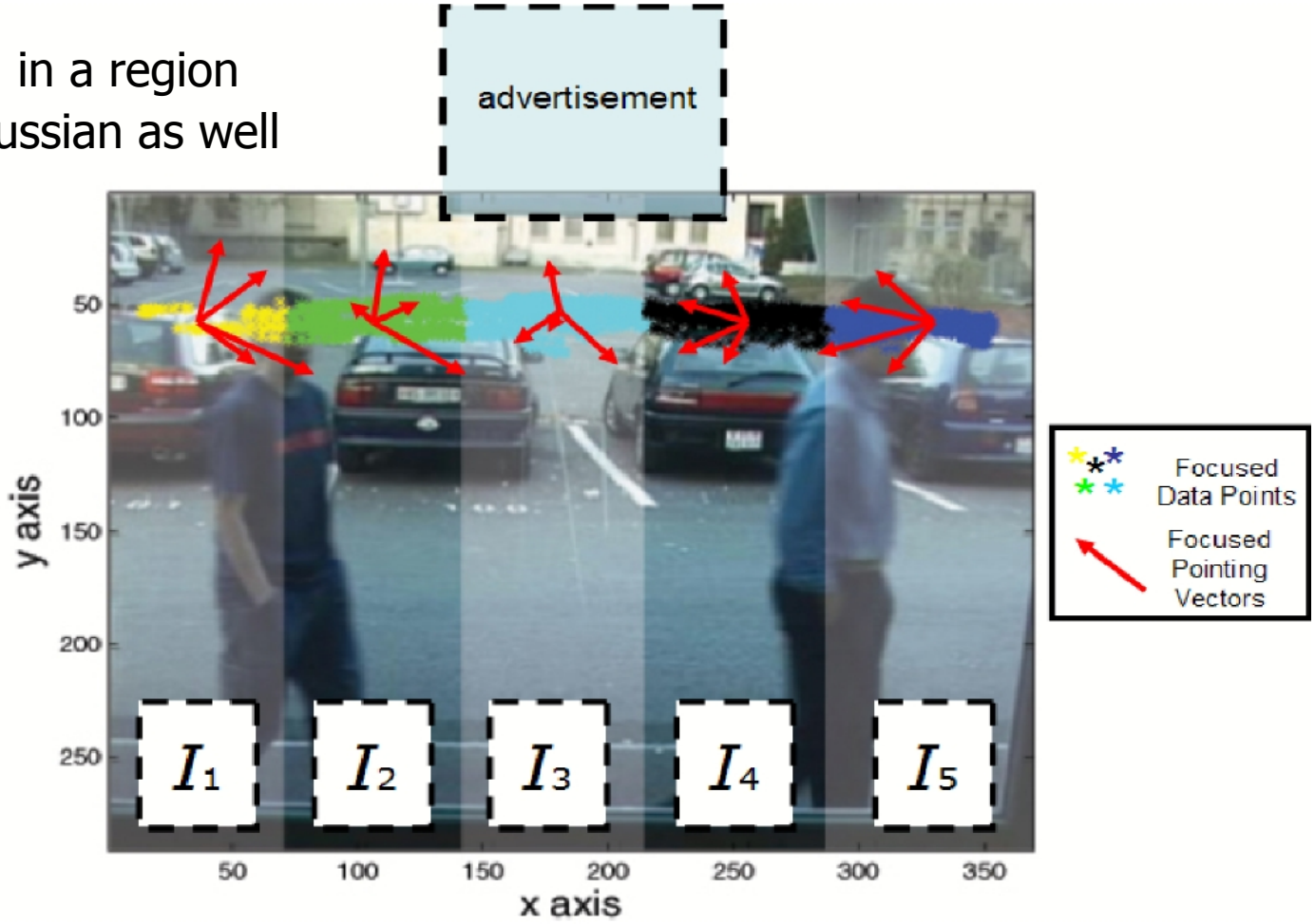


Wandering Focus of Attention of people
[Smith et al, PAMI 2008]

