Contextual Recognition of Visual Focus of Attention in Meetings

Jean-Marc Odobez

(joint work with Sileye Ba)

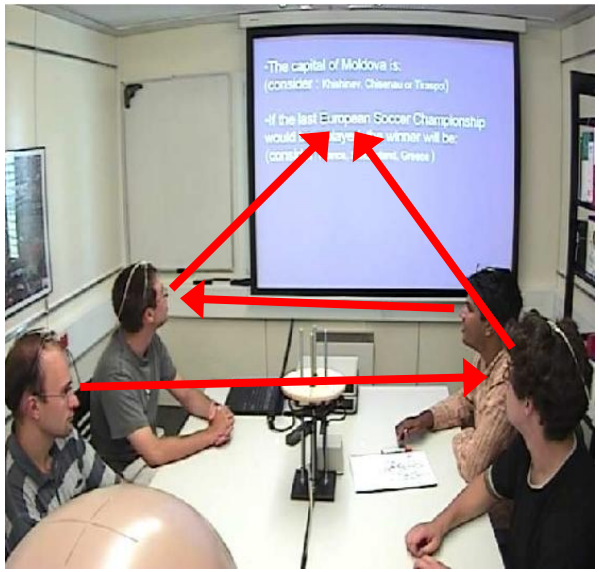
Idiap Research Institute

Joint IM2-AS summer Institute, Riederalp, 1-3 september 2008

Visual focus of attention (VFOA)

- Focus of attention: defined by the **eye gaze**

  "where and at whom or what
  a person is looking at"



- Non-verbal signal which conveys rich information about a person
  - what is he interested in
  - what is he doing (e.g. manipulation)
  - how does he explore a new environment ?
  - reaction to different stimuli

- Gaze is a strong social interaction cue
  - regulate conversation
  - personality traits
    - express intimacy, empathy
    - exercise social control => leadership, social status (cf D. Gatica's talk)

Communication role of gaze in conversations

- Important **non-verbal** interaction cue, with different functions in conversations [Kendon, Goodwin]
  - establish relationship (through mutual gaze)
  - monitoring and regulating the course of interactions
  - examples, in face-to-face conversations
    - listeners show their attention by orienting their gaze to the speaker
    - speakers use their gaze to indicate whom they address and secure their attention
    - when a speaker ends his utterance, he tends to look at the next speaker

    ⇒ gaze is a **turn holding/yielding/taking** cue

    ⇒ gaze interaction **patterns** define some **codes** useful to organize conversations

This presentation

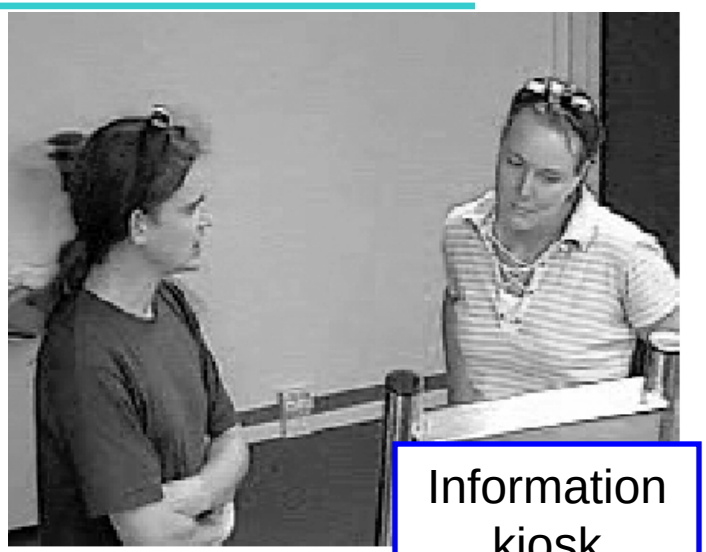
present a model of the gaze/speaking turn relationship

Adressee recognition

- **Addressee** recognition
  "who is a person talking to ?"

  important task in several contexts

- Human-computer interaction
  - information kiosk, robots
  - presence of several people
  - artificial agent: important to know whether it/he is addressed or not



Information kiosk

- psychosocial studies (dyadic, multiparty **face-to-face conversation**)
  => **gaze is a good predictor** of addressee-hood

- presence of artefact
  - object playing central role in a given task (e.g. manipulation) attracts attention
  - overrules trends of eye gaze behaviour observed in face-to-face conversations

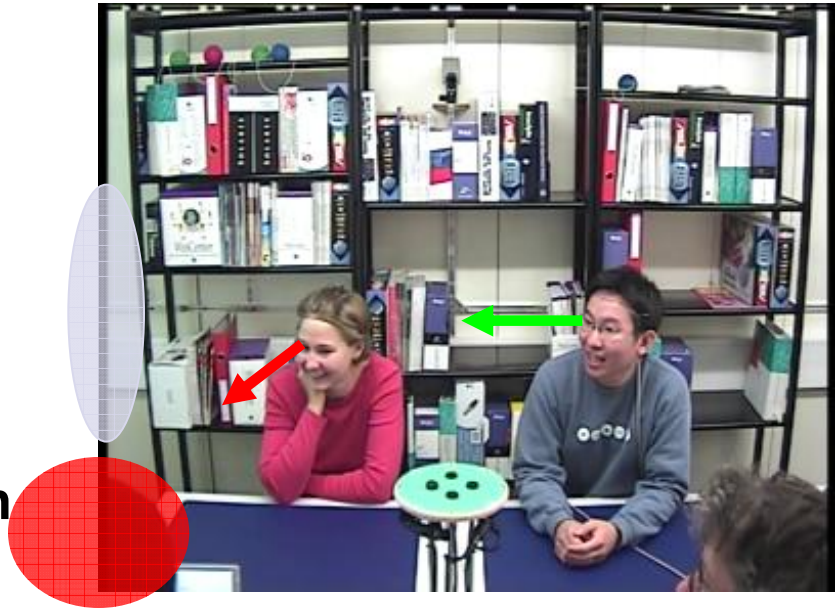
  => this has to be taken into account when modeling VFOA in meetings

Estimating Gaze Direction

- VFOA defined by eye gaze
  => head orientation + eye-in-head orientation

- HCI oriented gazing estimation approaches
  - head mounted system, high resolution iris image
    => invasive, restrict mobility
    interfere with natural conversation

- alternative :
  - **use head pose as surrogate**
    - psychological evidence
      people do exploit head pose to infer the VFOA of other people
    - empirical evidence:
      working in simple settings
  - In meetings, **exploit interaction with other cues**



Main issues

- Head pose is the main cue
  - how and well can we estimate it ?

