IDIAP RESEARCH REPORT

Multi-Person Visual Focus of Attention from Head Pose and Meeting Contextual Cues

Sileye O. Ba [a] Jean-Marc Odobez [a]

IDIAP–RR 08-47

August 2008

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne

IDIAP Research Report 08-47

Multi-Person Visual Focus of Attention from Head Pose and Meeting Contextual Cues

Sileye O. Ba Jean-Marc Odobez

August 2008

Abstract. This paper presents a model for visual focus of attention (VFOA) and conversational estimation in meetings from audio-visual perceptual cues. Rather than independently recognizing the VFOA of each participant from his own head pose, we propose to recognize participants' VFOA jointly in order to introduce context dependent interaction models that relates to group activity and the social dynamics of communication. To this end, we designed a dynamic Bayesian network (DBN) , whose hidden states are the joint VFOA of all participants, and the meeting conversational events. The observation used to infer the hidden states are the people's head poses and speaking status. Interaction models are introduced in the DBN by the use of the conversational events, and the projection screen activity which are contextual cues that affect the temporal evolution of the joint VFOA sequence, allowing us to model group dynamics that accounts for people's tendency to share the same focus, or to have their VFOA driven by contextual cues such as projection screen activity or conversational events. The model is rigorously evaluated on a publicly available dataset of 4 real meetings of a total duration of 1h30 minutes.

1 Introduction

In society, important decisions are in general taken during meetings. In meetings two or more people come together to discuss about a previously defined topic in order to disseminate information, or to reach an agreement. Due to the importance of meetings, meeting analysis became an important field in computer science. [23] presents research for the recognition of meeting actions, [39, 34, 17] present methods to estimate the most dominant person in a meeting, [19] present research for the recognition of people's addressee in meetings, [42, 4] present methods for the recognition of people'visual focus of attention in meetings, [33] presents methods for the recognition of conversational structures. As a key aspect of meetings, conversations have been widely studied by social scientists [10, 24, 14]. Although in conversations verbal signals (speech) are the main medium, non-verbal signals such as facial expressions, body postures, gestures, gaze are also very important as they convey important information that enhance the smoothness of the conversation flow [1].

In face to face meeting, gaze is an important non-verbal communication cue with functions such as establishing relationships (through mutual gaze), regulating the course of interaction, expressing intimacy, and exercising social control [22, 5]. A speaker's gaze often correlates with his addressees, i.e. the intended recipients of the speech [18]. Also, for a listener, monitoring his own gaze in concordance with the speaker's gaze is a way to find appropriate time windows for speaker turn requests [10, 27]. Thus, realtime feedback about the meeting participants gaze activities can positively affect the participants behaviour, improve group cohesiveness, and participants satisfaction wich can lead to more efficient meetings [21]. When conversation occur through a mediated device, which is the case of remote meeting situation, a person engaged in a distant discussion with a group of co-located people, does not perceive the non-verbal communication signals (specially the gaze) indicating the reactions of meeting participants to propositions and comments [20, 38]. As shown in [38], in negotiation meeting, the lack of perception of the non-verbal signal, because of the use of mediated device, leads to a misunderstanding of the other people's priorities, and delays agreement achievements. Thus, in video conferencing situation, investigations have been conducted toward the enhancement of gaze perception [25, 29].

Gaze estimation requires either invasive head mounted sensor or accurate estimates of the eye features [26]. In meeting rooms where high resolution eye images are not available, Stiefelhagen *et al* [42] proposes to estimate people's gaze denominated visual focus of attention (VFOA) from their head poses. The study in [42] considered the estimation of people's gaze in meetings with four participants in which, for each meeting participant, only the 3 other participants were the potential visual targets. Stiefelhagen *et al* [42] proposes to model a person VFOA hidden state as a Gaussian mixture model (GMM) with observations the person's head pose. The meeting participants' speaking status (speaking or not speaking) was used to set priors on visual targets to be the focus of the person. Using head poses as observation was based on the assumption that people's head pose is highly correlated with their gaze. Using people's speaking status as priors models people's tendencies to focus at speakers more than at non speakers. Generalizing [42], Otsuka *et al* [33] propose a model for temporal conversation structure that jointly take into account the gazes of all the meeting participants and the conversational events. Similarly to [42] head pose was used as observation for VFOA hidden state, however speaking status was not used directly as prior but as observation for the conversational events hidden states. With the conversational events state influencing the VFOA state dynamics. Similar four participants meetings with the only visual targets being the meeting participants as in [42] was also considered.

Nowadays, most of the meeting rooms are equipped with tables to support laptops, and projection screen for presentations. More general meeting scenarios than the one in which the studies [42, 33] can be considered to conduct gaze behaviour studies. In general meeting scenarios, people's VFOA targets are not restricted to the meeting participants, but can be the table when people are manupulating their laptops, the projection screen when people are gazing at information on the projection screen. VFOA studies in these scenario is more complex because there is more potential visual targets (five instead of three). The increase of the number of visual targets leads an increase of the ambiguities between the head poses defining the targets [4]. Similar head poses can be used to focus at different targets. The general scenario comprising visual targets that are not only the meeting participant is also more complex because the assumption about the conversation structure, more precisely the relationship between the gazing behaviours and the speaking patterns, is no longer valid [8]. The way people gaze at a speaker when there is no new information on the projection screen, is not the same than the way people gaze at a speaker right after a new slide is displayed on the projection screen.

In this paper we present investigations about VFOA recognition in general meeting scenario comprising a projection screen and a table as visual targets. The method we propose for recognition is based on a complete modeling the conversation structure taking into account the presence of the projection screen. We propose to perform the recognition of all participants VFOA jointly together with the meeting conversational events by the use of interaction models that exploit contextual cues relating to group activity or group communication properties. A first property is that the meeting context has a strong effects on the visual focus of attention and the conversational events. For example, depending on the time elapsed since the last slide change focusing at the projection screen is more or less probable. Also, certains type of conversational events are more probable. If a slide change has just occured it is more likely to be in a monologue situation, namely one of the meeting participant is making a presentation. A second property is the dependency of people's attention to conversational events and contextual situation. For example during a monologue, it is more probable that the listenners focus at a speaker than at a non speaker. Also, depending on the activity of the slide focusing on the slide can be more or less probable than focusing at a speaker. The dynamic Bayesian network (DBN) we propose includes these properties about interaction in groups by defining as contextual information a projection screen activity variable storing the time that has elapsed since the last slide change. This slide activity variable is an input the our DBN. The state space is a mixed state constituted by two main components. A first component which models the joint focus of all the meeting participants. The second component models the conversational events. Explicit dependencies between the components of the state variable and the input variables are learned. Head poses are observed for the VFOA states inference and speaking time proportions are observed for the conversational events. Two nature of head pose informations are used. Our VFOA recognition method is evaluated using the AMI Corpus dataset [7] which is a publicly available dataset.

The novelties of our research are the following:

- we propose a slide activity variable to introduce contextual information in DBN for meeting activity analysis.
- we evaluate our model on a publicly available database allowing for performances comparisons
- we analyze the effects of each contextual information on the VFOA recognition performance

The remainder of this paper is organized as follows. Section 2 discusses investigations that are related to the research conducted in this paper. Section 3 describes the task as well as the data used for evaluation. Section 4 presents the way the observations used in our modeling are extracted. Section 6 describes the architecture of the DBN modeling the multi-person VFOA and conversational events using head pose. Section 7 gives the way the parameters of the model are set and how inference is conducted to estimate the optimal sequence of VFOA and conversational events given the sequence of observations. Section 8 presents the procedure used to evaluate the model we propose and the results of the experiments, and Section 9 gives conclusions.

2 Related Work

The topic of this paper is about mutlti-person VFOA and conversational event estimation in meetings. It relates to research about automatic recognition of small groups human interaction in face to face meetings. Along the same line, without being exhaustive, studies have been conducted in computer science to analyze or estimate meeting actions, floor control, dominance, gazing (focus of attention), addressing. Efforts are also devoted towards the estimation of human activities in offices [30], lecture rooms, or video surveillance situations [31]. Althought very interesting, we considere these investigations to be beyond the scope of this paper. A review of computer based human behavior recognition can be found in [35]. In this paper, we focus on human behavior recognition in meetings. First we present research about meeting interactions recognition, then we review research about visual focus of attention estimation.

2.1 Meeting interaction recognition

[12] gives an overview about the research conducted towards the analysis of group interaction in multi-party face to face conversations. [23] presents investigations about the automatic recognition of group actions in meetings. The objective was during a meeting to recognize whether during a meeting a person was seated making a monologue, a standing making a presentation at the projection screen, was standing at the white board, or people were picking notes. The method was based on first recognizing the single person action using audio visual features (speech activity, pitch, speaking rate, and head and hand blobs), and from the single actions method recognize people's interactions using hidden Markov models (HMMs). The proposed model was evaluated on the M4 dataset.
Other researchers concentrated on studying in meeting not a broad class of group action but on a single type of action. Dominance is one of those actions. Dominance is a high level concept related to influence and leadership. Using audio features only, [39] attempted successfuly to estimate people's dominance level from easily detectable features feeded to a support vector machine (SVM) classifier trained to attribute dominance ranks to the meeting participants. Following similar vein of research, [17] proposed to use an unsupervised model using non verbal audio visual features to find the most dominant person in a meeting. Both [39] and [17] used the AMI corpus database [7] which is publicly available for research use.
Another meeting conversation event that has been studied is floor control. [8] conducted an analysis of floor control in meetings use the VACE meeting corpus database [9]. This study showed that verbal cues such as verbal backchannels, discourses markers specifying discourses boudaries (beginning, end), and visual cues such as gaze, non verbal backchannels (nodding) are strongly correlated to floor control.
In a conversation, an addressee is the person to whom the speaker utterance is intended. The act of addressing or being addressed is also a multi modal phenomenon. From the speakers utterance and gaze, insight are given about his intended addressee. Listenners express their attention, by gazing at the speakers and sending backchannel whenever required. Investigation about addressee detection in face to face meeting has been conducted in [19] using the M4 and the AMI corpus databases. In [19] gaze, utterance and conversational context features were feeded in a dynamic Bayesian network and naive Bayesian classifier to predict people addressee.
Following the review about conversational group activities, a commom aspect of the investigations is that models such as HMM or DBN are very popular because they allow to take into account the group interaction dynamics and the temporal aspect of the interactions. Another common shared aspect is that the models involved multiple person and are clearly multi-modal. Although speech is the main communication channel, interaction in group is also made through the visual channel with body posture, gestures, gaze. Thus a complete study of interaction in group has to be done by processing the audio visual information produced by the group. Also focus of attention (gaze) has been found to be one of the important visual cues for understanding group interactions.

2.2 Visual focus of attention recognition in meetings

More closely related to our study are the research about recognizing people's visual focus of attention in meetings. Stricly speaking, estimating people's gaze requires either invasive head mounted sensors or high resolution people's face image. In the context of meetings where people are not recorded from close up view cameras, high resolution image of the face are not available. However research conducted about VFOA estimation showed that in four person meeting setting in which, for a person the only potential VFOA targets are the other people, gaze can be reliabily estimated from head orientation

Figure 1: Meeting recording setup. The first image shows the central view that is used for slide change detection. The second and the third image show the side camera that are used for people's head pose tracking. The last image shows a top view of our the meeting room to illustrate seats and focus target positions. Seat numbers will be used to report the VFOA recognition results of people seated at these locations.