- Regression based approach: x = horizontal position

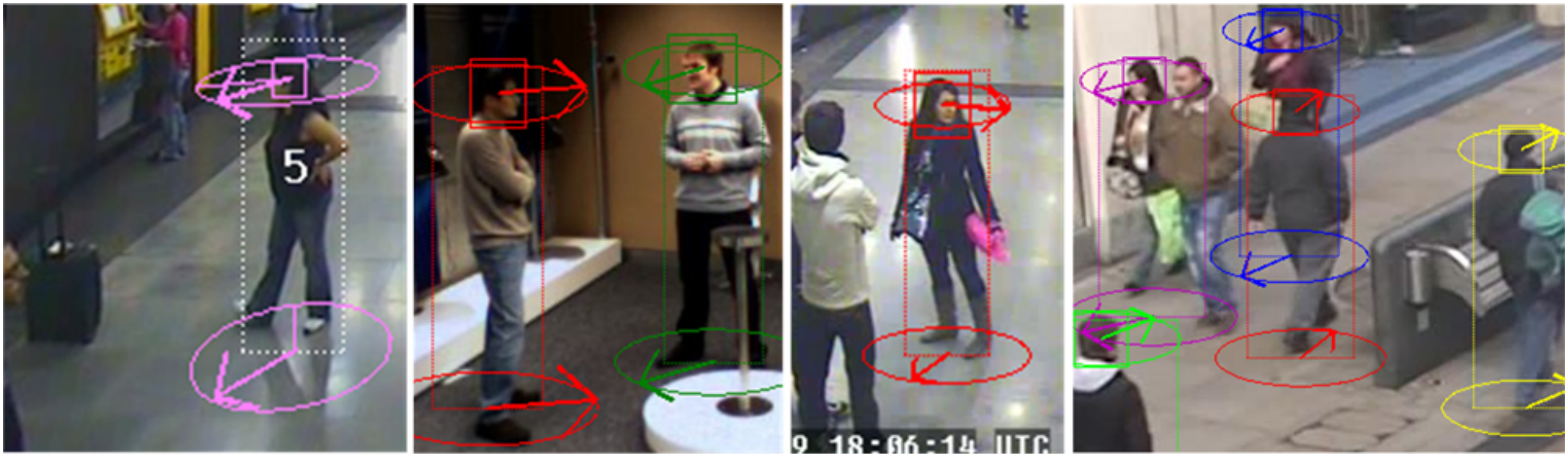
$$p(h|focus = 1) = \sum_k p(h|x \in I_k, f = 1)p(x \in I_k)$$

  - for each region Ik, head pose probability is modeled with a Gaussian
  - probability of being in a region modeled with a Gaussian as well