- VFOA modeling
  - how do we define the VFOA ?
  - how can we estimate the VFOA solely from head pose ?
  - **multi-party** VFOA recognition
    - audio-visual contextual cues
    - interaction models between gaze and speaking turn patterns (conversational event)
    - influence of group activity

Joint Head Location and Pose Tracking [Ba 2005]

State model : $X_t = (S_t, r_t, k_t)$

2D transform

Translation+scaling

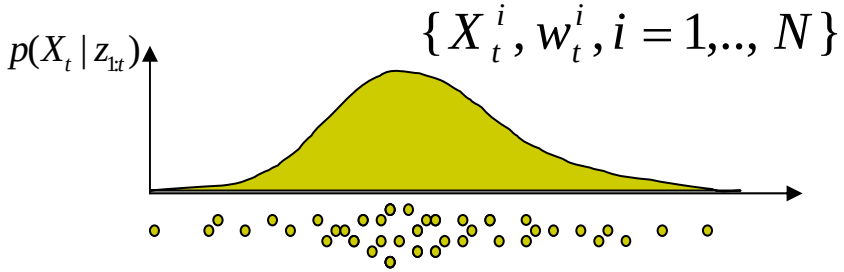
roll

Out of plane head rotation

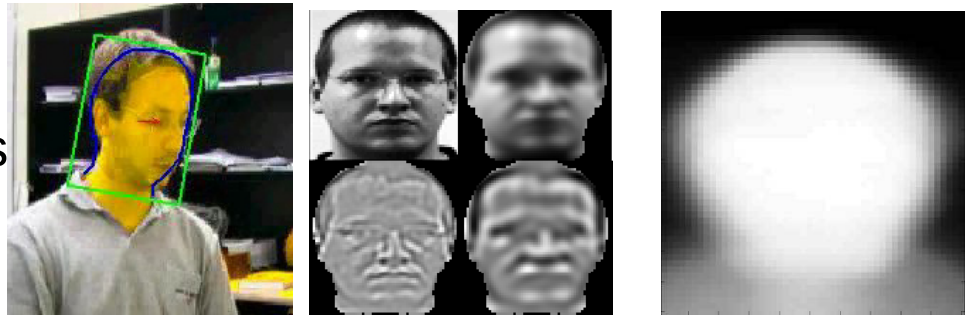
pose exemplar (index)

Joint Head Location and Pose Tracking [Ba 2005]

- Bayesian tracking with sampling approximation (particle filters)



- Joint optimization of location and pose
  not head tracking then pose estimation



- Appearance-based likelihood models
  - pose dependent/independent
  - various features

- Sampling exploits output of a head detector
  automatic (re)initialization and failure recovery

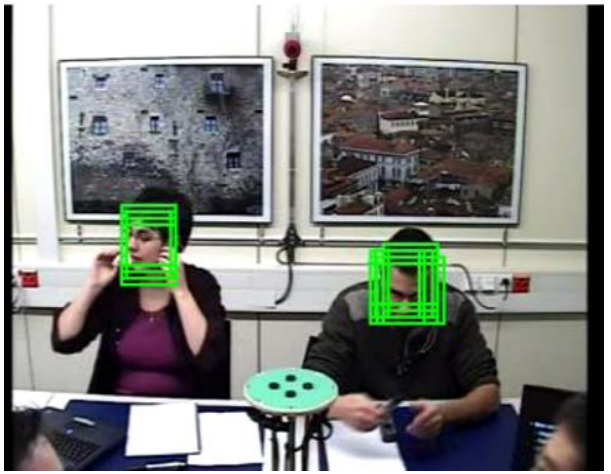


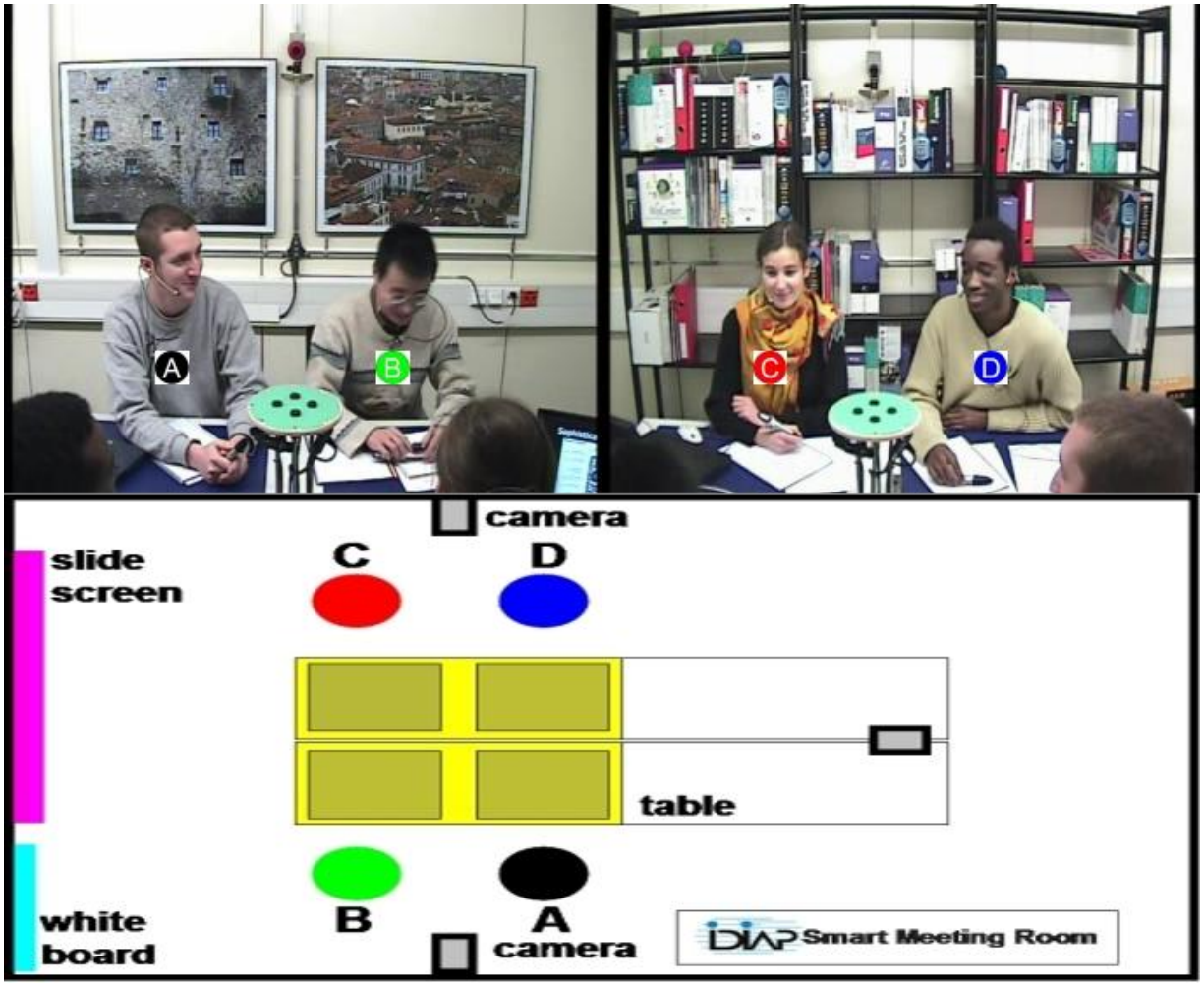
Illustration of head pose tracking

- around 10 degree error in pan
- tilt more difficult to estimate
- large variation across people (some people easier to track)

Multi party VFOA recognition using contextual cues

- how do we define the VFOA ?
  - set-up and task description
  - analysis of evaluation dataset