[42]. To estimate a person's focus, [42] used head pose as observation for a Gaussian mixture model (GMM) and audio cues defining priors on people to be the focus target. Using audio priors allowed to model people's tendencies to focus on speakers. [41] when attempting to estimate people's addressee in meetings proposed to consider as the focus for a person, the other persons lying in his field of view. The head pose of the person was estimates and used to define his field of view. A limitation of the method proposed in [41] is that the VFOA of a person is not defined by his head pose but his eye gaze. However when annotating evaluation data, the head direction was assimilated to the focus of attention. Another limitation of [41], and also [42], is that the focus of attention of a person was estimated independently of the focus of the other person while people focus are correlated. [33] addressed this limitation by proposing a probabilistic framework to jointly recognize the gaze pattern of four meeting participants together with their conversational regime with a Markov switching model using as observation the four people head poses and and their speaking status. Both [42, 33] assumed the setup to be a four persons meeting in which the potential VFOA targets are the four meeting particpant. This assumption left aside a whole range of meeting setup in which people have other potential VFOA targets. Because the introduction of other VFOA targets can significantly change people's gazing behavior, generalizing the current state of the art VFOA recognition model to scenarii where there are other focus targets than the people is not straightforward. The research presented in this paper has as one of its goal to conduct investigations towards this generalization.

3 Dataset and Task

In this Section, we describe the dataset that we used to conduct our research. We the define precisely the main task of the paper (VFOA estimation), describe the ground truth annotation procedure. We then provide global VFOA statistics of the ground truth and comment them in order to provide some insight about the data.

3.1 Dataset description

Our data source is the AMI corpus[1]. In this corpus, recorded meetings were following a general task-oriented scenario. More precisely, four people with different roles (project manager, marketing expert, user interface and system designers) were involved in the creation and design of a new television remote control. The different phases of the design process were discussed in a serie of four meetings, where the different participants would present their contributions according to their expertise, discuss design alternatives, market issues, and take decisions. During these meetings, people were behaving naturally, taking notes, using laptops, making presentation with slides projected on a screen, and possibly manipulating a prototype of the remote control.

[1]www.idiap.ch/mmm/corpora/ami

Fig. 1 shows the physical set-up of the meeting room. Amongst the different sensor recordings which are available in the corpus, we used the following: the video streams from the two cameras facing the participants (Fig. 1 center images) and of a third camera capturing a general view of the meeting room (Fig. 1, left). As audio sensor, we used the close-talk microphones.

Our database consists of four meetings from the above AMI corpus, for which VFOA annotation was available and in which people were remaining seated during the entire meeting. The meeting durations ranged from 15min to 27min, for a total of 1h30min. In total, twelve different people were featured in these meetings, making the head pose tracking a challenging task.

3.2 VFOA recognition Task and Data analysis

Our main objective is to estimate people's VFOA in meeting scenarios such as described above. Although in principle VFOA is given by a 3D eye gaze direction, studies about gaze [16] have shown that humans tend to look at specific targets (location/objects, humans) which are relevant to the task they are solving or of immediate interest to them. Thus, we have defined the set of interesting visual targets for a given participant seated at seat k, denoted $\mathcal{F}_k$, as comprising 6 VFOA targets: the 3 other participants $\mathcal{P}_k$ (e.g. for seat 1, $\mathcal{P}_1 =$\{seat2,seat3,seat4\}), as well other targets $\mathcal{O} =$\{Table, Slide Screen, Unfocused \}. The later target (Unfocused) is used when the person is not visually focusing on any of the previously cited targets.

Data annotation analysis

The meeting participants' VFOA were annotated based on the set of VFOA labels defined above. Annotators used a multimedia interface, with access to the sound recordings and all camera views, including close-up cameras. The later happened to be useful in some ambiguous cases (esp. when watching people with glasses).

Fig 2 gives the VFOA statistics, where we have grouped the VFOA labels corresponding to participants into a single label 'people'. As can be seen, looking at people only represents 45% of the data, while looking at table or slide represents a significant proportion of the VFOA. The label 'Table' corresponds to two main situations i) when people use their laptop or look at a remote control prototype placed on the table (this happens in one meeting) ii) when people look downwards without actually changing their head pose while listening to a speaker. In particular, situation ii) has been found to have increased w.r.t. our previous study on 7 to 10min long meetings [4]. These VFOA statistics contrast with other setups and places our work in a different context than studies investigating VFOA estimation [42, 32]. In these studies, the scenarios are reduced to a discussion between meeting participants and the meeting duration is shorter (e.g. around 8min in [32]). While already challenging, these scenarios reduce the chances to observe natural behavior such as the behavior ii) mentioned above, and the interactions between people are more simple due to the absence of artifacts or contextual objects which can attract the visual attention and modify the gazing pattern of participants. In addition, in [42, 32], the targets are contrived to be only other meeting participants. As some targets are more difficult to recognize than others, this difference will have effects on the performance. This is indeed the case for the label 'Table' due at least to the situation ii) described above, and to the fact that, in contrast to [42, 32], we can no longer rely only on the head pan, but also need to use the head tilt -which is known to be more difficult to estimate from images- to distinguish different VFOA targets.

4 Audio-visual features for VFOA modeling

Three types of features were used in our model: audio features describing people's speaking status, the projection screen activity features, and the head pose features characterizing people head orientation. In the following we describe how these features were extracted.

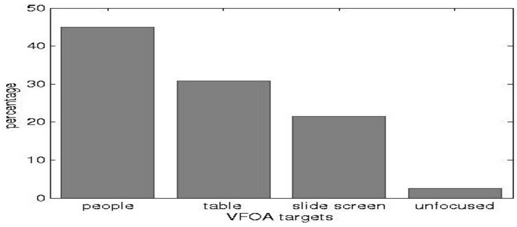


Figure 2: Distribution of VFOA labels in the AMI Corpus.

4.1 Audio Features

The Audio features are estimated from close-talk microphones attached to each meeting participants. At each time step, real value speaker energy defined as the root mean square amplitude of the audio signal over a window of 40 milliseconds (ms) with a 10 ms time shift, is extracted for each participant k. At each time step t, the speaker energy is thresholded, as illustrated in Fig 3(a), to give the speaking status s_t^k which value is 1 if participant k is speaking and 0 otherwise. To model the conversation, we are interested in knowing who are the current main speakers. To obtain more stable features, we smoothed the instantaneous speaking status by averaging them over a temporal window of W frames [2]. Our audio feature for person k is thus:

$$\tilde{s}_t^k = \frac{1}{W} \sum_{l=-\frac{W-1}{2}}^{\frac{W-1}{2}} s_{t+l}^k \tag{1}$$

that is, the proportion of time person k is speaking.

4.2 The projection screen activity

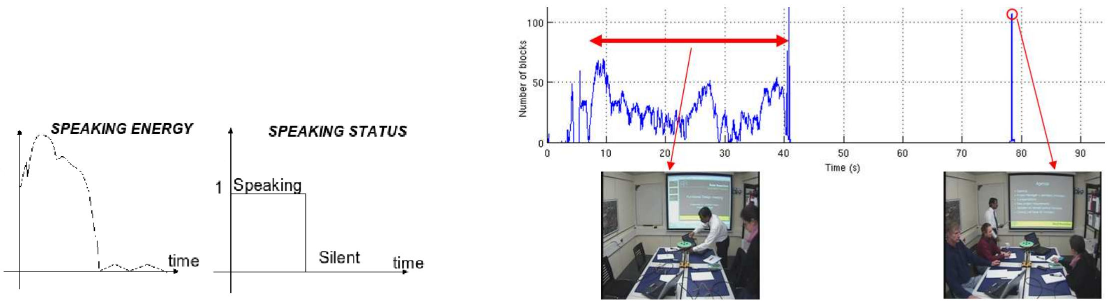
In our scenarios, people are discussing slides projected on a screen and presented by one of the participant. As we have described earlier, the presentation of these slides will have an important effect on the conversation behaviours, and especially on the gaze patterns. Intuitively, the effect will be more important when the slide has just been displayed, and will decrease as time goes by. Thus, we decided to use as slide activity feature a_t the time that ellapsed since the last slide change occurred [3]. We thus need to detect the instants of slide changes, from which deriving a_t is straightforward.

Slide changes are detected by processing the video stream from the central view camera (see Fig 3(b)). We exploited a compressed domain method proposed by Yeo *et al* [43]. The method relies on the residual coding bit-rate, which is readily extracted from compressed video recordings. This residual coding bit rate captures the temporal changes which are not accounted for by the block translational model. In the case of slide transitions, there is no translation involved, yet there are very distincts frame differences which turn out to be highly correlated with the residual coding bit-rate. Thus, the approach we used consisted simply in thresholding the number of blocks in the slide area of the image which have a sufficiently high residual coding bit-rate. In addition to the block count thresholding, a non-maximal suppression step over 2 second windows was applied, and a minimum difference between the peak value and the average number of detected blocks over one second windows on both sides of the peak was enforced.

Figure 3(b) illustrates how the approach can handle people walking in front of the projection screen. First, translational motion of people is somehow accounted for by the block translational component,

[2] The temporal window W is taken to be odd to allow for a window symetric with respect to its center.

[3] A slide change is defined as anything that modifies the slide-screee display. This can correspond to full slide transitions, but also to partial slide changes, or to switch between a presentation and some other content (e.g. the computer desktop).



(a) Binarization of real values of speaking energy.

(b) Slide change detection example.

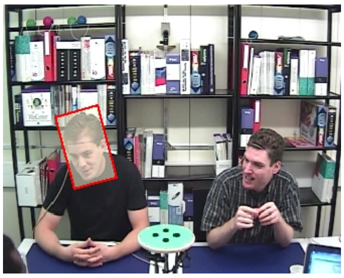
Figure 3: Fig. 3(a) Audio observations. Fig. 3(b) Illustration of slide change detection based on thresholding the count of the number of blocks with high residual bit rate while handling changes due to people passing in front of the projection screen.

thus reducing the number of high residual bit-rate blocks detected. Second, as shown in Fig. 3(b), when there is a person walking in front of the projection screen, there is a high number of detected blocks over an extended period of time. In contrast, for a slide transition, the block counts exhibit sharper peaks. The postprocessing steps described above account for this behaviour. Evaluated on 12 meetings, this method provided a slide change detection recall of 90% with a precision of 95%.

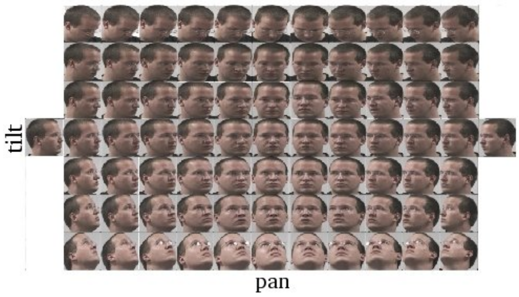
4.3 The head poses

To estimate the head pose, we used an improved version of the tracking method described in [3]. It relies on the Bayesian formulation of the tracking problem. Denoting the head configuration state at time t by X_t and the observations by Y_t, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of the state given the observation sequence $Y_{1:t} = (Y_1, \ldots, Y_t)$. In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters (PF), which consists of representing the filtering distribution using a set of N_s weighted samples (particles) $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrives. In [3], we applied such a framework to the joint tracking of the head location and pose.
More precisely, the state space contains both continuous variables L_t and a discrete variable θ_t. L_t represents the head location, vertical and horizontal scales, and an in-plane rotation that allows to localize the head in the image, as illustrated in Fig.4(a). The discrete index $\theta_t \in \Theta$ denotes an element of the discretized set of possible out-of-plane head poses shown in Fig. 4(b).