Result example

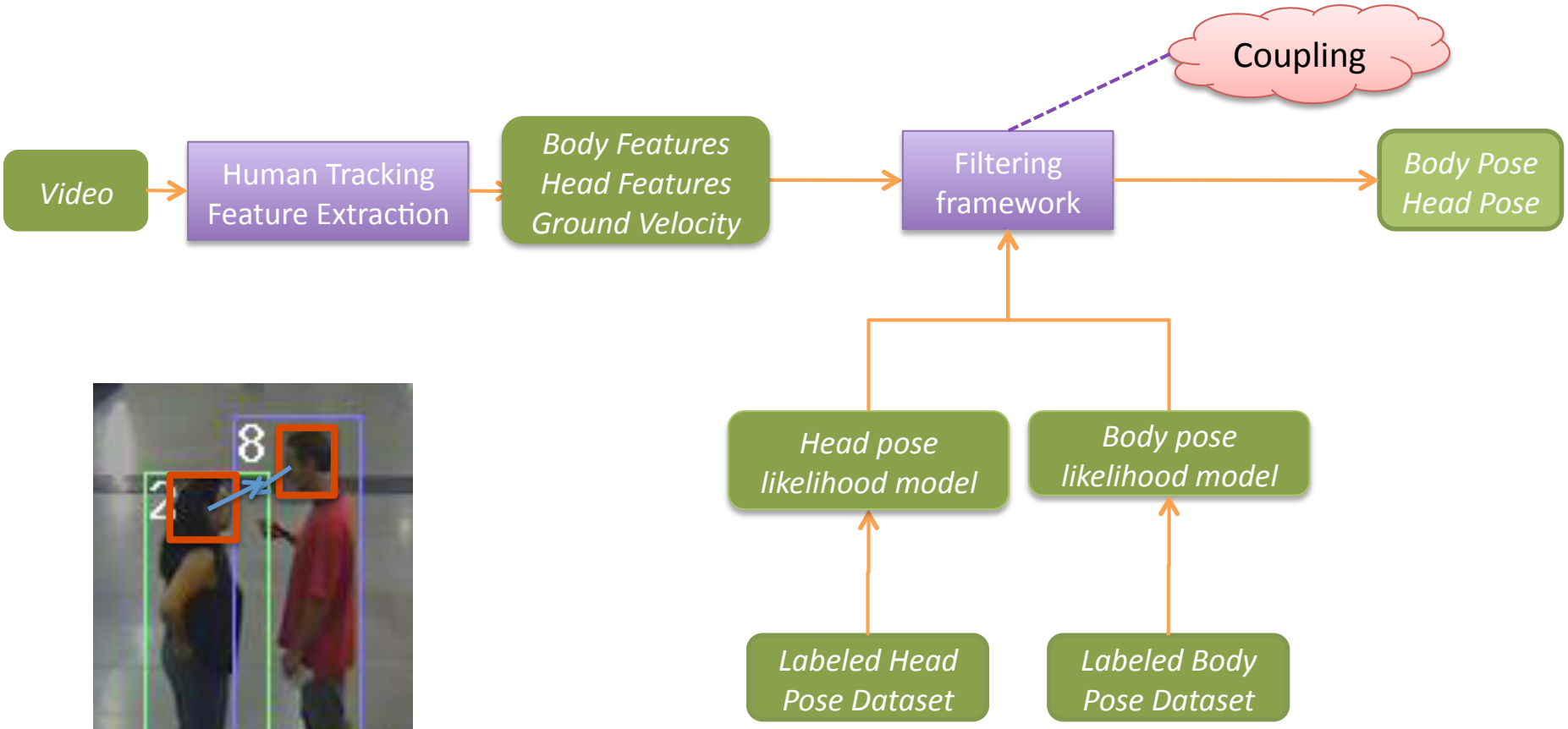
Head and body pose estimation



Estimation is challenging

- Exploit coupling during filtering
- Exploit coupling for adaptation

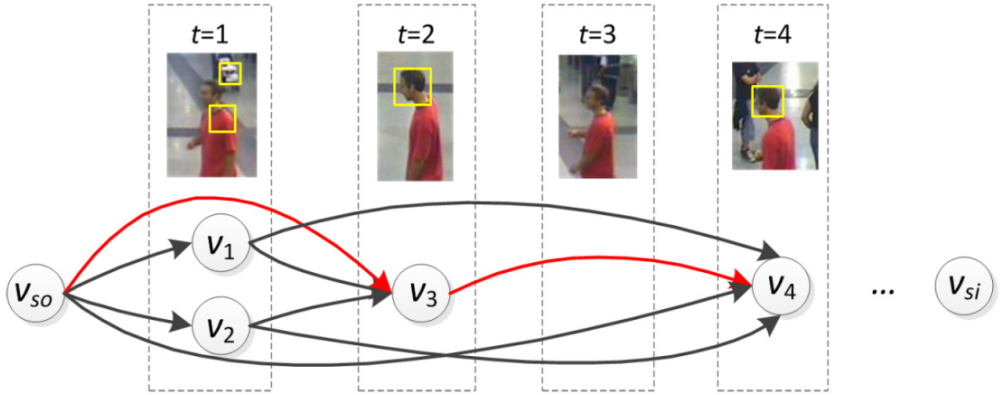
Filtering framework with cue coupling (1)



Filtering framework with cue coupling (2)

Body and Head Location Tracking

- Body location tracking: Multi-human tracking by CRF model [Heili 2011]
- Head location tracking
  - HoG-SVM head detector employed on extended body region
  - Detection based tracking: path probability optimization