Meeting set-up

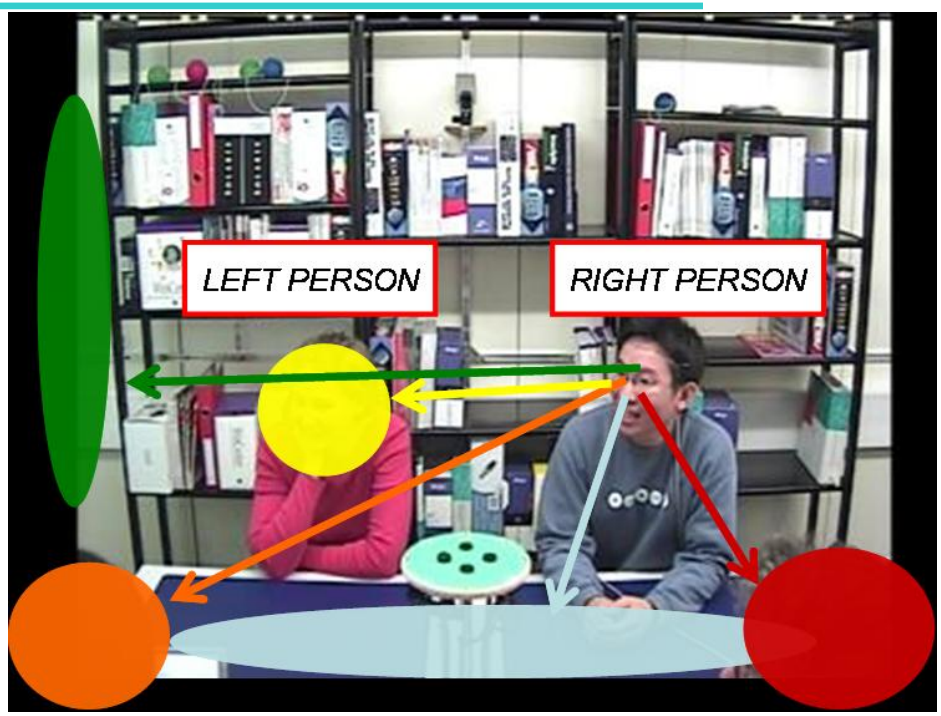


- meeting setup

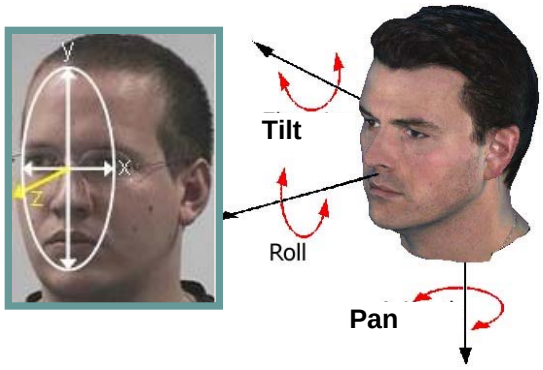
Task description

- Task: estimate a person focus of attention f_t

- FOA set : 6 labels
  - other participants
  - slide screen/white board
  - table
  - unfocused



- Input:
  - head pose features h_t (pan/tilt angles at time step t)
  - other contextual cues



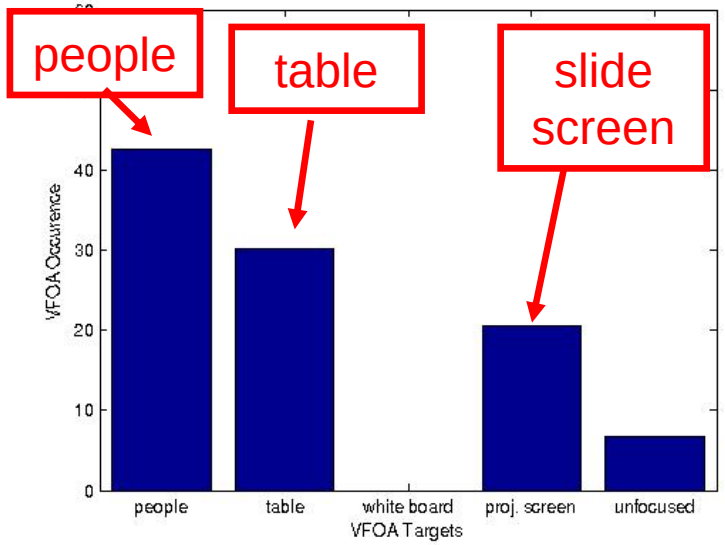
Data samples

VFOA statistics analysis

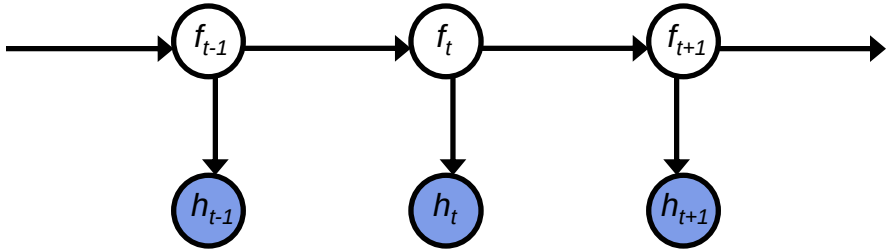
- 4 full meetings, people seated, total: 90 min data
  - people presenting, discussing

- real people behavior
  - laptop and object manipulation
  - large variation of body poses, gaze behavior, gestures

- VFOA analysis
  - only 43% looking at people
  - around 30% looking at table
  - ⇒ people use their laptop
  - ⇒ avert gaze
  - ⇒ 'long-meeting' effect
    - people listen while looking down at table (without changing head pose)
    - bored people

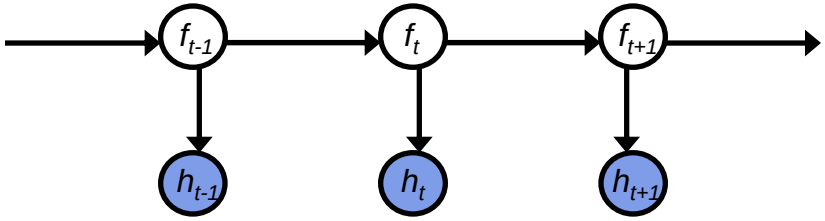


Independent VFOA recognition



- Input: head pose of one person
- Ouput: recognized VFOA for this person

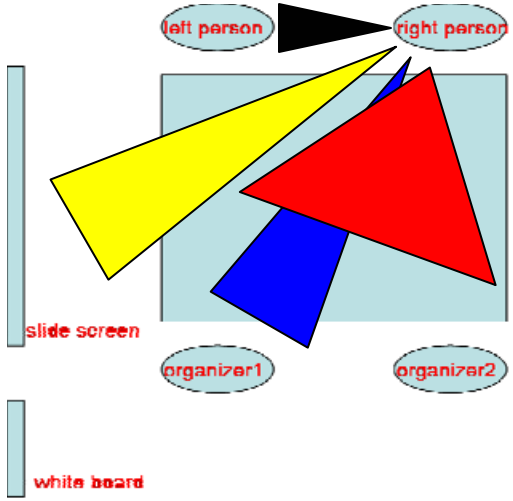
FOA modeling using HMM



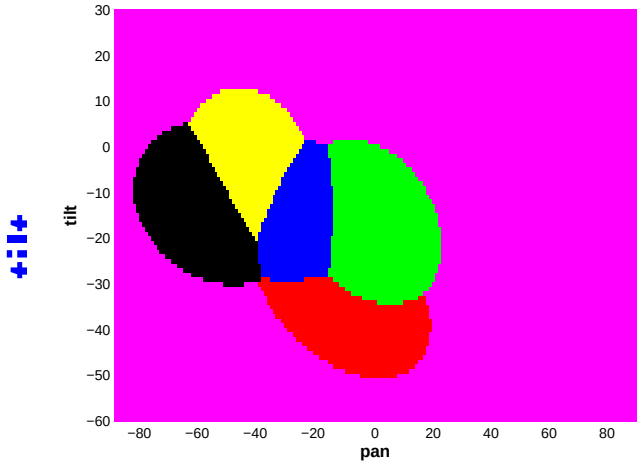
- **observation likelihood $p(h_t \mid f_t)$**
  - Gaussian distribution for regular label