As image observations Y, we used texture (output of one Gaussian and two Gabor filters) and skin color features (plus background subtraction features, see below) computed at predefined sample locations from image patches extracted from the image and preprocessed by histogram equalization. Fig 5 shows the features we used for head pose tracking. For each element of the discrete pose space Θ, a texture and color appearance lieklihood model was learned using the images of the Prima-Pointing database [15], which contains 15 individuals recorded over 91 different poses. These pose-dependant likelihood models were used to evaluate the likelihood of texture and skin color observations given the state values.
Two main improvements w.r.t. the method in [3] were used. First, an additional silhouette likelihood term was used. It was based on the correlation between a background subtraction map and a silhouette mask. While this likelihood model was not pose-dependent, it helped in providing better localization information and to reduce tracking failures. In addition, in addition to the state dynamics, we also exploited the output of a head detector to propose new sample locations in the current frame given past particles. This detector was useful to automatically (re)initialize the tracker and avoid failures.



(a) Head localization (red bounding box) parameterized by the variable L_t



(b) Discretized head pose space Θ (sample images from the Prima-Pointing database [15])

Figure 4: Head localization and pose space.

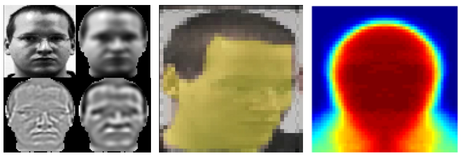


Figure 5: Texture, skin color and silhouette observations for tracking.

One specificity of the approach in [3] was to use a Rao-Blackwellization approach to increase the sampling efficiency, which results in a reduction of the number of samples for similar tracking performance. The Rao-Blackwellized particle filter (RBPF) consists of applying the standard PF algorithm to the continuous variables L while applying an exact filtering step over the pose exemplar variable θ, *given a sample of the localization variables*. In this way, the posterior distribution of the state given the observations can be written as:

$$p(L_{1:t}, \theta_{1:t}|Y_{1:t}) \quad = \quad p(\theta_{1:t}|L_{1:t}, Y_{1:t})p(L_{1:t}|Y_{1:t}) \tag{2}$$

In practice, only the sufficient statistics $p(\theta_t|L_{1:t}, Y_{1:t})$ of the first term in the right hand (RHS) side is computed and is involved in the PF steps of the second term of the RHS. Thus, in the RBPF modeling, the probability density function (PDF) in Equation 2 is represented by a set of particles

$$\{L^i_{1:t}, \pi^i_t, w^i_t\}_{i=1}^{N_s} \tag{3}$$

where $\pi^i_t(\theta) = p(\theta|L^i_{1:t}, Y_{1:t})$ is the pdf of the pose exemplars $\theta \in \Theta$ given a sequence of localizations and a sequence of measurements, and $w^i_t \propto p(Y_t|L^i_t)$ is the weight of the particle estimated through the PF approach. Given the state samples, we computed the head pose output by the tracker as the mean of the particles as $\hat{\theta}_t = \sum_{\theta\in\Theta} \theta(\sum_{i=1}^{N_s} w^i_t \pi^i_t(\theta))$.

5 Model Overview

In this Section, we provide a global overview of our model. We first define the conversational event variables which we have introduced to make the link between people's VFOA and their speaking status, and then present overviews of the probabilistic model as well as of our system.

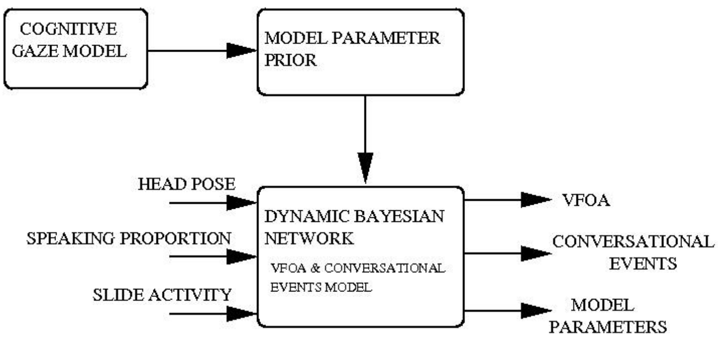


Figure 6: Overview of the model proposed for VFOA and conversational event estimation.

5.1 Conversationel events

To model the interactions between people's VFOA and their speaking status, we propose to introduce a new variable that provide a status on the current meeting conversation structure. When only considering speaking patterns, a natural and common way to describe the conversation status is to define "floor holding" patterns. Basu *et al* [6] followed this approach, by introducing variables that indicate whether the floor was dominated by a single person (e.g. in a monologue) or not, and who was the dominating person, in two-people interactions. When considering the gaze, Otsuka *et al* [32] showed that the people joint gaze patterns in group conversation could be clustered into gaze "convergence" patterns, in which people mainly looks at the speaker, and gaze "dyadic" patterns, in which people mainly look at the two persons involved in a dyadic discussion. Otsuka *et al*'s definition of gaze patterns made the link between people's gaze and their speaking status of people [32].

In this paper, we follow a similar approach, and introduce a set $\mathcal{E} = \{E_i, i \in 1...16\}$ of 'conversational events' characterizing the floor holding status, i.e. who are the persons currently holding the floor. Since we are dealing with meetings of 4 people, the set comprises 16 events which can be divided into 4 types: $\mathcal{T}(E_i) \in \{silence, monologue, dialogue, discussion\}$. where $\mathcal{T}(E_i)$ denotes the conversation type of the event E_i. In addition, the set $\mathcal{I}(E_i)$, indicates the people actively involved in the event, e.g., for a monologue, the main speaking person.

5.2 System overview

A global representation of our system is provided in Fig. 6. The main part is composed of a DBN model which describes the relationship between all random variables (cf next Subsection). More precisely, the model takes as input a set of observation variables as well as the set of all DBN model parameters λ, from which the set of all hidden variables is inferred, and an estimate of λ is given. The observation variables comprise the head poses observations for all participants $o_t = (o_t^1, o_t^2, o_t^3, o_t^4)$, their speaking proportion $\tilde{s}_t = (\tilde{s}_t^1, \tilde{s}_t^2, \tilde{s}_t^3, \tilde{s}_t^4)$, and the slide-screen activity variable a_t, as described in Section 4. The hidden variables comprise the joint VFOA state $f_t = (f_t^1, f_t^2, f_t^3, f_t^4)$ of all participants, where f_t^k denotes the VFOA of participant k at time t, and the conversational event $e_t \in \mathcal{E}$. Importantly, Fig 6 enhances the fact that some prior probability $p(\lambda)$ on the DBN parameters need to be defined. In particular, defining a prior on the parameters relating the people head pose observations to their gaze VFOA target appeared to be crucial to automatically adapt the DBN model to the specific head orientations of individuals in meeting. To set this prior values, we used the cognitive gaze model proposed in [4], as will be described in Section 7.

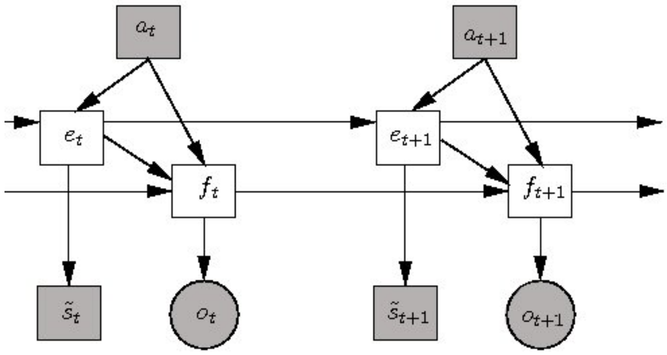


Figure 7: Dynamic Bayesian network graphical model. Squares represent discrete variables, circles represent continuous variables. Observation are shaded, hidden variable are unshaded.

5.3 Joint multi-party VFOA modeling with a DBN

To estimate the joint VFOA of people, we rely on the DBN model displayed in Fig. 7, and according to which the joint distribution of the variables is given by:

$$p(f_{1:T}, e_{1:T}, \lambda, o_{1:T}, \tilde{s}_{1:T}, a_{1:T}) \propto p(\lambda) \prod_{t=1}^{T} p(o_t|f_t)p(\tilde{s}_t|e_t)p(f_t|f_{t-1}, e_t, a_t)p(e_t|e_{t-1}, a_t). \quad (4)$$

where λ denotes the set of model parameters. All the terms appearing in the Eq 4 are defined in Section 6. From a qualitative point of view, the DBN expresses the probabilistic relationships between our random variables and reflects the assumptions we made on the way random variables are generated and their stochastic dependencies. In the current case, an important assumption is that the conversational event sequence is the primary hidden process that governs the dynamics of gaze and speaking patterns, i.e. gaze patterns and people utterance are independent given the conversational events. Indeed, given a conversational event, speakers are clearly defined and speaking patterns can be thus sampled according to the $p(\tilde{s}_t|e_t)$ term. At the same time, as people tend to look at the current speakers [32], the conversational event will also directly influence the dynamics of the people's gaze through the term $p(f_t|f_{t-1}, e_t, a_t)$, which increases the prior probability of looking at the VFOA targets corresponding to the people who are actively involved in the current conversational event. However, people do not always gaze at the current speaker. Argyle [2] has shown that objects that plays a central role in a task that people are doing attracts the visual attention, thereby overruling the trends for eye gaze behaviour observed in 'direct' human-human conversations. This is the case of the slide-screen in our meetings: during presentations, people often look at the projection screen rather than at the speaker, specially right after a slide change. One novelty of our approach is to take into account the impact of this activity on the gaze pattern through $p(f_t|f_{t-1}, e_t, a_t)$, by modulating the prior on looking at the current speaker(s) as a function of the duration a_t since the last slide change occurence.
In addition, the model assumes that people head pose are only dependent on the current gaze target (term $p(o_t|f_t)$) and temporal information is introduced in the system to favor an overall smoothness, as detailed in the following section.

6 Joint multi-party VFOA modeling with a DBN

We present below the different probabilistic terms involved in the DBN of Fig. 7 and defining the posterior density function in Eq 4. We present below the conversational events dynamics $p(e_t|e_{t-1}, a_t)$, the VFOA dynamics $p(f_t|f_{t-1}, e_t, a_t)$, the the observation models for the conversational event $p(\tilde{s}_t|e_t)$ and for the VFOA $p(o_t|f_t)$, and $p(\lambda)$, the prior distribution on the model parameters.

6.1 Conversational events dynamics

We assumed the following factorized form for the conversational events dynamics:

$$p(e_t|e_{t-1}, a_t) = p(e_t|e_{t-1})p(e_t|a_t). \tag{5}$$

The first term in this Eq 5, $p(e_t|e_{t-1})$, models the conversational event temporal transitions, and was defined to enforce temporal smoothness. We modeled this term assuming that some probability mass to remain in the same state $p(e_t = E_i|e_{t-1} = E_i) = p_e$, and spreading the remaining probability mass for the transitions to the other events $p(e_t = E_j|e_{t-1} = E_i) = \frac{1-p_e}{15}$. The second term in Eq. 5 models the prior knowledge about the conversational events given the slide activity. So not to overfit specific meeting situation, we assumed that this prior is only depedent on the event types, and is learned through counting after discretization of the a_t variable. Fig 8(b) gives the plot of the priors learned from our dataset using all the meetings. As can be seen this figure, monologues and dialogs are more frequent then silences and group discussions. Overall, this curves do not show a great dependency of the event probability w.r.t. a_t, except for silence which are much more probable right after and long after a slide change, conversely to monologues which are less frequent right after a slide change.