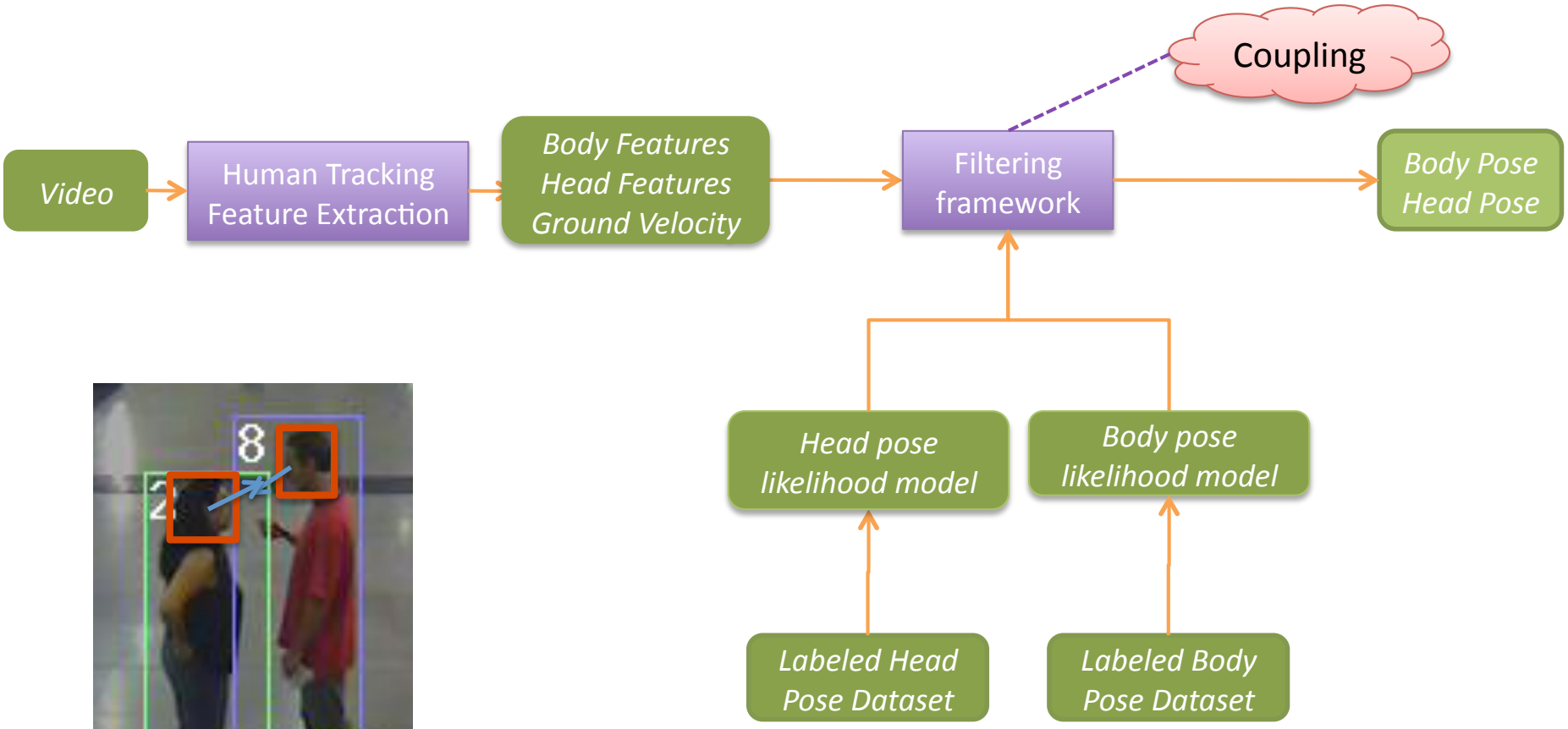


$$p_{ij} = p_{siz}(v_j|v_i)p_{loc}(v_j|v_i)p_{app}(v_j|v_i)p_{mis}(v_j|v_i)$$

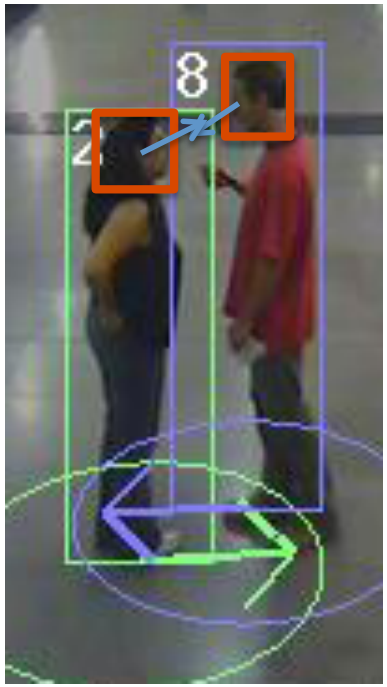
Continuity (size, location, appearance)

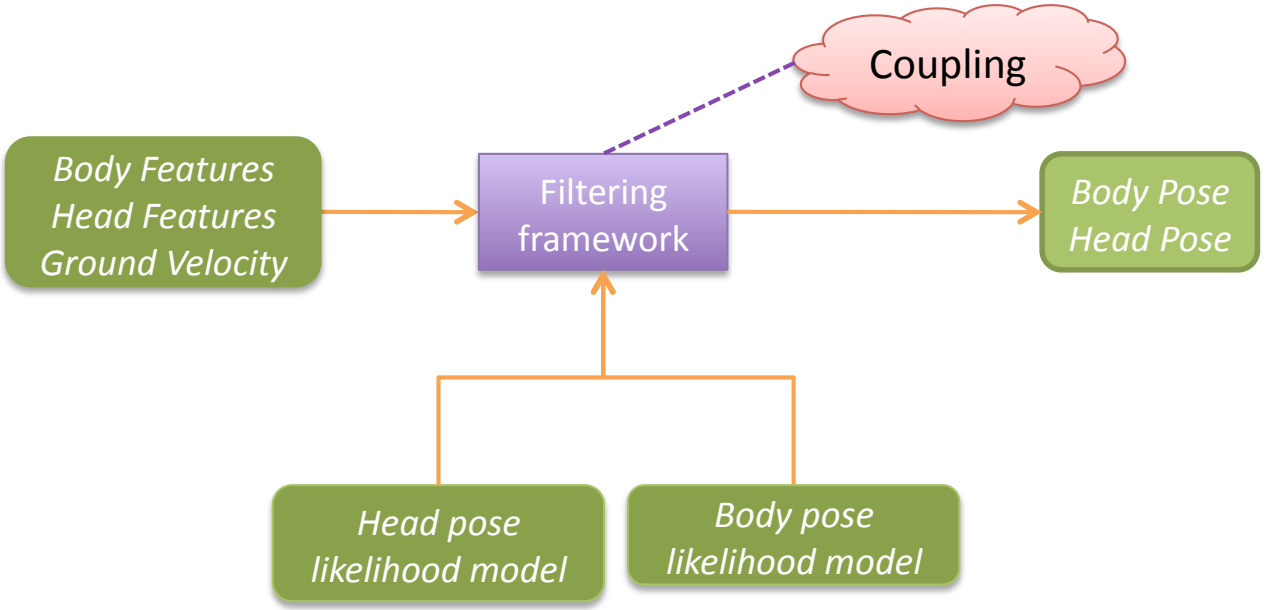
Miss detection

Filtering framework with cue coupling



Filtering framework with cue coupling (3)