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

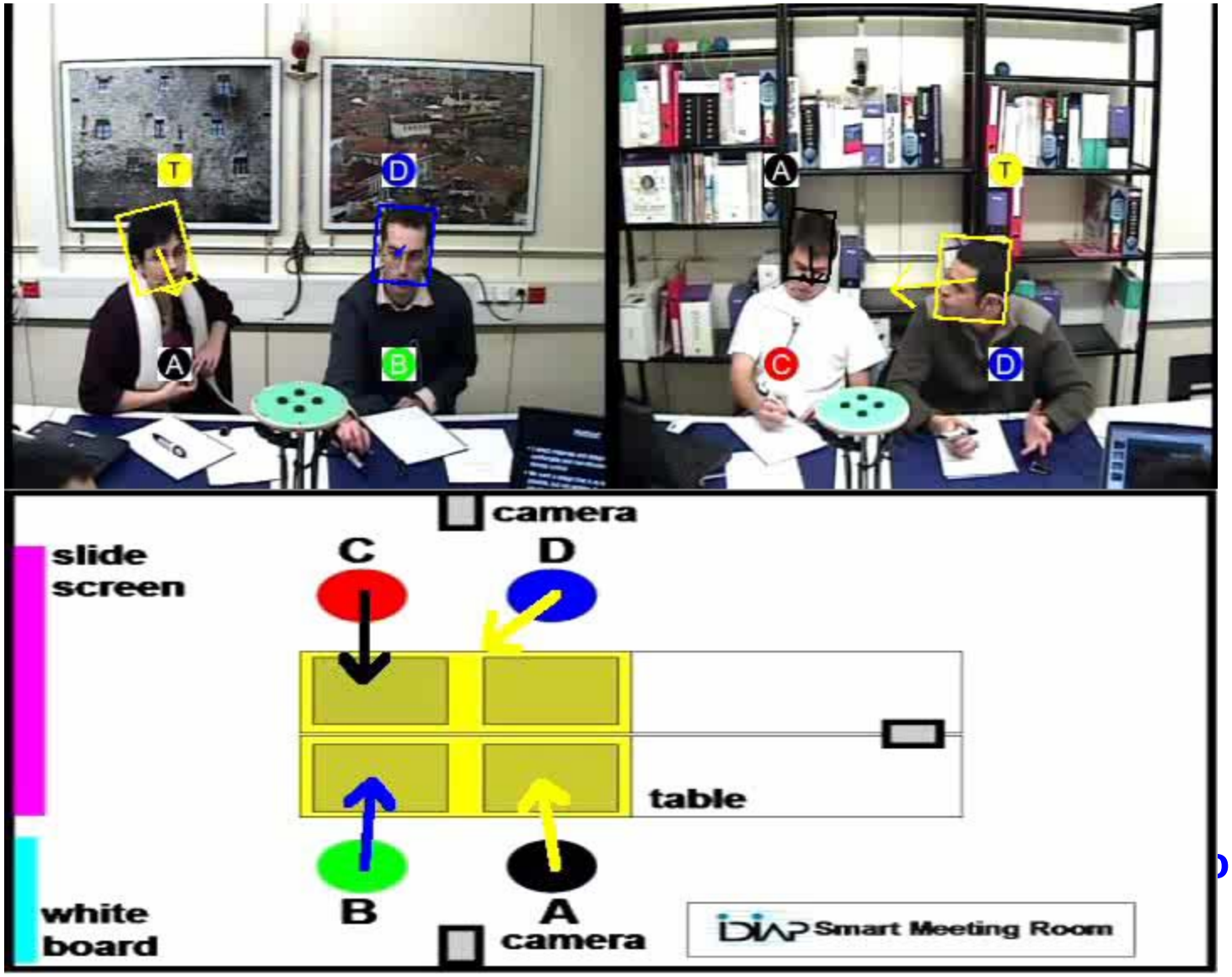
  - uniform distribution for unfocused label



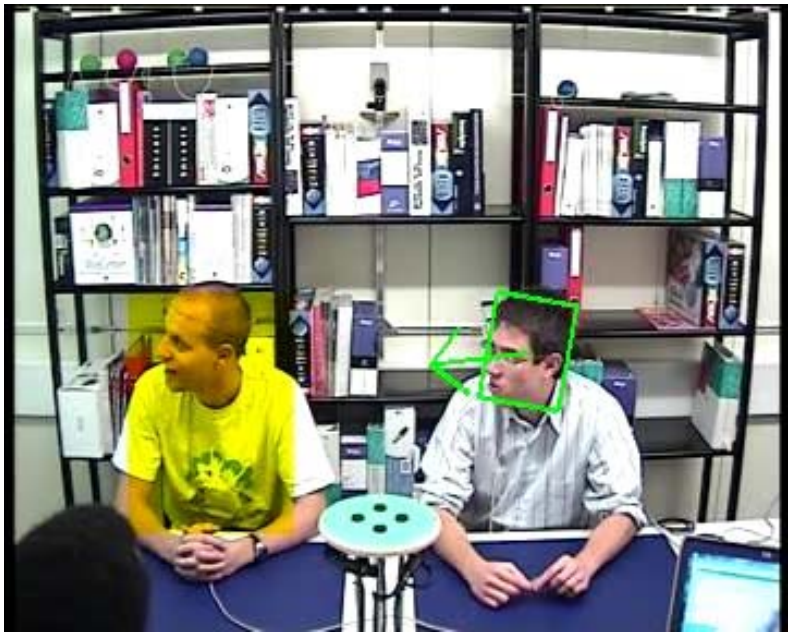
- **dynamic model $p(f_t \mid f_{t-1})$**
  - transition between the different VFOA states
  - set to favor smooth VFOA sequences
  - no other prior



illustration



Ambiguities

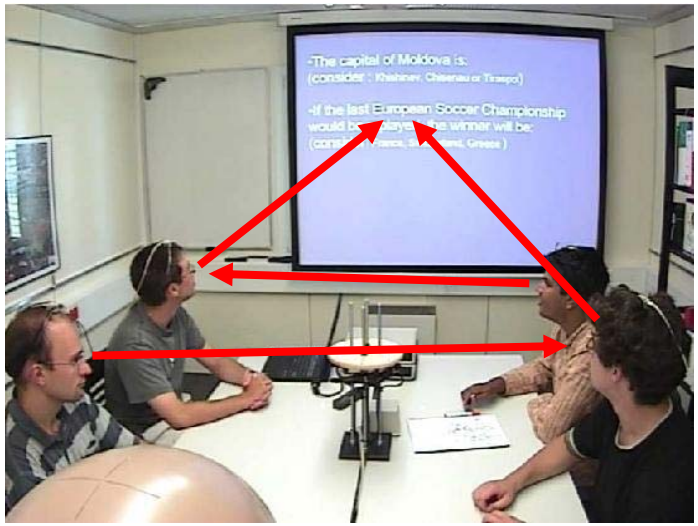


⇒ modeling people interactions and context should help

Multi-party VFOA recognition using contextual cues

- **task:** recognize the VFOA of all participants

- **social interaction provides context**
  - we often share the same VFOA
  - when a person speaks, we tend to look at her/him
  - when a new slide is displayed, we tend to look at it