6.2 The VFOA dynamics

We assume the following factorized form for the VFOA dynamics:

$$p(f_t|f_{t-1}, e_t, a_t) \propto \Phi(f_t)p(f_t|f_{t-1})p(f_t|a_t, e_t) \tag{6}$$

where $\Phi(f_t)$ is a distribution modeling the probability of people sharing the same focus targets, $p(f_t|f_{t-1})$ models the temporal transitions between VFOA states, $p(f_t|a_t, e_t)$ models the probability to observe a joint VFOA state given a meeting context defined by the slide screen activity. The factorization made in Eq. 6 is based on the assumption that the group prior $\Phi(f_t)$ models all the dependencies between the VFOA of meeting participants, while the other terms only model the effect of the conditional variable on the current focus. Below, we describe all the terms defining the VFOA dynamics.

6.2.1 The shared focus prior $\Phi(f_t)$

This term models people's inclination to share VFOA targets. Fig. 8(a)) depicts the distribution of frames w.r.t. the number of people that share the same focus. As can be seen in this figure, people are sharing more often the same focus than if their focus were independent. Thus, we have set $\Phi(f_t)$ as:

$$\Phi(f_t) = \Phi(SF(f_t) = n) \propto \frac{d_n}{c_n} \tag{7}$$

where $SF(f_t)$ denotes the number of people that share the same focus in the joint state f_t, and d_n and c_n are defined in Fig. 8(a). Qualitatively, the shared focus prior will favor states with shared focus according to the distribution observed on training data.

6.2.2 The joint VFOA temporal transition $p(f_t|f_{t-1})$

is modeled assuming the independence of people's VFOA states given their previous focus [4]:

$$p(f_t|f_{t-1}) = \prod_{k=1}^{4} p(f_t^k|f_{t-1}^k). \tag{8}$$

[4]The dependencies between people's focus is assumed to be modeled by the joint focus prior.

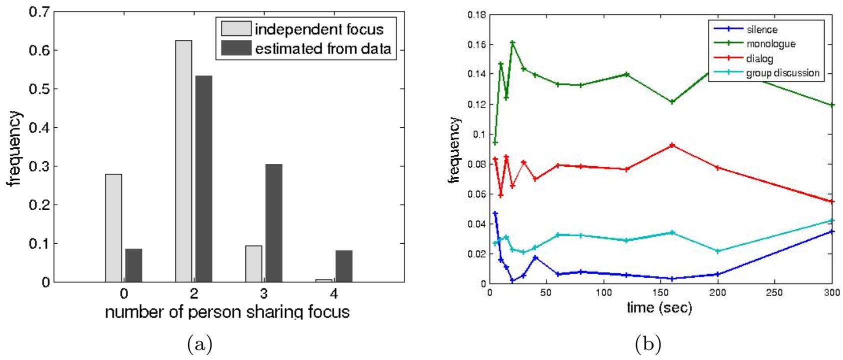


Figure 8: Left: Shared focus. Distribution of frames where n persons are focused on the same VFOA target. Light bars: distribution c_n assuming people VFOA are sampled independently from the marginal label distribution of Fig. 2. Dark bars: distribution d_n measured on the data. Right: frequency of meeting event given the time since the last slide change.

The VFOA dynamics of a person k is modeled by a transition table $B^k = (b_{k,i,j})$, with $b_{k,i,j} = p(f_t^k = j|f_{t-1}^k = i)$. These tables are automatically adapted during inference (cf Subsection 6.4), with prior values defined to enforce some temporal smoothness, with high probability to remain in the same state and lower probability to transit to other focus.

6.2.3 The VFOA meeting contextual priors $p(f_t|a_t, e_t)$

The introduction of this term in the modeling represents one of the main contribution of the paper. It models the prior of focusing at VFOA targets given the meeting context (conversational event, slide activity). To define this term, we first assume the conditional independance of people VFOA given the context:

$$p(f_t|a_t, e_t) \propto \prod_{k=1}^{4} p(f_t^k|a_t, e_t). \tag{9}$$

We then have to define the prior model $p(f_t^k = l|a_t = a, e_t = e)$ of any participant. Intuitively, this term should establish a compromise between the known properties that: (i) people tend to look at speaker(s) (ii) during presentations, people look at the projection screen (iii) speaker's gazing behaviour might be different than listener's one. Thus, to follow this intuition, we introduce the following notations: $ki(k, e)$ is a function that maps into $\{inv, not\_inv\}$ and which defines whether participant k is involved or not in the conversational event e; $i(l, e)$ is a function that maps a VFOA target l to its type in $\mathcal{FT} = \{slide, table, unfocus, inv, not\_inv\}$, in function of the event e. Defining i is straightforward: $i(l, e) = l$ if $l \in \{slide, table, unfocus\}$, and $i(l, e) = ki(l, e)$ if l is participant target. Then, we define (and learn from training data) the table $T(i, ki, t, a) = p(i|t, a, ki)$ providing the probability, for a participant whose involvment status in a conversational event of type t (i.e either silence, monologue, dialog, discussion) is ki, of looking at a VFOA target type i (either an environmental target or a participants involved -i.e. speaking- or not -i.e a side participant- in the event) given the event type t and the time a that elapsed since the last slide change. Then, the prior model is simply defined as: $p(f_t^k = l|a_t = a, e_t = e) = T(i(l, e), ki(k, e), t(e), a)$.
The table T can be learned directly by counting the corresponding configuration occurences in a training data set and normalizing appropriately.

To obtain smoother versions of the contextual priors, we fitted using gradient descent a function of the form:

$$T(i, ki, t, a) \approx g_{i,ki,t}(a) = \vartheta_1 e^{-\vartheta_2 a_t} + \vartheta_3. \tag{10}$$

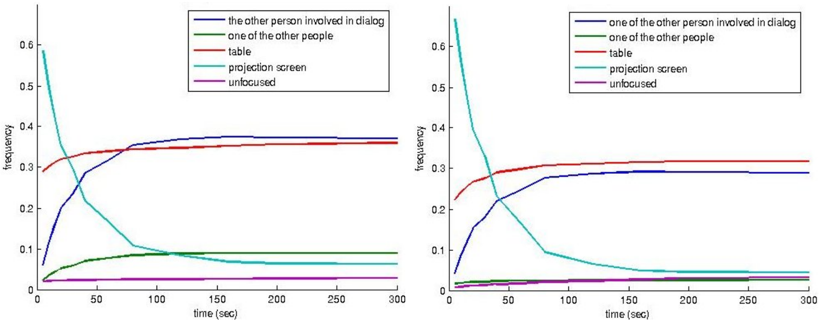


Figure 9: Fitted contextual prior probabilities of focusing to a given target, in function of time a since last slide change, when the conversational event is a dialog, and for: Left: a person involved in the dialog. Right: a person not involved in the dialog.

to the tables learned from training data. In practice, we noticed that this family of exponential functions, depending on the parameters $\vartheta_i,\ i = 1, ..., 3$ (one set for each configuration (i, ki, t)) were providing a good approximation to observed exponential increase or decrease of probability in function of the ellapsed time a.

Fig 9 gives provides an interesting example of the fitted priors when the conversational event is of type $t = dialog$. For the table target, we can see that its probability is always high and not so dependent on the slide context a. However, whether the person is involved or not in the dialog, the probability of looking at the projection screen right after a slide change is very high, and steadily decreases as the time a since last slide change increases. A reverse effect is observed when looking at people: right after a slide change, the probability is low, but this probability keeps increasing as the time a increases as well. As could be expected, we can notice that the probability of looking at the people involved in the dialog is much higher than looking at the side participants. For the later target, we can notice a different gazing behaviour depending on whether the person is involved in the dialog or not: people involved in the dialog focus at the side participants, a fact which is known in the social literature, whereas a side participant almost never looks at the other side participants.

6.3 Observation models

Two models need to be defined and are successively described. First, the audio model describing the speaking observation probabilities for the different conversational events. Second, the probability of observing a given head pose given the VFOA target.

6.3.1 The speaking proportion observation model

This term $p(\tilde{s}_t|e_t)$ was defined as:

$$p(\tilde{s}_t|e_t = E_j) = \prod_{k=1}^{4} \mathcal{B}(\tilde{s}_t^k; \eta_{k,j}, 1 - \eta_{k,j}) \tag{11}$$

were we have assumed that people's speaking proportion $\tilde{s}_t^k$ were independent given the event E_j, and $\mathcal{B}(\tilde{s}_t^k, \eta_{k,j}, 1 - \eta_{k,j})$ is a Beta distribution with parameters $\eta_{k,j}$ and $1 - \eta_{k,j}$ defined as:

$$\mathcal{B}(x, p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}. \tag{12}$$

The parameter $\eta_{k,j}$ is the probability that person k speaks during the W frames defining the event E_j. The Beta distribution is the standard distribution to model probablities on proportion values.

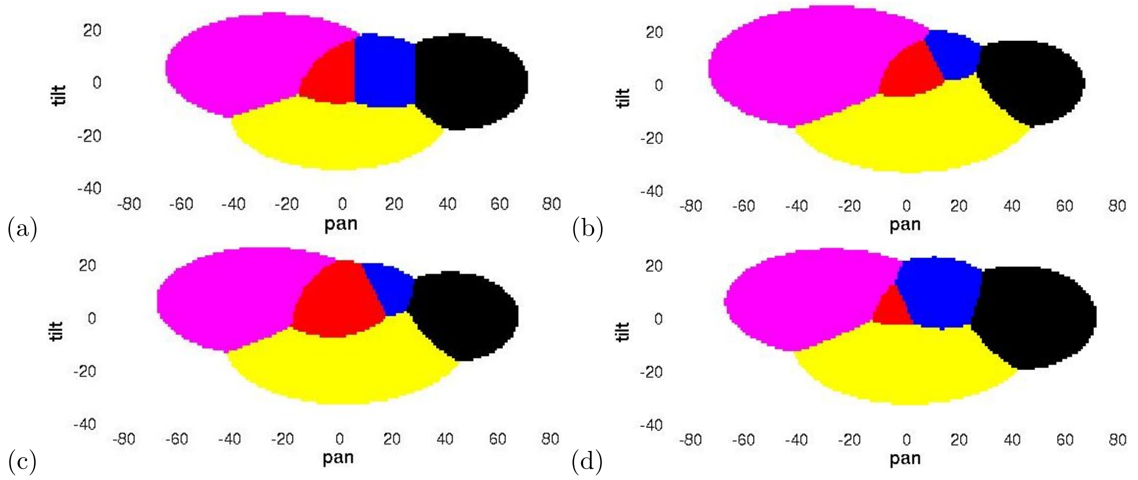


Figure 10: Qualitative effects of the contextual priors on the VFOA decision maps for a person sitting at seat 2. Each image indicates through color labels (see Fig.) the VFOA target that maximizes the probability $p(o_t^k|f_t^k)p(f_t^k|e_t, a_t)$ for the corresponding head pose. In (a), decision map when using only the observation model $p(o_t^k|f_t^k)$. In (b), the context is silence, and a slide change just occured (a_t is low), which results in an emphasis of the slide screen (magenta area is bigger). In (c), a monologue is performed by the person at seat 1 (bigger red area)and a_t is large. In (d), a dialog occurs between the participants at seat 3 (blue area) and seat 4 (black area).