Filtering framework with cue coupling (4)

$$p(\mathbf{s}_t|\mathbf{z}_{1:t}) \propto \underbrace{p(\mathbf{z}_t|\mathbf{s}_t)}_{\text{likelihood}} \int \underbrace{p(\mathbf{s}_t|\mathbf{s}_{t-1})}_{\text{dynamics}} p(\mathbf{s}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{s}_{t-1}$$

- State space

$$\mathbf{s}_t = [\mathbf{x}_t, \dot{\mathbf{x}}_t, \theta_t, \alpha_t,]$$

Position + speed on ground plane

Body orientation (on ground plane)

Head orientation

- Likelihood for body and head

Filtering framework with cue coupling (5)

$$p\left(\mathbf{s}_t|\mathbf{s}_{t-1}\right) = p\left(\alpha_t|\theta_t, \alpha_{t-1}\right) p\left(\theta_t|\theta_{t-1}, \dot{\mathbf{x}}_t\right) p\left(\mathbf{x}_t, \dot{\mathbf{x}}_t|\mathbf{x}_{t-1}, \dot{\mathbf{x}}_{t-1}\right)$$

Standard AR model

Head pose dynamics

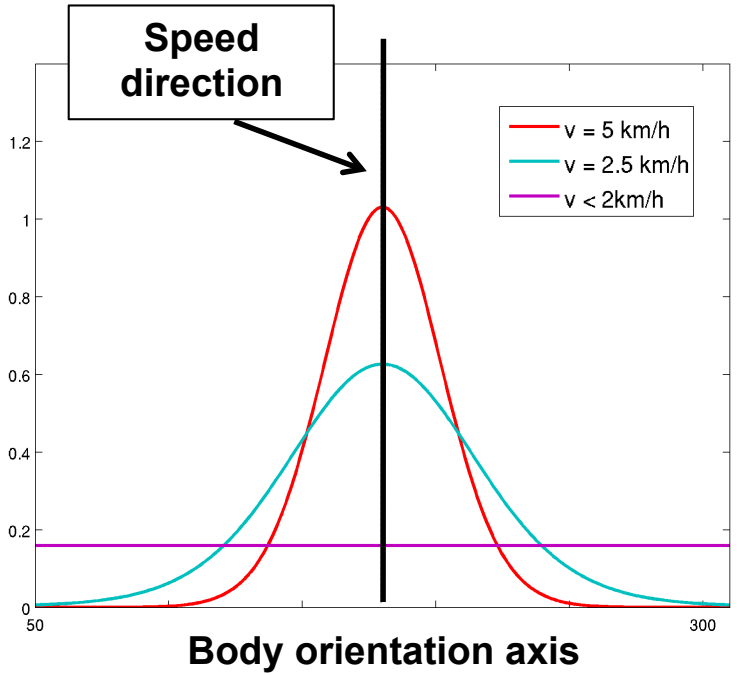
Body pose dynamics

- Dynamics introduces coupling

  E. g. body orientation dynamics: two terms

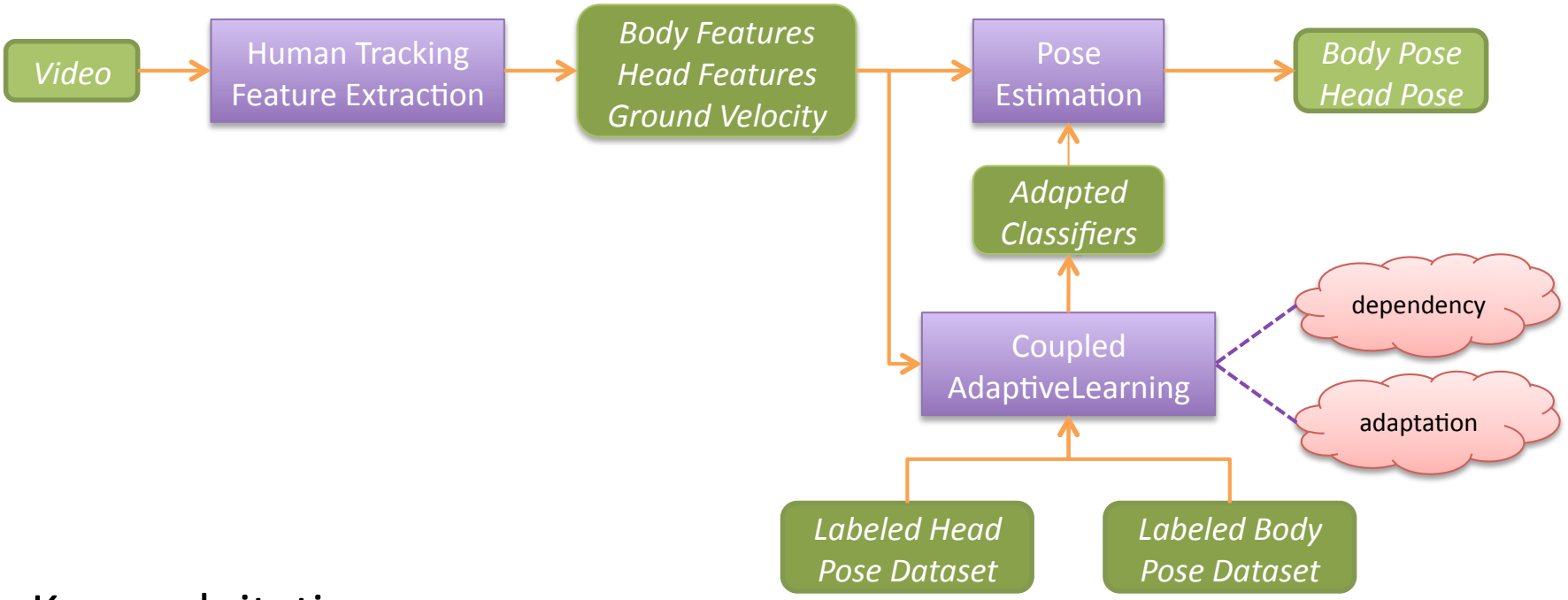
  (i) favors alignment of body orientation and speed direction