goal: **integrate this knowledge into a principled model**

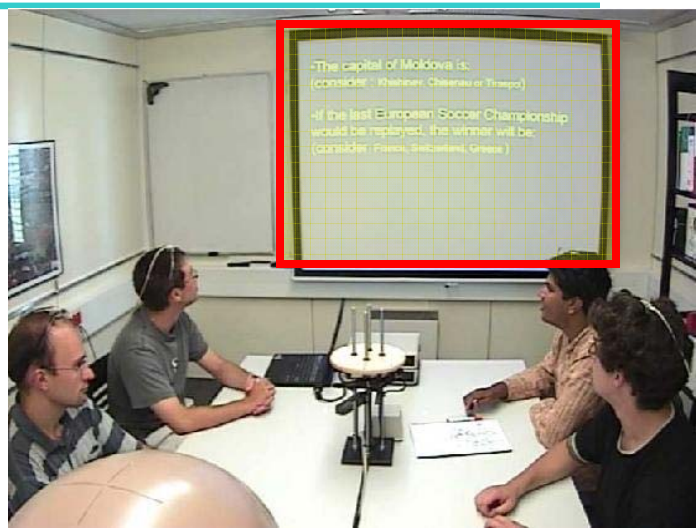
- **Cues:** head pose **and** audio-visual contextual cues
- interaction models between gaze and speaking turn patterns (conversational event)
- influence of group activity (slide presentation)

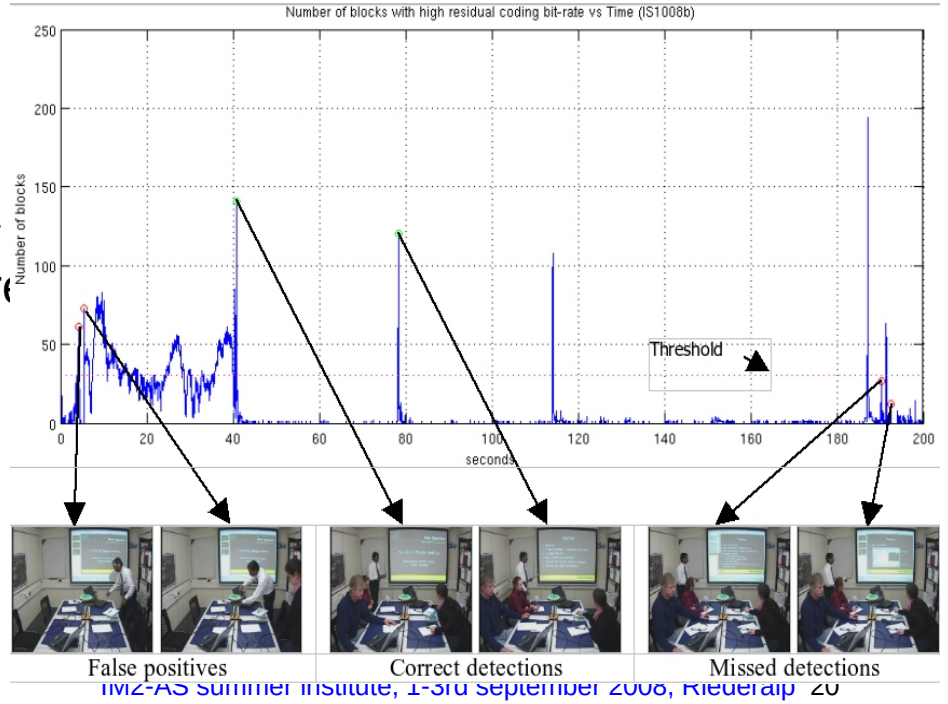
Contextual cue a_t : slide activity modeling

- rely on automatic detection of slide changes
  - activity features extracted from central camera
  - thresholding => slide change instants
  - computation of variable a_t

“Elapsed time since the last slide change”

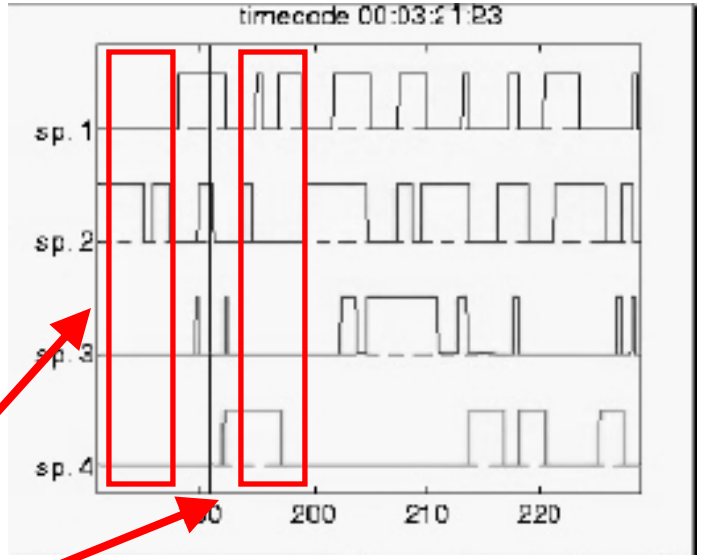
(Note: a slide change corresponds to any new material displayed on the slide are





Speaking activity

- Speaking activity of each person
  - thresholding signal energy of close talk microphone
  - **Cue:** proportion of speaking time over a temporal window



[0 , **0.8** , 0 , 0]

[**0.5** , 0.1 , 0 , **0.6**]

=> more robust than instantaneous measures of speech

Interaction modeling : conversational events

- Characterization of the communication flow

  => introduction of new (hidden) variables e_t

Conversational events

- mainly defined from speech cue
- all possible combinations of speaking/silence per participant => 16 events
  - event type:
    - silence
    - monologue
    - dialogue
    - discussion (3 or 4 people)
  - who is involved