6.3.2 The head pose observation model

This term $p(o_t|f_t)$ is the most important term for gaze estimation, since people VFOA is primarily defined by their head pose. Assuming that given their VFOA state, people head poses (defined by a pan and tilt angles) are independent of each other, this term can be factorized as $p(o_t|f_t) = \prod_{k=1}^{4} p(o_t^k|f_t^k)$. Then we use the same model as in our previous work [4] to model the observation model of individual people. For regular VFOA targets (which are not unfocused) the head pose distribution is modelled as a Gaussian distribution to account for the head pose variability when looking at visual targets:

$$p(o_t^k|f_t^k = i) = \mathcal{N}(o_t^k, \mu_{k,i}, \Sigma_{k,i}). \tag{13}$$

where $\mu_{k,i}$ are the mean and covariance $\Sigma_{k,i}$ of the likelihood when person k looks at the target i. For the case when the person is unfocused, we simply model the pose distribution as a uniform distribution $p(o_t^k|f_t^k = \text{ unfocused }) = u$.

The pose observation model can be illustrated by plotting 'decision maps', i.e. by drawing for each pose which gaze target is the most probable. This is shown in Fig. 10, which also illustrate the influence of the context (conversation, slide) on the decision maps.

6.4 Priors on the model parameters

In the Bayesian framework, we can specify some prior knowledge about the model by specifying a distribution $p(\lambda)$ on the model parameters. Then, when the observation are given, during inference, optimal parameters of the model are estimated jointly with the optimal VFOA and conversational event sequences.

Our model contains parameters related to the conversational events (those defining the terms $p(e_t|e_{t-1}), p(e_t|a_t), p(f_t|a_t, e_t)$ and $p(\tilde{s}_t^k|e_t)$), and parameters involving only the VFOA state variable. In this study, we assumed that the first ones were set or learned a priori (as described in previous Subsections) and remained fixed

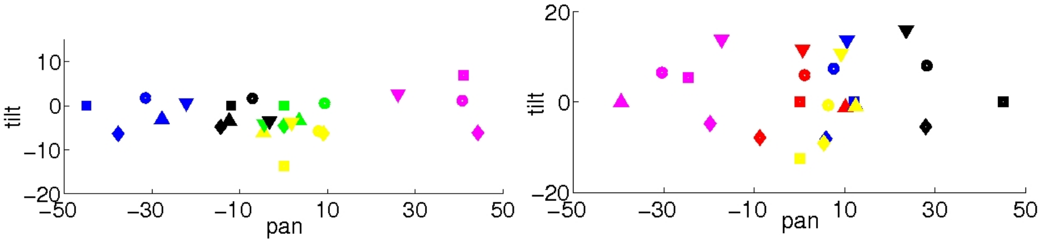


Figure 11: VFOA centers predicted using the cognitive model (square) and the people's data (diamonds, triangles up and down, circles) for people seated at seat 1 (left) and seat 2 (right). This figure illustrate the inter-person variabilitities about head pose used to gaze at visual targets.

during inference, implicitely assuming a Dirac prior distribution over these parameters. The parameters related to the VFOA state are the transition probability tables $B = \{B^k\}_{k=1..4}$ defining the VFOA dynamics of each individuals, and the means $\mu_k = (\mu_{k,i}, i = 1, ..., 5)$ and covariance matrices $\Sigma_k = (\Sigma_{k,i}, i = 1, ..., 5)$ specifying for each person k the Gaussian distribution head pose observation models. Thus, the parameter set which was allowed to vary in order to adapt to the observations during inference was defined as $\lambda = (\lambda_k), k = 1, ..., 4$, with $\lambda_k = (\mu_k, \Sigma_k, B_k)$. In our previous work [4], we showed that allowing adaptation of these parameters was important to improve performance. This is particularly true for the Gausian distribution means $\mu_{k,i}$ representing the mean head pose for person k to look at target i, since we observed large variations between the values computed empirically using the VFOA ground truth data for different people. These variations can be due to either people specific ways of looking at given target (people turn more or less their head to look at targets), or the introduction of bias in the estimated head pose by the head pose tracking system. These variabilities are illustrated in Fig. 11.

We provide below the main elements on how we specify the prior and set the prior parameters. More details can be found in [4]. We used natural conjugate priors for mathematical convenience.

VFOA transition probability prior. For the transition probabilities, this corresponds to using a Dirichlet distribution $\mathcal{D}$ on each row of the transition matrices B^k, defined as

$$\mathcal{D}\left(b_{k,i,1}, ..., b_{k,i,J} | \nu q_{k,i,1}, ..., \nu q_{k,i,J}\right) \propto \prod_{j=1}^{J} b_{k,i,j}^{\nu q_{k,i,j} - 1}.$$

where ν is a scale factor, and $q_{k,i,j}$ are the prior parameter values that we can give to the transition parameters $b_{k,i,j}$. In the experiments, we set $\nu = T$ (where T is the sample size, i.e. the sequence length), so that the contributions to the transition parameter estimation of the prior values and of the data during inference are of equal importance. We defined the prior transition values to enforce smoothness in the VFOA trajectories. Thus, the probability to remain in the same state is favored with respect to transit to another state: $q_{k,i,i} = 0.9$, while the remaining of the probability mass is uniformly spread among the transitions to the remaining VFOA target, i.e. $q_{k,i,j} = \frac{0.1}{J-1}$, for $j \neq i$.

Head pose prior. The conjugate prior for the Gaussian mean $\mu_{k,i}$ and covariance matrix $\Sigma_{k,i}$ is the Normal-Wishart distribution, $\mathcal{W}(\mu_{k,i}, \Sigma_{k,i} | \tau, m_{i,k}, d, V_{i,k})$ [4, 13]. The prior is defined by two types of parameters: the prior values for the mean and covariance ($m_{i,k}$ and $V_{i,k}$)), and scale factors τ and d which specify how much deviation from the prior values are allowed. In the experiments, we used $\tau = \frac{T}{J}$, meaning that equal importance was given to the prior values and the data when estimating the Gaussian mean $\mu_{k,i}$ during inference. We set $d = \frac{5T}{J}$, indicating that in this case not much variations from the prior covariance values was allowed. As for the prior covariances, we used the same values than in our previous work [4], which are taking into account the physical size of the object and the

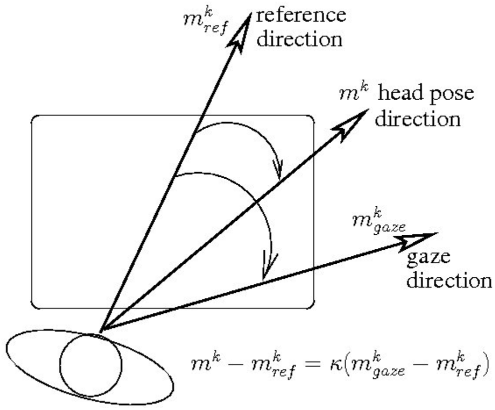


Figure 12: Relationship between the gaze direction associated with a VFOA target and the head pose of a participant. Assuming that the participant has a prefered reference gazing direction (usually in front of him), the gaze rotation towards the VFOA target is made partly by the head and partly by the eye.

head tracking errors (between 12 and 17 degrees for pan and tilt for the standard deviation).

Defining the $m_{k,i}$ using a Cognitive model. To set the prior means $m_{k,i}$, we used our model [28, 4], which relies on findings in cognitive sciences about gaze shift dynamics [11, 16]. These investigations have shown that, to achieve a gaze rotation towards a visual target from a reference position, only a constant proportion κ of the rotation is made by the head, while the remaining part is made by the eyes. This is illustrated in Fig 12. Thus, denoting by m_k^{ref} the reference direction of person k, by $m_{k,i}^{gaze}$ the gaze direction associated with the VFOA target i, the head pose mean $m_{k,i}$ we are looking for is defined by:

$$m_{k,i} - m_k^{ref} = \kappa(m_{k,i}^{gaze} - m_k^{ref}). \tag{14}$$

Following [28, 4], we used $\kappa = (0.5, 0.4)$. The gaze directions can be computed given the approximate position of the people and objects in the room. To set the reference direction m_k^{ref}, we considered two alternatives.

Manuel reference setting: the first one, used in [4] was to set it manually, by assuming that the reference direction roughly lies at the middle between the VFOA target extremes. For seat 1 and 2, this corresponds to looking straight in front of them (e.g. for seat 1, looking towards seat 2, see Fig. 1). For seat 3, and 4, this corresponds to looking at the nearest person to the slide screen on the opposite side (e.g. for seat 4, looking at seat 1).

Automatic reference setting: The reference direction corresponds to the average head pose of the recording. This approach still follows the idea that the reference split a person's gazing space, but allows to adapt the reference to individual people and meeting. From another point of view, it can be seen as the head direction which minimizes the energy to rotate the head during the meeting.

7 Bayesian model inference

Our inference problem is to estimate the conversational events $e_{1:T}$, the joint VFOA $f_{1:T}$, and the model parameters λ from the set of observations $(a_{1:T}, s_{1:T}, o_{1:T})$. In our probablistic framework, this is done by maximizing $p(f_{1:T}, e_{1:T}|a_{1:T}, s_{1:T}, o_{1:T})$ the posterior distribution of the hidden variables given the observation variables (see Eq. 4). Given the large dimensionality of our state space $(e_t, f_t))$ and the presence of hyperparameters λ, the use of direct maximization methods such as the Viterbi algorithm [37] leads to an untractable solution. Thus, we exploited the hierarchical structure of our

1. **Initialization: first estimate of the conversational events**. $n = 0$:

$$\hat{e}_{1:T} = \arg\max_{e_{1:T}} \prod_{t=1}^{T} p(\hat{s}_t|e_t)p(e_t|e_{t-1})p(e_t|a_t) \tag{15}$$

2. **Main loop:** repeat a) and b) until $n = N$

   (a) estimate VFOA state sequence and model parameters:

$$(\hat{f}_{1:T}, \hat{\lambda}) = \arg\max_{(f_{1:T},\lambda)} p(\lambda) \prod_{t=1}^{T} p(o_t|f_t, \mu, \Sigma)p(f_t|f_{t-1}, B)p(f_t|\hat{e}_t, a_t) \tag{16}$$

   (b) estimate conversation events sequence:

$$\hat{e}_{1:T} = \arg\max_{e_{1:T}} \prod_{t=1}^{T} p(\tilde{s}_t|e_t)p(e_t|e_{t-1})p(e_t|a_t)p(\hat{f}_t|e_t, a_t) \tag{17}$$

3. **Final estimate**: $(\hat{f}_{1:T}, \hat{e}_{1:T}, \hat{\lambda})$

Figure 13: Approximate inference algorithm.

model to define an approximate inference method which consists in the iteration of two main steps: estimation of the conversational events given the VFOA states, and reversely, estimation of the VFOA states and models parameters given the conversational events. This procedure guarantees at each time step to increase the joint posterior probability distribution. Similar inference procedures have already been proposed in [40, 36] for switching linear dynamic systems. Fig. 13 summarizes the different steps of our inference algorithm. Details about the algorithm are given below.