  (ii) Speed dependent coupling



Coupled adaptative classifier

- However : Pre-trained models perform poorly on test data
  - Change of view point, different appearance, not enough training data

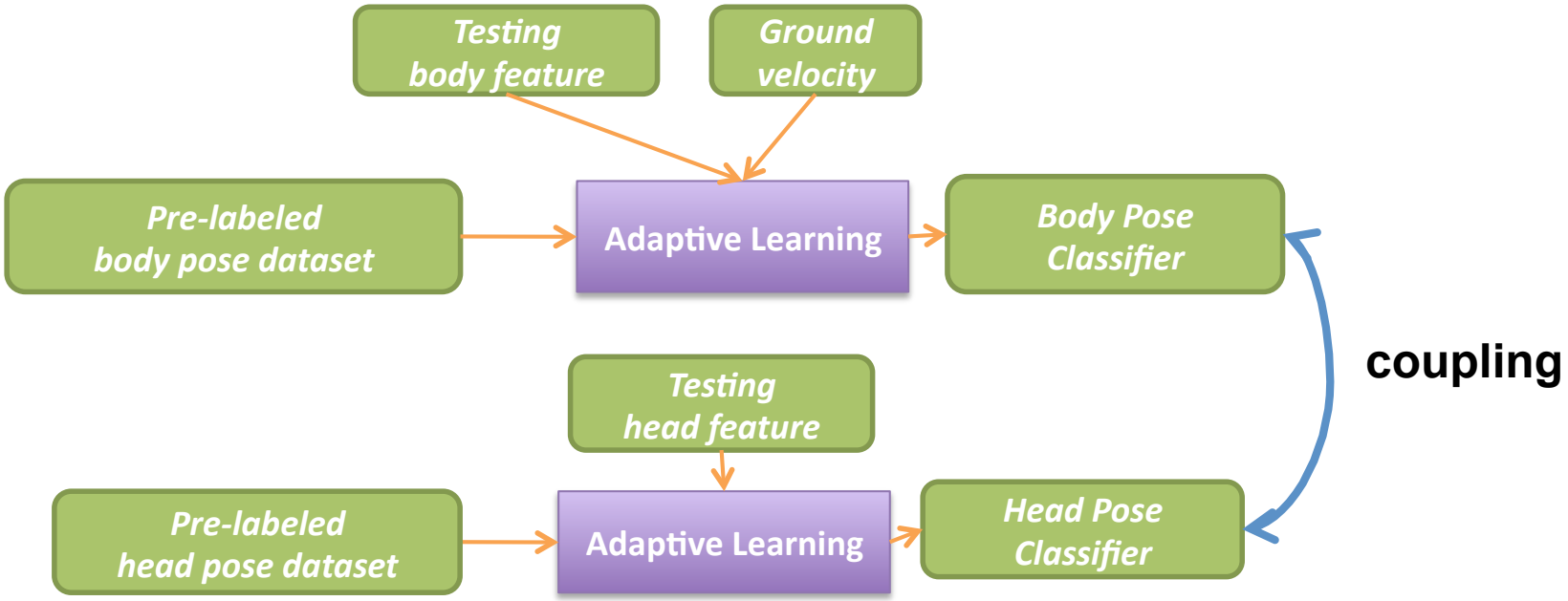


- Key exploitation
  - The adaptation of classifiers using unlabeled testing data
  - The inter-cue coupling (velocity, body pose, head pose)

Chen, Odobez, CVPR 2012

Coupled adaptative classifier (2)

- Adaptation of head and body pose classifiers
- Exploitation of the coupling between the two classifiers
  - Set constraints on adaptation



Coupled adaptive classifier learning (1)