- also relate to VFOA activity

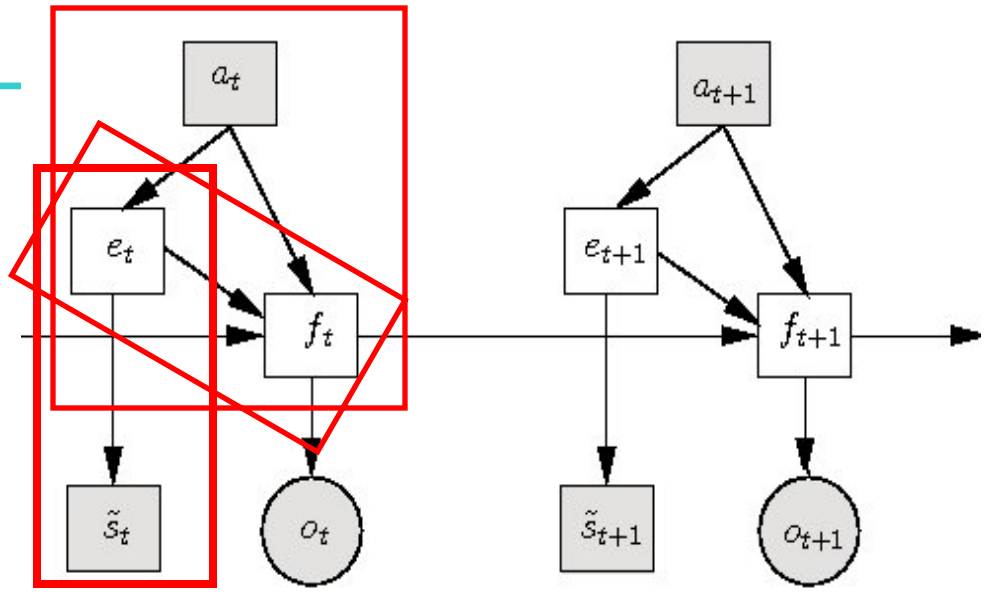
e.g. monologue by person D

Multi-party Dynamic Bayesian Network Model

- hidden states:
  - joint focus of all participants

$$f_t = (f_t^1, f_t^2, f_t^3, f_t^4)$$

  - conversational events e_t



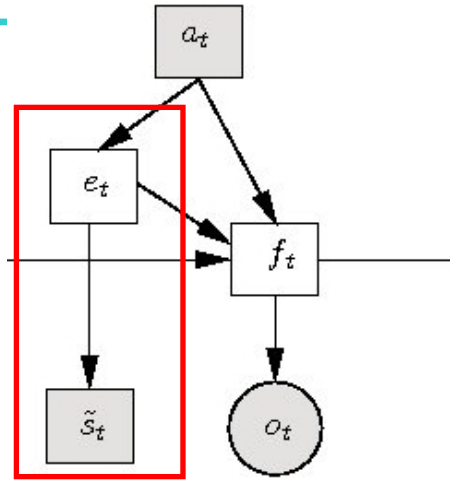
- observations
  - head pose o_t pan/tilt angle for all people
  - slide activity a_t elapsed time since last-slide change
  - speaking activity proportion of speaking time, for all people
- assumption
  - conversational event controls
    - speaking activity
    - dynamics of gaze
  - this control is modulated by the slide activity

  => conversation activity, esp. gaze, varies with group activity

Speaking activity likelihood

- assumption
  - people speaking activity independent given the conversational event

$$p(\tilde{s}_t|e_t = E_j) = \prod_{k=1}^{4} B_j(\tilde{s}_t^k|\eta_{j,k}, T)$$

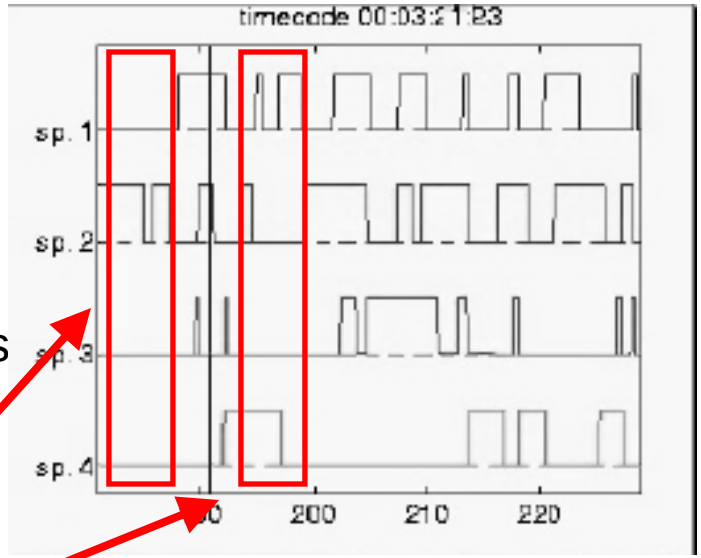


  - individual speaking distribution: Beta distribution characterized by ideal speaking proportions

e.g. event: monologue of person 2

$$\eta_j = [0.05, 0.9, 0.05, 0.05]$$

  - high likelihood if observations near ideal values



e.g. window 1 [0 , **0.8** , 0 , 0]

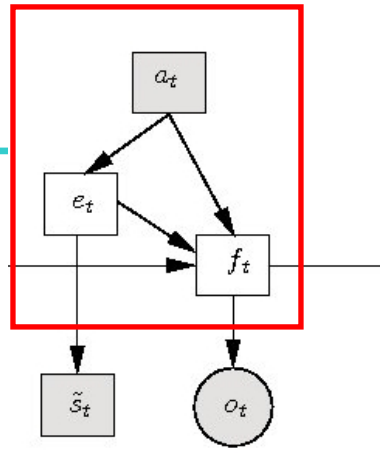
e.g. window 2 [**0.5** , 0.1 , 0 , **0.6**]

context

- intuition: a new slide?
  - people turn their attention to it
  - after some time, attention shifts back to the discussion

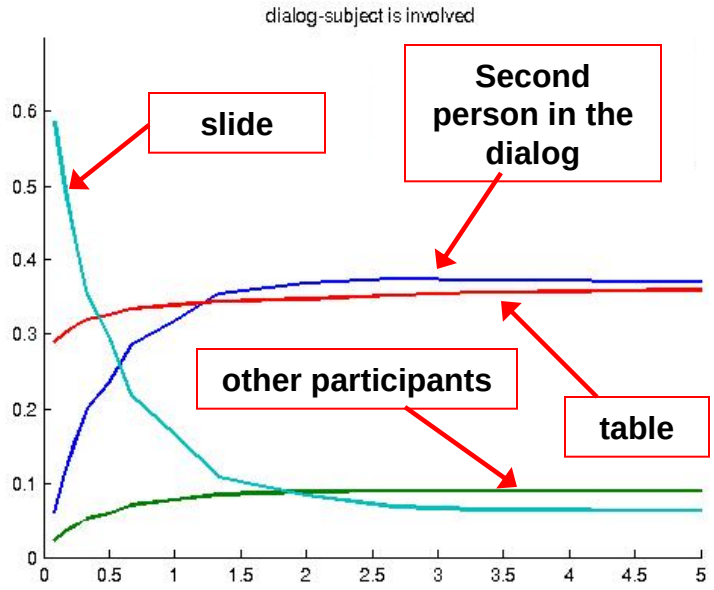
- intuition: person 1 makes a monologue ?
  - people look at him
  - exceptions
    - long monologues (audience looks at table)
    - person talks while a slide is displayed (audience looks at slide)
    - ⇒ Interaction between slide context and conversation activity needs to be taken into account

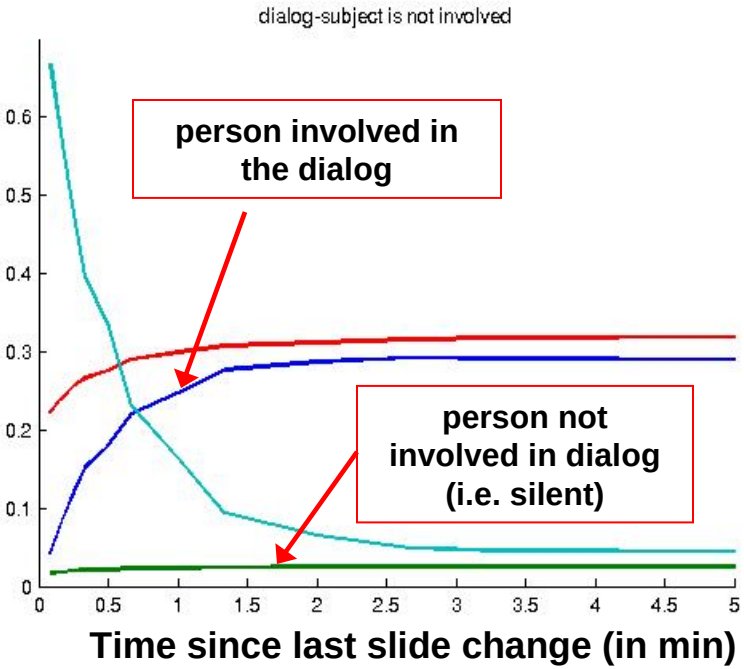
Context



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

- **joint influence** of slide and conversational event 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