Conversational event estimation: It is conducted at two places. First, at initialization: conversational events are inferred solely from the speaking variables and the slide context (see Eq. 15). This is easily done using a standard Viterbi algorithm. Second, in the iterative loop, the re-estimation is done by maximizing $p(e_{1:T}|\hat{f}_{1:T}, s_{1:T}, a_{1:T})$, see Eq. 17, which can be done with a Viterbi algorithm. VFOA states play the role of observations and the corresponding term $p(\hat{f}_t|e_t, a_t)$ allows to take into account that for a given joint VFOA state, some conversational events are more probable than others. For example, when all the participant are looking at a given participant it is more likely that this person is speaking.

VFOA states and model parameters estimation. This is done in the step 2a, by maximizing $p(f_{1:T}, \lambda|\hat{e}_{1:T}, s_{1:T}, o_{1:T}, a_{1:T})$. The inference is conducted by exploiting the Maximum A Posteriori (MAP) framework of [13], which consists in first estimating the optimal model parameters by maximizing the parameter likelihood given the observations. Then, given these parameters, Viterbi decoding can be used to find the optimal VFOA state sequence [37].

Given our choice of prior, the MAP procedure relies on an EM algorithm similar to the one used in HMM model parameter learning, and iterates an E step (computation of VFOA state expectations) and a M step (re-estimation of optimal model parameters) until convergence. These 2 steps are summarized below. First, notice that when the prior on the joint state is not defined, the posterior we need to maximize can be written as

$$\prod_{k=1}^{4} p(f^k_{1:T}, \lambda_k|\hat{e}_{1:T}, s_{1:T}, o_{1:T}, a_{1:T}) \tag{18}$$

which means that the MAP procedure can be applied independently to each participant k. Thus, in the E step, we compute the expected value $\gamma^k_{i,t}$ for participant k to look at target j at time t, as well as his expectation $\xi^k_{i,j,t}$ of being in state i and j at time t and $t-1$, given the observations and the current parameter values:

$$\gamma^k_{i,t} = p(f^k_t = i | o^k_{1:T}, \hat{e}_{1:T}, a_{1:T}, \hat{\lambda}_k) \tag{19}$$

$$\xi^k_{i,j,t} = p(f^k_{t-1} = i, f^k_t = j | o^k_{1:T}, \hat{e}_{1:T}, a_{1:T}, \hat{\lambda}_k) \tag{20}$$

which can be computed using the classical Baum-Welsch forward-backward algorithm [37]. Given these expectations, the parameter re-estimation formula in the M step are.

$$\begin{aligned}
\mu_{k,i} &= \frac{\tau m_{k,i} + \sum_{t=1}^T \gamma^k_{i,t} o^k_t}{\tau + \sum_{t=1}^T \gamma^k_{i,t}} \\
\Sigma_{k,i} &= \frac{\tau(\mu_{k,i} - m_{k,i})(\mu_{k,i} - m_{k,i})' + V_{k,i} + \sum_{t=1}^T \gamma_{k,i}(o^k_t - \mu_{k,i})(o^k_t - \mu_{k,i})'}{d - p + \sum_{t=1}^T \gamma_{k,i}} \\
b_{k,i,j} &= \frac{\nu q_{k,i,j} - 1 + \sum_{t=1}^T \xi^k_{i,j,t}}{\nu - J + \sum_{l=1}^J \sum_{t=1}^T \xi^k_{i,l,t}}
\end{aligned} \tag{21}$$

Qualitatively, in the re-estimation, prior values are combined with data measurements to obtain the new values. For instance, $\mu_{k,j}$ is a linear combination of the prior values $m_{k,j}$ and of the pose observations o^k_t which most likely corresponds to the target j according to $\gamma^k_{j,t}$. The re-estimation formula are the same than in our previous work [4]. Indeed, the only but crucial difference with [4] lies in the computation of the expectations, for instance $\gamma^k_{j,t}$. While in [4], this term was only depending on the pose values (i.e. $\gamma^k_{i,t} = p(f^k_t = i | o^k_{1:T}, \hat{\lambda}))$, here the expectation takes into account the contextual information (slide, speaking status), see Eq. 20. This will increase the reliability that a given measurement is associated with the right target in the parameter adaptation process, leading to more accurate parameter estimates.
Finally, when a global prior $\Phi(f_t)$ is used, the MAP procedure can still be applied. The only difference is that we must now compute expectation over the joint VFOA space, e.g. $\gamma_{i,t} = p(f_t = i | o_{1:T}, e_{1:T}, a_{1:T}, \hat{\lambda})$. The needed individual expectation can be computed by marginalization, e.g. $\gamma^k_{j,t} = \sum_{i=(f^1_t,..,f^4_t)/f^k_t = j} \gamma_{i,t}$. The computation of the expectation requires the use of the forward-backward algorithm [37] which has an $O(TN^2)$ time complexity where T is the sequence length and N the size of the hidden space set. When N is large, as in the case where the joint focus of the four person are jointly decoded, computing these expectation can be computationally very expensive.

8 Evaluation Setup and Experiments

In this section, first we presents the experimental setup we followed to evaluate the model we propose for people's VFOA estimation based on a full conversational structure modeling including contextual information. Second we give the performances achieved by our models.

8.1 Experimental setup

The experimental setup we followed to evaluate the models we propose is the following. The evaluation dataset is composed of four recordings described in Section 3.1. In turn one of the recordings is used as evaluation data, the other recordings are used to train model parameters such as the tables modeling the VFOA dependencies to the slide activitites and the conversational events.

The performances of our algorithms are measured in term of frame based recognition rate (FRR) which is the percentage of frames that are correctly classified. The FRR is computed for each one of the meetings and the overall performance is taken as the average of the FRR over the four meetings. This ways of computing the overall FRR is motivated by our goal not to make long meetings weight more than short ones.

	without adaptation					with adaptation				
recordings	seat 1	seat 2	seat 2	seat 4	average	seat 1	seat 2	seat 2	seat 4	average
1	52.2	48.9	32.9	45.1	44.8	48.8	53.1	30.5	29.3	40.4
2	58.5	39.4	34.3	35.9	42	55.2	34.8	22.2	36	37
3	40.9	34.4	16.3	32	31	37.1	35.8	18.3	33.6	31.2
4	23.5	57.7	48.3	42.4	42.2	24	58.4	47.9	46.6	44.2
average	43.8	44.5	32.9	38.8	**40**	41.3	45.5	29.7	36.4	**38.2**

Table 1: VFOA recognition results from head pose without adaptation and without contextual cue.

8.2 Results

8.2.1 Global results

Without contextual priors Tab. 1 gives the results obtain with the model that performs VFOA estimation from head pose without the use of contextual priors. When no context is used the average FRR is about 40% when no unsupervised adaptation is used. When unsupervised adaptation is used the average performance is about 38%. Unsupervised model adaptation leads to a decrease of performances.

With contextual priors Tab 1 gives the results of the method that recognize VFOA from head pose using contextual prior. The left half of Tab 2 gives results when slide activity only is used as context, the right half of Tab 2 gives results when only conversational events priors is used. The use of slide activity priors leads to an FRR of 45% which represents an improvements of 5% with respect to (w.r.t) when no prior is used (see Tab 1). The use of conversational event context leads to a FRR of 50.7%, which represents a performance improvements of about 10.7%. wrt to when no context is used. The use of conversational event context leads also to 3% of improvements wrt to the use a slide activity context. This may be understood because the conversational event allows to set an informative priors on four VFOA targets (the four people), while the slide activity priors allows to set a direct prior on a single target (the slide screen).
Tab 3 gives the VFOA recognition results of the method using the full contextual information (projection screen activities and conversational events). For this method the FRR is about 55.1% which shows that using jointly the slide and conversational priors gives better results than using only one of them. Fig 14 shows the benefit of the use of the contextual priors on the VFOA recognition performances for the VFOA targets people [5], table and slide screen. The use of a slide activity context leads to improvements in the recognition of the slide targets, the use of conversational events context results in improvements in the recognition of the people. The joint use of both the contextual cues results in improvements for the slide and people's recognition.
Fig. 15 gives the average confusion matrixes for the method that recognize VFOA only from head pose (first rows), and the method that uses the full contextual priors (second row) for the four seating locations. We can notice on the confusion matrixes that there are more confusions for seat 3 and 4 than for seat 1 and 2 for both the methods. We can notice also that the confusion matrixes of the method using the full context have brighter diagonal, meaning that there are less confusions for this method.

8.2.2 Influence of the model parameters

In this section we analyze the effects of the various parameters of the models namely, the size of the window defining the conversational events, the explicit modeling of the VFOA group dynamics, the

[5]In our setup there are four persons, here they are grouped into a single target to study the effects of the contextual cues on the VFOA recognition performances.

	adaptation and slide prior					adaptation and conversational event prior				
recordings	seat 1	seat 2	seat 2	seat 4	average	seat 1	seat 2	seat 2	seat 4	average
1	58.7	58.4	42.5	56.6	54	59.7	67.7	53.5	41.7	55.7
2	66.8	44	47.1	53.2	52.8	65.2	19.8	38.1	38.8	40.5
3	35.1	36.2	18.3	33.4	30.8	74.7	47.9	26.1	59.1	52
4	42.3	56.2	45.2	37.7	45.3	50.1	59.6	56.6	52.7	54.8
average	50.7	48.7	38.3	45.2	**45.7**	62.5	48.8	43.6	48.1	**50.7**

Table 2: VFOA recognition results from head pose with adaptation and slide activity or conversational event contextual prior.

	without adaptation					with adaptation				
recordings	seat 1	seat 2	seat 2	seat 4	average	seat 1	seat 2	seat 2	seat 4	average
1	59	62.8	48.2	58.7	57.2	59.3	67.9	53.9	58.4	59.9
2	58.2	40.7	41.4	50	47.6	67.2	50.5	47.1	54.8	54.9
3	50.7	47.7	23.5	55.8	44.4	70.4	47.5	21.6	58.6	49.5
4	44.8	58.8	64.1	51.2	54.5	53	59.3	62.3	49.4	56
average	53	52.5	44.3	53.9	**50.9**	62.5	56.3	46.3	55.3	**55.1**

Table 3: VFOA recognition results from head pose with full contextual prior (slide and conversational events) wit hand without adaptation.

causality of the conversational events, the model adaptation, the reference of the cognitive model.

Window size The window size of the conversational events is the duration that is consider to compute the speaking proportion that are used as observation for the conversational events estimation. To study the effects of this parameter on the VFOA recognition performances we considered fives windows sizes: 1,3,5,9,15 seconds. Fig 16 gives the average performances for seating locations and recordings. We can notice that augmenting the size of the window defining the conversational events leads to a decrease of performance. This can be explained by the fact that taking wider conversational windows has as effects to loose information about when speaking events occurred. Thus, using the conversational events priors do allows to set VFOA priors on speakers.

Explicit group dynamics The explicit group dynamics models people tendencies to share VFOA target. Although, the use of contextual cues implicitly models partially models people tendencies to focus at the same targets, it only models when people are jointly focusing on the speakers or on the projection screen when there is a recent slide change. The use of an explicit group dynamics allows to model when people are jointly focusing at non-speakers or at the slide while there is no recent slide activity. Fig 17 shows the performances when using an implicit group dynamics models (only contextual priors), and when using the explicit group dynamics model described in Section XX. In average, the method using an explicit group dynamics achieves a FRR of 55.6% while the method using an implicit model achieves a FRR of 55.1%. Thus including explicit group dynamics is useful. However, when an explicit group dynamics is used, the state space becomes very large with a dimension of 6^4 while the method using an implicit models has a state space dimensionality of 4×6. Knowing for search methods such as Viterbi search, the computational cost is an exponential function of the state space dimension, searching the state space for a model with an explicit dynamics modeling is computationally very expensive.