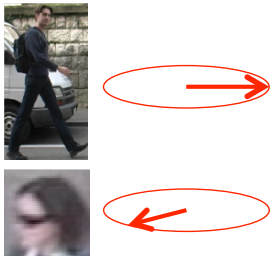
- Task: Learn classifiers $\mathbf{y}^b = f^b(\mathbf{x}^b : \mathbf{W}^b)$, $\mathbf{y}^h = f^h(\mathbf{y}^b : \mathbf{W}^h)$

  pose vector, feature, parameter

  – Kernel approach: mapping to infinite RKHS

  Goal: Learning W^b and W^h

- Data
  - Labeled external data
    - $\mathcal{D}_b = \{(\mathbf{x}_i^b, \mathbf{y}_i^b)\}$, $\mathcal{D}_h = \{(\mathbf{x}_i^h, \mathbf{y}_i^h)\}$
  - Unlabeled target data
    - $\mathcal{D}_t = \{(\tilde{\mathbf{x}}_i^b, \tilde{\mathbf{x}}_i^h, \mathbf{v}_i, u_i)\}$



Two separate labeled datasets for body and head pose respectively

body feature, head feature, velocity direction, velocity reliability flag

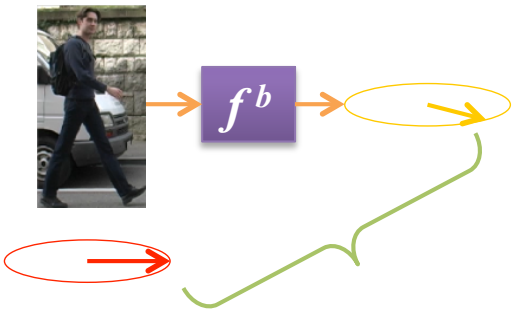
Unlabeled test data where we exploit the coupling information

- Problem formulation: Energy minimization
  - $E(\mathbf{W}^b, \mathbf{W}^h) = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r$

Coupled adaptive classifier learning (2)

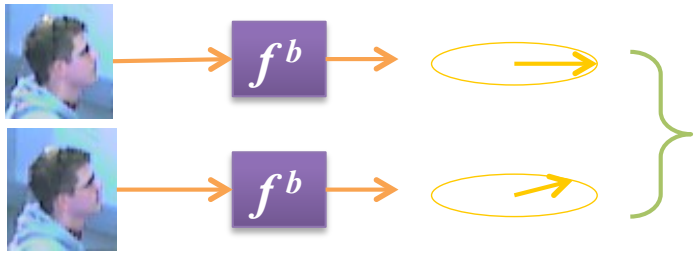
- Label factor E_l
  - The classifiers should respect the labeled dataset

$$E_l^b = \frac{1}{n_b}\sum_{i=1}^{n_b}\left\|\mathbf{M}f^b(\mathbf{x}_i^b) - \mathbf{M}\mathbf{y}_i^b\right\|_F^2, E_l^h = \cdots$$



- Manifold structure factor E_m
  - The mapping between feature space to pose space should be continuous and smooth
  - Close data points in feature space map to close points in pose space

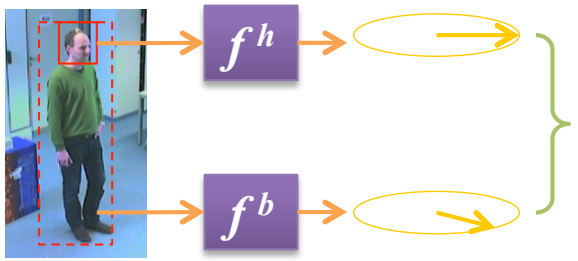
$$E_m^b = \frac{1}{\sum_{i\neq j} s_{ij}^b}\sum_{i\neq j} s_{ij}^b\left\|f^b(\mathbf{z}_i^b) - f^b(\mathbf{z}_j^b)\right\|_F^2, E_m^h = \cdots$$



Coupled adaptive classifier learning (3)

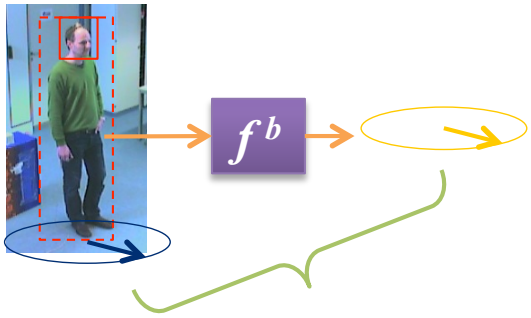
- Body pose and head pose coupling factor E_c^{bh}
  - Body pose and head pose should be similar

$$E_c^{bh} = \frac{1}{n_t}\sum_{i=1} n_t \left\| \mathbf{M} f^b(\tilde{\mathbf{x}}_i^b) - \mathbf{M} f^b(\tilde{\mathbf{x}}_i^h) \right\|_F^2$$



- Velocity and body pose coupling factor E_c^{vb}
  - Body pose should be close to velocity direction, provided that the velocity magnitude is large enough

$$E_c^{vb} = \frac{1}{\sum u_i}\sum_{i=1}^{n_t} u_i \left\| \mathbf{M} f^b(\tilde{\mathbf{x}}_i^b) - \mathbf{M}\mathbf{v}_i \right\|_F^2$$