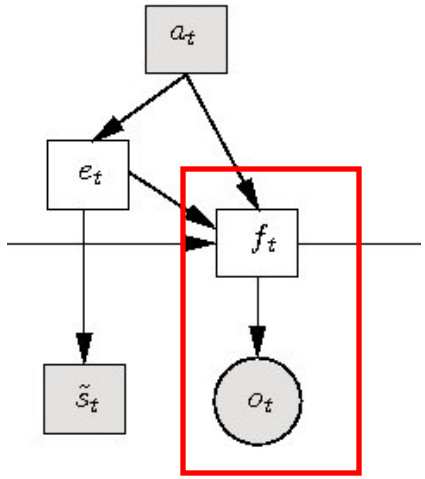




Head pose – VFOA relationship

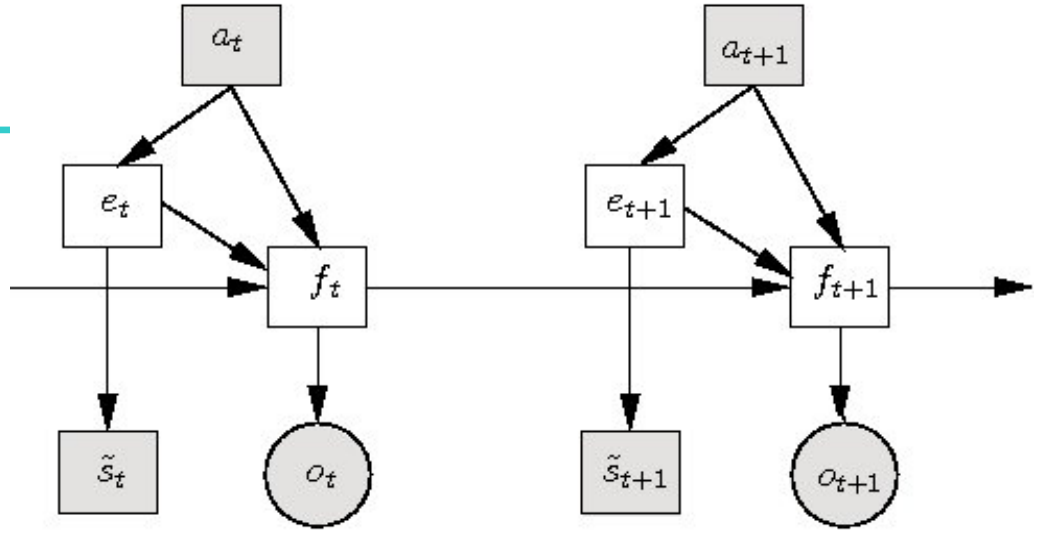
- Assumption
  - head observations independent given VFOA of all participants

$$p(o_t|f_t) = \prod_{k=1}^{4} p(o_t^k|f_t^k)$$



  - Same Gaussian model as with independent case

Bayesian inference

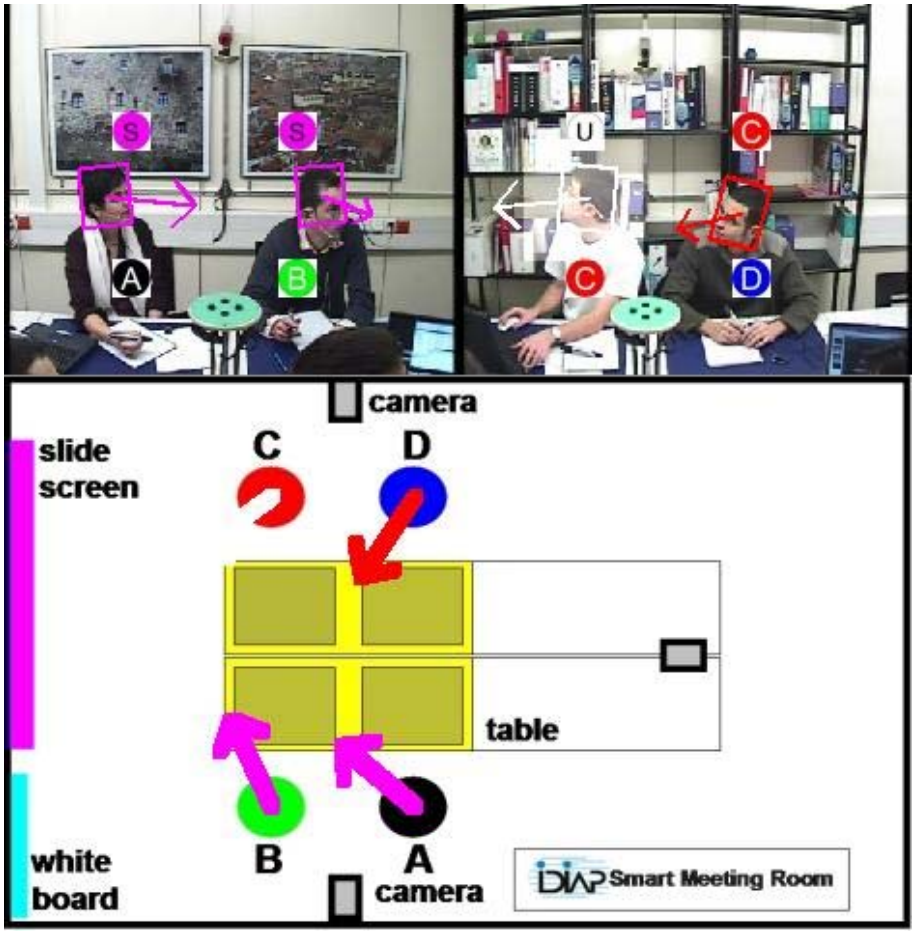


- Maximization of joint posterior distribution of hidden variables given observations

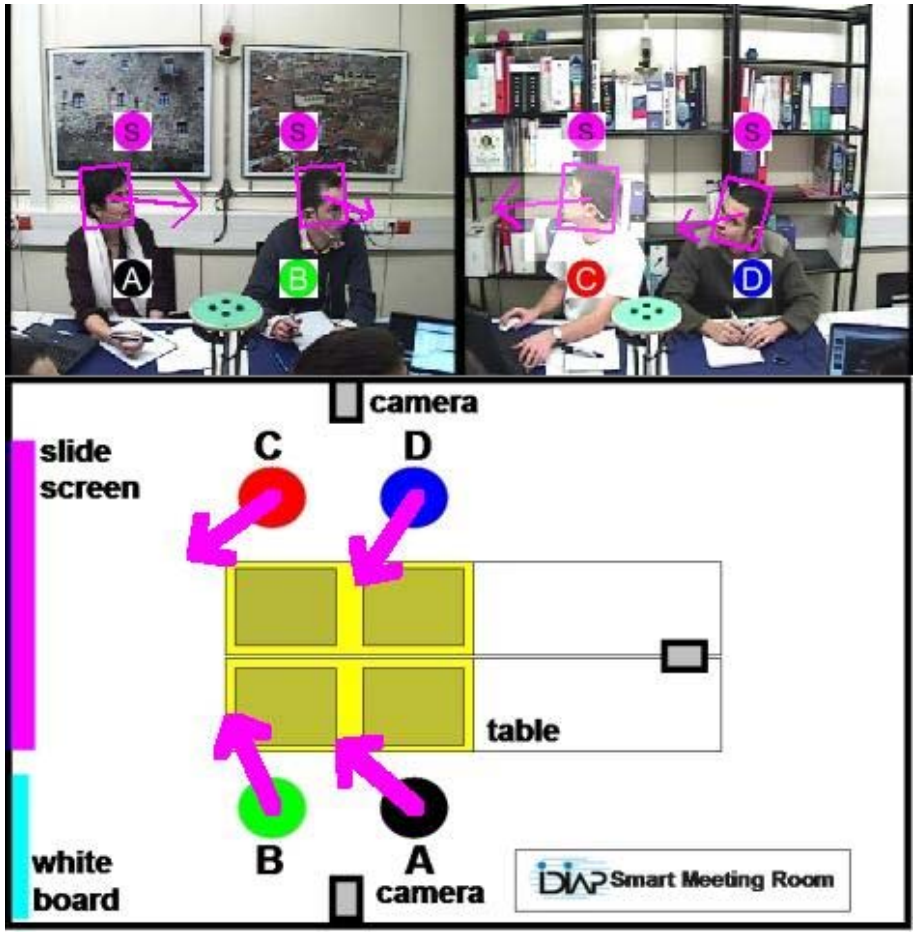
$$p(e_{1:T}, f_{1:T}, \lambda | \tilde{s}_{1:T}, o_{1:T}, a_{1:T})$$