Non-causal or causal conversational events The causality of the conversational events observation refer to whether or not when computing the observation for a frame t, information about time later then t are used. In our case where offline processing is done using frames information from the

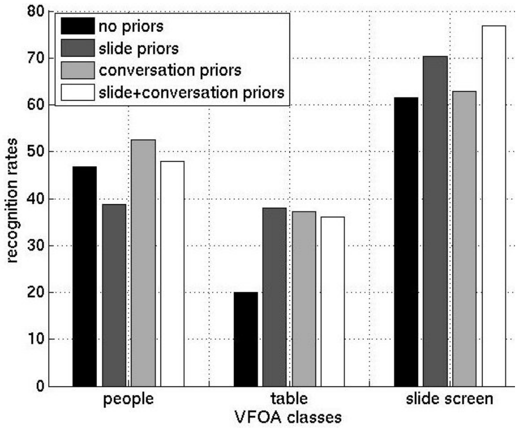


Figure 14: Effects of VFOA contextual priors on recognition for people, table and slide screen. The use of slide activity prior improves the recognition of the slide screen target and the use of conversational event priors improves the recognition of people as VFOA targets.

future in not an issue. Thus it is possible for us to use non-causal conversational events. In a situation of online processing information from the future cannot be used, the conversational events have to be causal. The average FRR for the method using a causal event is about 54.6%, while the method using a non-causal events achieves 55.1%. Thus, although the non-causal events achieves better results, the loss when using a causal event, 0.5%, is not very high.

Model adaptation The model adaptation refers to the joint inference of the optimal parameters of the model and the optimal VFOA and conversational events state sequence. Tab 1 when no contextual priors was used, model adaptation do not results in performance improvements. Because the adaptation is unsupervised there exit a risk that the model parameter drift because the VFOA center configuration were very complex due to our experimental setup with high potential ambiguities in the head poses defining the VFOA targets. Tab 3 gives the results using full prior without and with adaptation. This table shows that the average performance when full prior is used without adaptation is about 50.9% while when adaptation is used the performances are about 55.1%. Adaptation in the presence of contextual priors is beneficial contrarily to adaptation without priors. Fig 19 shows illustrations for VFOA center adaptations with and without contextual priors. As can be seen in this figure, when contextual priors is used the adapted VFOA centers are closer to the VFOA center that would be obtained if data of the same person were used to learn the VFOA centers.

Cognitive model reference The cognitive model reference refers is used to build the relationship between a person's gaze and his head pose. We followed two methodologies to set it. The first methodology considers as the gaze reference the average of the head pose of the person during an entire recordings. This setting followed the idea of a gaze reference being the center of mass of the person's head poses. The second methodology consisted in setting the reference manually using the geometry of the room. This setting requires knowledge of the room geometry. Fig. 20 gives the results for two settings of the reference. The average FRR when the gaze reference is manually set is about 53.6% while it is about 55.1% when it is taken as the average of the person's head pose. In Fig 20 we can see mainly that the difference of performance between the two ways of setting the reference is due to seat 4 of the recording 3. When the reference is set manually the performance on this data point is about 37.2% while the performances on this data point is about 58.6 % when the reference is taken a the mean of the head poses. As shown in Fig. 21, the initial VFOA centers are better predicted by the gaze model using the head pose average as reference, leading to better predictions after adaptation.

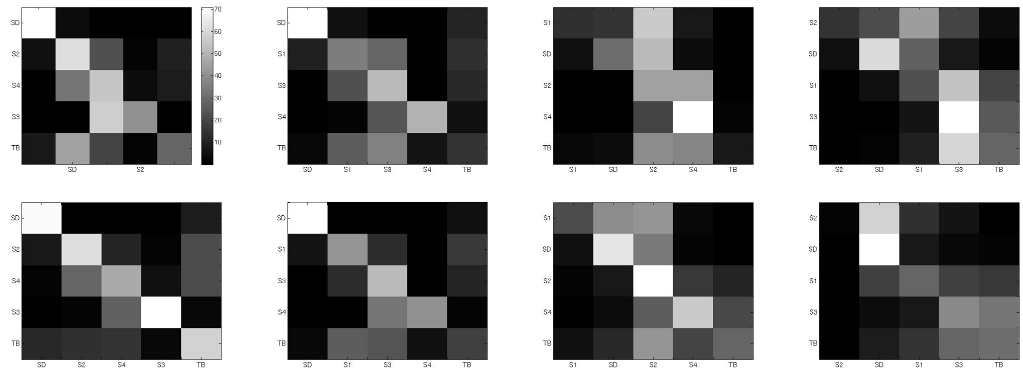


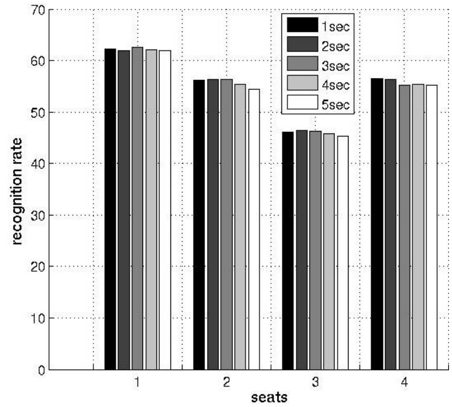
Figure 15: VFOA recognition confusion matrix without context (first row), with context (second row). Column 1,2,3,4 correspond to seat 1,2,3,4. More confusions are exhibited for seat 3 and 4. The use of context allows for less confusions (brighter diagonal for the confusion matrixes of the second row).

8.2.3 Illustrations

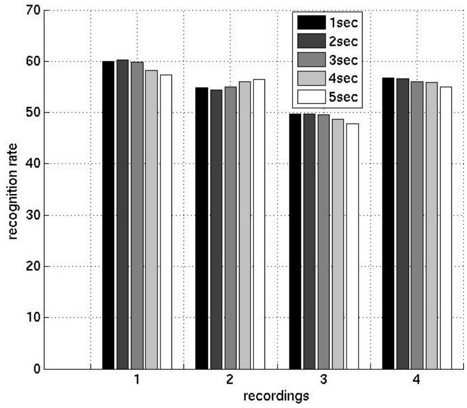
This Section gives qualitative illustrations about the behavior of the VFOA recognition method we proposed. Fig 22 gives a one minute sequence illustration for estimated speaking status, estimated conversational events, manually annotated VFOA, and estimated VFOA. This sequence illustrate the relationship between a recent slide activity and people more focusing at the slide. It also illustrate well the focus of the listeners being the speakers. The conversational events that our model estimate is more a smoothed version of the speaking status. Monologues are mainly periods in which mostly only one person is speaking while dialogs and discussions are periods where two or more people are sharing the floor. Thus, periods in which a speaker change occurs may be recognized as dialogs.

9 Conclusion

This paper presented a dynamic Bayesian network for the joint multi-party VFOA and conversational events recognition in meetings from head pose information and multi-modal contextual information. Contextual information was in troduced in our model by the use of an input variable conveying information about the activities of a projection screen (slide changes). Dependencies between people's focus of attention and the ongoing conversational event was modeled, as well as dependencies of the people's attention and conversational event to the projection screen activities. From our study the following conclusions can be drawn. First the use of contextual priors leads to significant improvements with respect to when no context is used. Secondly, in situation where unsupervised adaptation is used, the use of informative contextual priors is beneficial as it reduces the risk of the adaptation process to drift. Thirdly, although an explicit modeling of VFOA group dynamics leads to recognition performances improvements, its computational cost is high. In a practical situation, an implicit modeling of the group dynamics through the contextual priors might be more efficient when considering a good balance between recognition performance and computational efficiency. In this paper we only consider the projection screen activities and the conversational events effects on people's attention leading to an increase of performances recognition for the people and slide screen and sometimes a decrease of performance for the recognition of the table. Table activities could also be used as contextual priors. However the larger the set of context the more data would be required to learn the dependencies among the variable.



(a) average FRR for seats



(b) Average FRR for recordings

Figure 16: Average frame based recognitions for seating locations and recordings for varying conversational event based windows (1,3,5,9,15 secs). Wide conversational events windows leads to lower VFOA recognition performances.

Aknowledgements

This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the European union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811). This research was also funded by the U.S. Government VACE program. The authors thank Chuohao Yeo from UC Berkeley who built the compressed domain features based slide change detector that was used to incorporate contextual prior in our model. Finally , the authors also thank Dr. Daniel Gatica-Perez, Dr. Hayley Hung, and Dinesh Jayagopi from IDIAP Research Institute for their helpfull discussions.

References

[1] M. Argyle. *Bodily communication.* International University Press, 1975.

[2] M. Argyle and J.Graham. The central europe experiment - looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour*, 1:6–16, 1977.

[3] S. O. Ba and J.-M. Odobez. A Rao-Blackwellized mixed state particle filter for head pose tracking. In *Proc. ACM-ICMI-MMMP*, pages 9–16, 2005.

[4] S.O. Ba and J.-M Odobez. Recognizing human visual focus of attention from head pose in meetings. *IEEE Transaction on Systems, Man, and Cybernetics*, 2008.

[5] J. N. Bailenson, A.C. Beal, J. Loomis, J. Blascovitch, and M. Turk. Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. *Human Comm. Research*, 31(4):511–537, October 2005.

[6] S. Basu. *Conversational Scene Analysis.* PhD thesis, Massachusset Institute of Thechnology, 2002.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal. The AMI meeting corpus: A pre-announcement. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

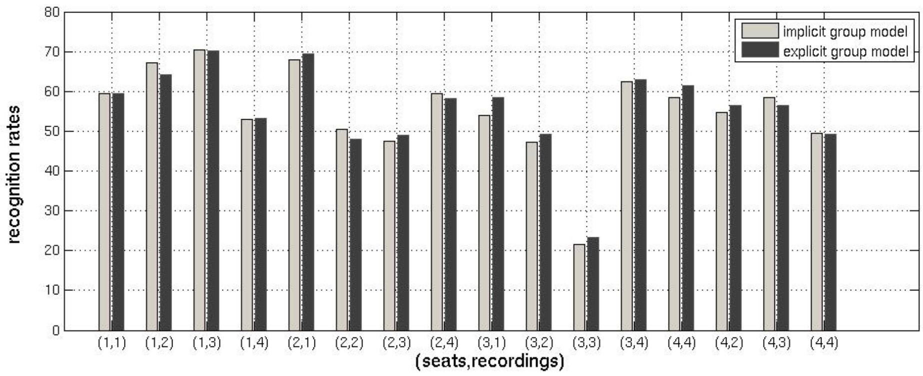


Figure 17: Implicit and explicit group dynamic modeling. Implicit model include only contextual priors, explicit models include priors on people sharing focus targets. In average, using an explicit group dynamics models allows to obtain better results. Both explicit and implicit group dynamics modeling achieve similar performances.

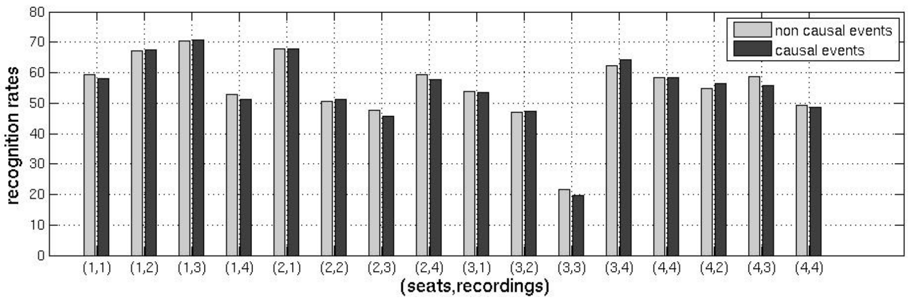


Figure 18: Non causal and causal conversational events. In average, the use of a non-causal conversational event gives better results. However the results for each evaluation data point (seat,recording) shows that the two setting achieves similar results.