- Regularization factor E_r
  - Better generality

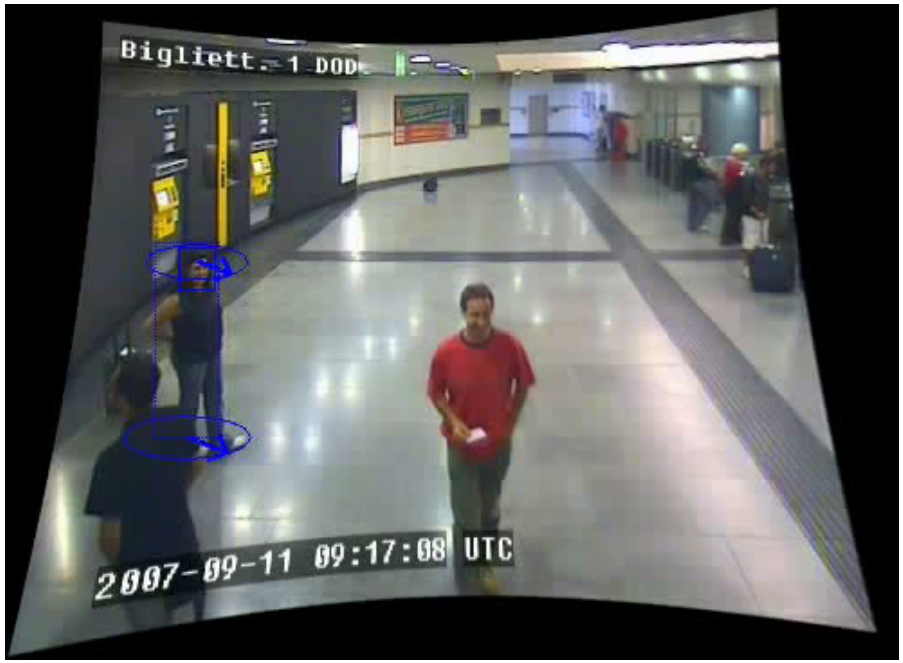
Optimisation : closed form solution !

Experiments

- CHIL data (CLEAR contest 2007)
  - Laboratory scenario
  - Static persons (velocity reliability $\frac{1}{n_t}\sum u_i = 0\%$)
  - Large body/head pose range
  - Ground-truth: ~7 min data
    - Body pose: Hand labeled; Head pose: provided by dataset.

Experiments

- Metro station data
  - Metro station surveillance video at Torino, Italy
  - Both static and moving persons (velocity reliability 24%)



Experiments

- Smart room dataset
  - Indoor scenario (velocity reliability 5%)
  - Ground-truth: hand labeld ~4 min data for body and head pose

Experiments

- TownCentre data [*Benfold 2011*]
  - Mainly moving people (velocity reliability 73%)
  - Ground-truth data: hand labeled ~15 tracks for body and head pose

Experiments

		Ours (default)	Walking direction	Ours (baseline)	Ours (no prior)	Ours (setting [5])	Coupled TF [7,16]	Ours +TF
CHIL	Body	35.3	78.7	50.7	80.7	80.7	44.5	37.7
	Head	36.0	79.5	56.9	85.1	85.1	46.7	35.2
Metro Station	Body	29.4	79.9	53.8	63.5	82.2	42.2	32.8
	Head	30.0	77.1	40.5	66.7	85.4	40.5	31.0
Smart Room	Body	23.6	66.3	59.9	63.9	63.5	36.3	24.9
	Head	23.6	66.7	29.4	68.2	66.7	33.8	23.9
Town Centre	Body	17.4	19.3	48.1	18.3	18.4	20.1	19.0
	head	18.4	22.9	44.8	19.4	20.5	24.9	25.0

Ours (default): $E = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r$

Ours (baseline): $E = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r$

Ours (no prior): $E = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vb} + \lambda E_r$

Ours (setting [5]): $E = E_l + \alpha E_m + \beta E_c^{bh} + \gamma E_c^{vh} + \lambda E_r$

- Walking direction is a bad indication [illegible]
- Our method significantly outperforms supervised learning exploiting only labeled data
- Prior information is important [illegible]
- Coupled temporal filtering does not further improve the results

Conclusion

- head-pose tracking methods - error from to 10-12 degrees
- DBN models for visual attention modeling
  - multimodal interaction (gaze, head pose, speech, people location)
  - contextual recognition (conversation, ‘gestural activity’, group activity…)
  - mapping function head pose <-> gaze
  - model parameter adaptation
    - benefit from context
- Coupled adaptation for body & head pose estimation

Future research

- Exploit/improve head pose tracker
- Extract gaze
- Apply extracted pose to the HRI or interaction analysis

THANK YOU FOR YOUR ATTENTION - QUESTIONS ?