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

illustration: group and slide activity



independent recognition
(head pose only)

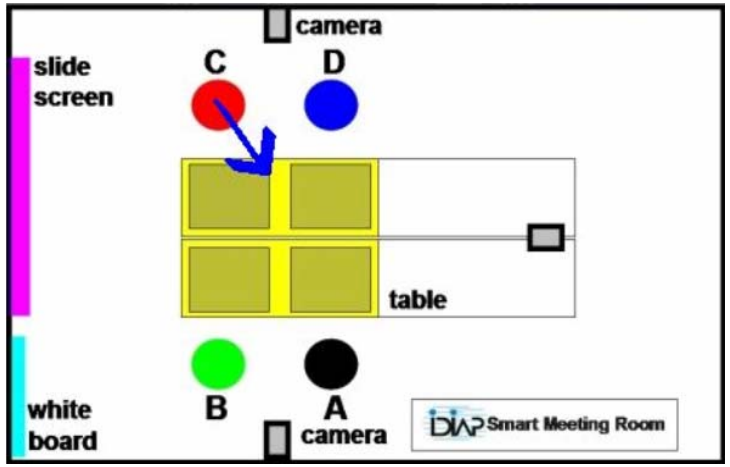


multi-party recognition
using contextual cues

Results

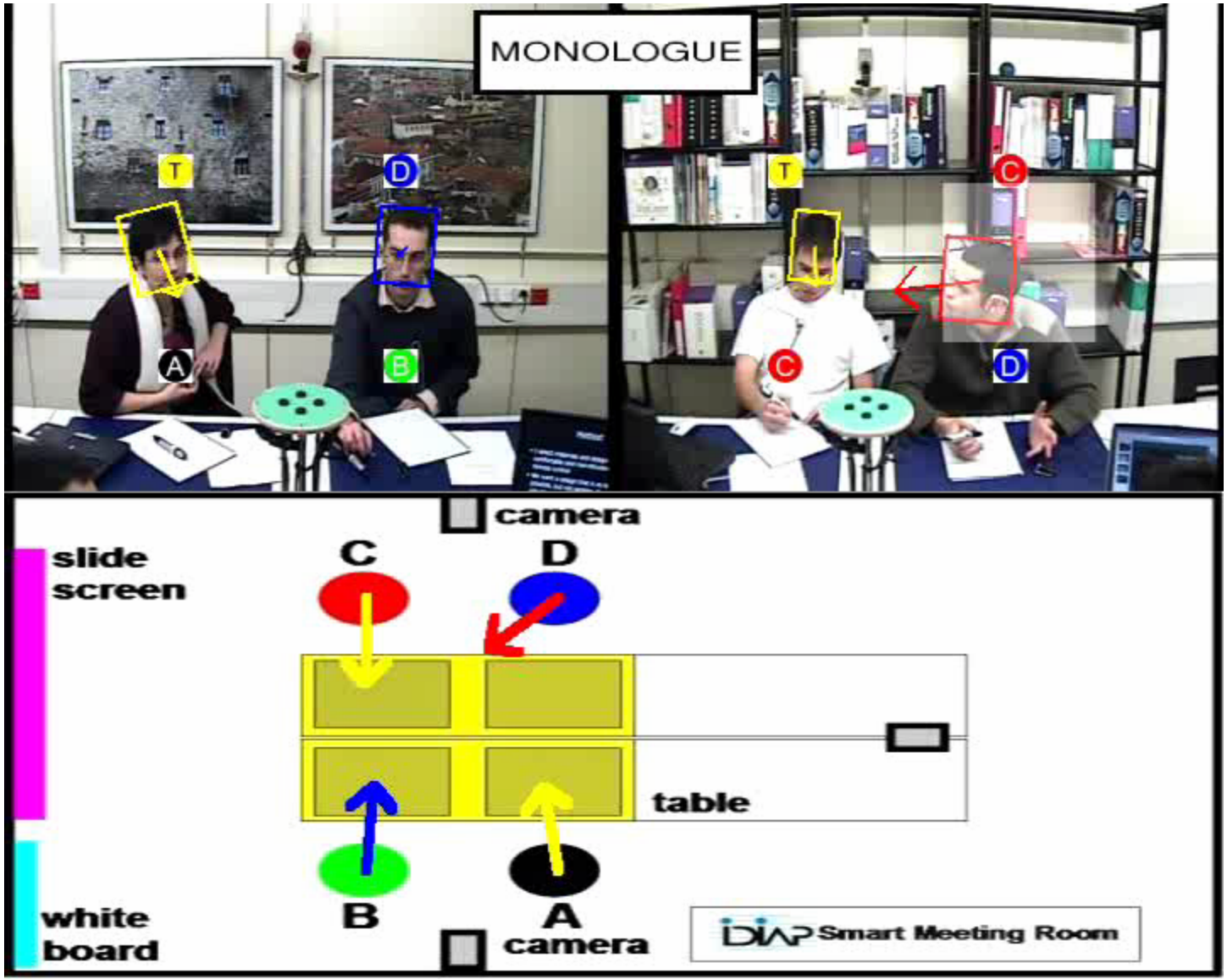
- performance measure : percentage of correctly recognized VFOA

position	A	B	C	D	**mean**
Baseline, **independent**	39.5	50.5	52.5	27.3	**42.4**
Multi-party, cognitive, **slide context**	45.6	51.4	50.7	39.8	**46.9**
Multi-party, cognitive, **conversational event context**	51.5	54.0	50.5	43.2	**49.9**
Multi-party, cognitive, **full context**	51.0	54.1	57.2	47.0	**52.2**



- baseline: 42% => challenging problem
- seats A and D: more VFOA ambiguities
- multi party
  - context helps:
    - **slide** context: **+ 4.5%**
    - **conversational** context: **+ 7.5%**
  - **full context**: **+ 10%** absolute improvement
  - higher improvement on seats with larger ambiguities

Demonstration video: full context



Conclusion

- **head-pose tracking**
  - difficult given image resolution
  - around 10-12 degree average error in pan, depending on people appearance

- **VFOA recognition**
  - **independent recognition from head pose (baseline)**
  - **multi-party VFOA using contextual cues**
    - Gaze/speech interaction modeling through conversational events
    - Accounting fro group activity (presentations)

- **future work – how to improve recognition ?**
  - Improve head pose estimation
    - previous study (independent recognition case)

    => importance decrease when using estimated head pose rather than GT pose

  - use other contextual cues (e.g. table activity)

  - model timing information (people tend to look more at speaker at beginning and end of speaker turn)

Thank you for attention

Questions ?