[8] L. Chen, M. Harper, A. Franklin, T.R. Rose, and I. Kimbara. A Multimodal Analysis of Floor Control in Meetings. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[9] L. Chen, R.T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Wel, T.X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R.Tuttle, and T. Huang. VACE multimodal meeting corpus. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[10] S. Duncan Jr. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[11] Edward G. Freedman and David L. Sparks. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77:2328–2348, 1997.

[12] D. Gatica-Perez. Analyzing group interactions in conversations. In *Proc. of the Int. Conf. Multisensor Fusion and Integration for Intelligent Systems*, pages 41–46, 2006.

[13] J.L. Gauvain and C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–213, 1992.

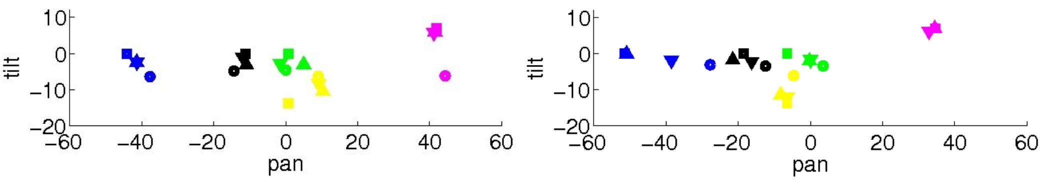


Figure 19: Effect on adaptation on Gaussian means defining the VFOA targets for two persons sitting at seat 1: (square) predicted using the model, upper triangle adaptation without contextual priors, lower triangle adaptation with contextual priors, circle predicted using annotated data of the subject. When contextual prior is used, the adapted VFOA centers are closer to the VFOA centers that are predicted using head pose data of the same person.

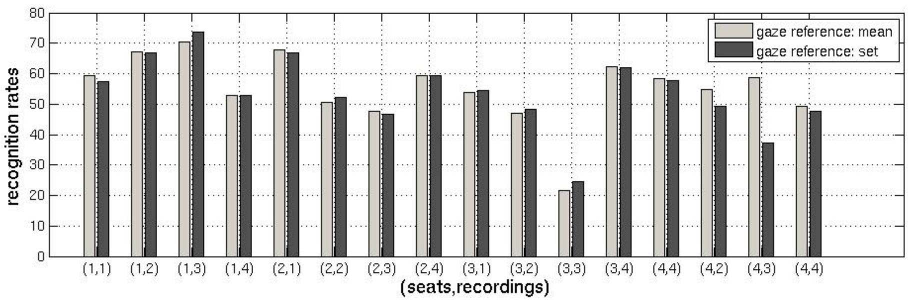


Figure 20: Gaze reference setting effects on the VFOA performances. The performance for the two methodologies to set the gaze reference are close except for seat 4 or recording 3 where the the method using the manually set gaze reference achieves lower results (37.2% against 58.6%).

[14] C. Goodwin and J. Heritage. Conversation analysis. *Annual Review of Anthropology*, pages 981–987, 1990.

[15] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures*, pages 183–191, 2004.

[16] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *TRENDS in Cognitive Sciences*, 9(4):188–194, 2005.

[17] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. O. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proc. of ACM Multimedia*, 2007.

[18] N. Jovanovic and H.J.A. Op den Akker. Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, 2004.

[19] N. Jovanovic, R. op den Akker, and Anton Nijholt. Addressee identification in face-to-face meetings. In *Proc. the 11th Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, 2006.

[20] M. Kouadio and U. Pooch. Technology on social issues of videoconferencing on the internet: a survey. *Journal of Network and Computer Applications*, 25:37–56, 2002.

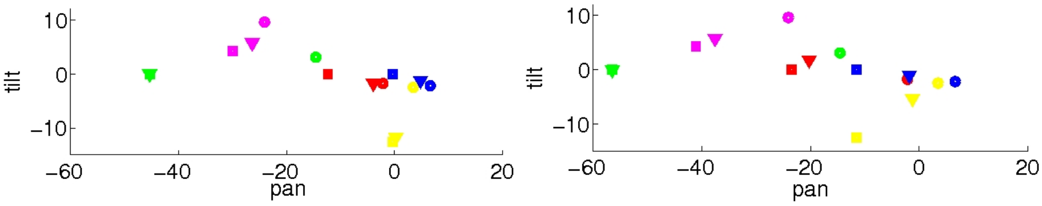


Figure 21: Effect on mean prediction when using a reference learned (left) and a reference manually.square=prediction, triangle=adapted, circle= from subject's data

[21] O. Kulyk, J. Wang, and J. Terken. Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2006.

[22] S.R.H. Langton, R.J. Watt, and V. Bruce. Do the eyes have it ? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–58, 2000.

[23] I. McCowan, D. Gatica-Perez, G.Lathoud, M. Barnard, and D. Zhang. Automatic analysis of group actions in meetings. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27:305–317, 2005.

[24] J.E. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[25] A. F. Monk and C. Gale. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. pages 257–278, 2002.

[26] C.H. Morimoto and M.R.M. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98:4–24, 2005.

[27] D. Novick, B. Hansen, and K. Ward. Coordinating turn taking with gaze. In *International Conference on Spoken Language Processing*, 1996.

[28] J-M. Odobez and S.O. Ba. A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of Focus of Attention from Head pose. In *Proc. of ICME*, 2007.

[29] T. Ohno. Weak gaze awareness in video-mediated communication. In *Proceedings of Conference on Human Factors in Computing Systems*, pages 1709–1712, 2005.

[30] N. Oliver and E. Horvitz. Layered representations for human activity recognition. In *Proc. the International Conference on Multimodal Interfaces*, 2002.

[31] N.M. Oliver, B. Rosario, and A.P. Pentland. A bayesian computer vision system for modeling human interactions. In *Proc. CVPR Workshop on Interpretation of Visual Motion*, 1999.

[32] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. of ICMI*, pages 191–198, 2005.

[33] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Conversation scene analysis with dynamic bayesian network based on visual head tracking. In *Proc. of ICME*, 2006.

[34] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proceedings of Conference on Human Factors in Computing Systems*, pages 1175–1180, 2006.

[35] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: a survey. In *Proc. the International Conference on Multimodal Interfaces*, 2006.

[36] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems (NIPS)*, pages 981–987, 2000.

[37] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in Speech Recognition*, 53A(3):267–296, 1990.

[38] H.-S. Rhee, H. Pirkul, V. Jacob, and R. Barhki. Effects of computer-mediated communication on group negotiation: Au empirical study. In *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, pages 981–987, 1995.

[39] R. Rienks and D. Heylen. Automatic dominance detection in meeting using easily detectable features. In *Proc. of the MLMI workshop*, 2005.

[40] R.H Shumway and D.S. Stoffer. Dynamic linear model with switching. *Journal of the American Statistical Society.*

[41] M. Siracusa, L.P. Morency, K. Wilson, J. Fisher, and T. Darrell. A multi-modal approach for determining speaker location and focus. In *Proc. of International Conference on Multimodal interfaces (ICMI)*, 2003.

[42] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4):928–938, 2002.

[43] C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.

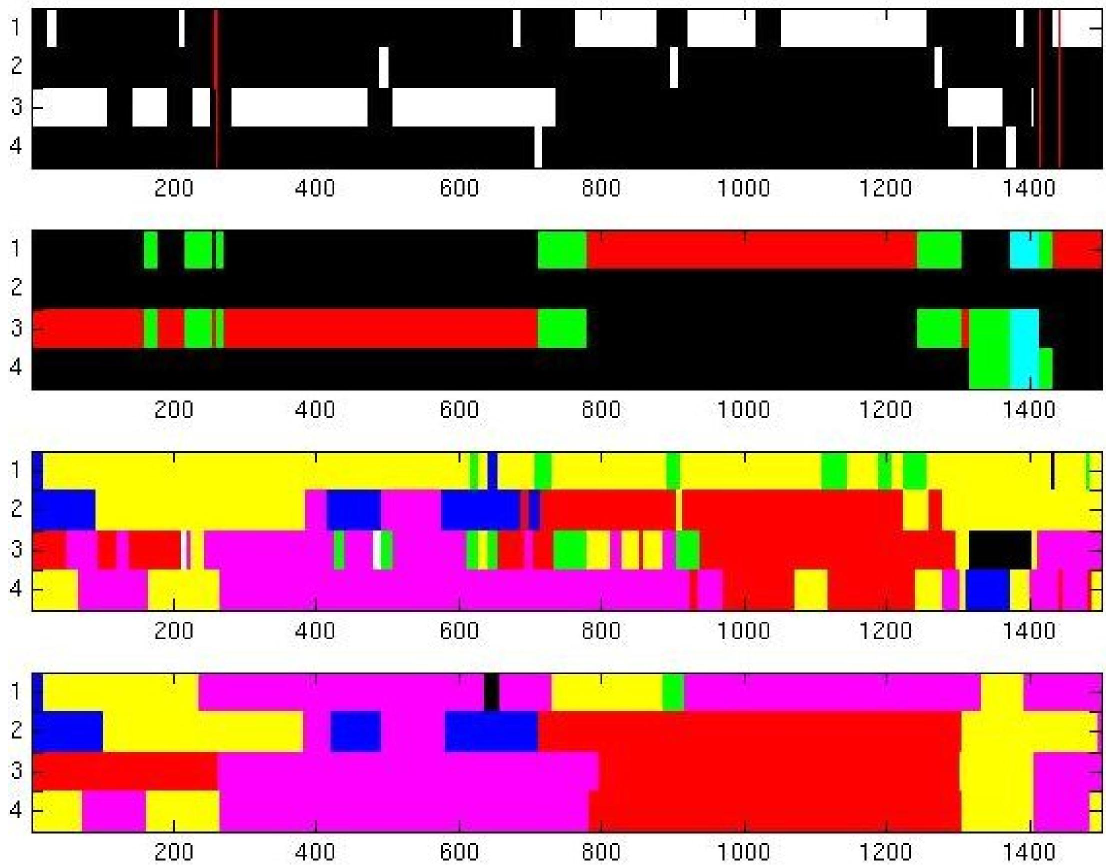


Figure 22: One minute (minute 11, recording 1) illustration for speaking status observation (top image), estimated conversational events (second image), manually annotated VFOA, estimated VFOA. For each image the abscissa is the timeline while the ordinate represent observation for the four seats. For the speaking status (top image) the color in a given time frame means that the person a a given seat is silent, white means the person is speaking. The red line highlight the time when a slide change occurs. For the conversational event (second image), the color red means the person is making a monologue, green tags two person is involved in a dialog, cyan tags three persons involved a a 3 person discussion. For the VFOA illustration (ground truth and estimates), red tags a person focusing at seat 1, green at seat 2, blue at seat 3, black at seat 4, yellow at the table, magenta at the slide screen and white stands for unfocused. The estimated conversational event is a smoothed version of the speaking status. After a slide change we can notice more presence of the slide as focus target and a speaker also is more a target.