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

Sileye O. Ba*, and Jean-Marc Odobez†, *Member, IEEE*

LabSTICC, Ecole Nationale des Télécommunications de Bretagne*, 29238, Technopole Brest-Iroise, France

Telephone	+33 2 29 00 15 70
Fax	+33 2 29 00 10 12
Email	sileye.ba@telecom-bretagne.eu

Idiap Research Institute†, Rue Marconi 19, CH-1920 Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL)†, CH-1015 Lausanne, Switzerland

Telephone	+41 27 721 77 26
Fax	+41 27 721 77 12
Email	odobez@idiap.ch
URL	www.idiap.ch

***Abstract*— This paper introduces a novel contextual model for the recognition of people's visual focus of attention (VFOA) estimation in meetings from audio-visual perceptual cues. More specifically, instead of independently recognizing the VFOA of each meeting participant from his own head pose, we propose to jointly recognize the participants' visual attention in order to introduce context dependent interaction models that relates to group activity and the social dynamics of communication. Meeting contextual information is represented by the location of people, conversational events identifying floor holding patterns, and a presentation activity variable. By modeling the interactions between the different contexts and their combined and sometimes contradictory impact on the gazing behavior, our model allows to handle VFOA recognition in difficult task-based meetings involving artifacts, presentations, and moving people. We validated our model through rigorous evaluation on a publicly available and challenging dataset of 12 real meetings (five hours of data). The results demonstrated that the integration of the presentation and conversation dynamical context using our model can lead to significant performance improvements.**

***Index Terms*— Visual focus of attention, conversational events, multi-modal, contextual cues, dynamic Bayesian network, head pose, meeting analysis.**

I. Introduction

In human societies, meetings whether formal or informal are important daily life activities. It is the place where several people come together to exchange and disseminate information, discuss some predefined topics, come to an agreement and take important decisions. Due to this importance, there has been a recurrent development of technology oriented tools to support meeting effectiveness such as meeting browsers [1] or the use of automatic speech transcription to perform automatic content linking [2] (i.e. relevant documents are automatically retrieved according to the discussed topic) or create summaries [3]. At the same time, meetings are arenas where people interact and communicate, laugh, argue, feel and express emotions. Thus, following social scientists [4], computer scientists have explored more deeply the social aspects of meetings through automatic analysis of different non-verbal communication cues. For instance, decomposing meetings in different phases based on turn-taking patterns and meeting activities cues [5] or extracting people's status or role [6] can lead to the design of efficient tools for computer enhanced human-to-human interactions.

Analyzing meetings requires the ability to understand the behaviors that are exhibited during human interaction. Among these behaviors, gaze represents one of the fundamental non-verbal communication cues with functions such as establishing relationships, expressing intimacy, and exercising social control. The role of gaze as a turn holding, taking or yielding cue to regulate the course of interaction has been emphasized by social studies [7]. For instance, speakers use their gaze to indicate whom they address and secure the listeners' attention [8], while listeners turn their gaze towards speakers to show their attentiveness and to find suitable time windows to interact [4]. Thus, an appropriate perception and usage of such gazing codes is important for smooth communication. Kulyk *et al.* [9] showed that real-time feedback about meeting participants' gaze activities positively affects the participants' behavior, resulting in improved group cohesiveness and participant satisfaction. In video conferencing situations, where the lack of perception of non-verbal signals (especially the gaze) can lead to communication misunderstandings [10], studies have been conducted towards the enhancement of gaze perception [11] to favor the engagement of remote participants.

Researchers have investigated the recognition of gaze in meetings. Roughly speaking, the eye-gaze tracking systems that are used in human computer interaction [12] are not appropriate for analyzing the conversation of several people in a room. They can be intrusive or interfere with natural conversation because they restrict head mobility and orientation. As an alternative, head pose can be used as an approximation

Silèye Ba was with the Idiap Research Institute while this work was done.

for gaze, as supported by psycho-visual evidence [13] and empirically demonstrated by Stiefelhagen *et al.* [14] in a simple setting. However, in realistic scenarios, the level of head pose ambiguities (the same head pose can be used to look at different targets) makes the recognition of the visual focus of attention (VFOA) solely from head pose a difficult task [15].

Using context for VFOA recognition. When analysing group interactions, context plays an important role as the same non-verbal behavior can have a different interpretation depending on the context. In general, the context relates to the set of circumstances in which an event takes place and which are relevant to derive a precise understanding of this event. For VFOA recognition, the circumstances correspond to whatever can attract or affect the gaze of people, and can be classified into two main categories: a social interaction context (how people are conversing), and a task context (what are the people or the group doing). The knowledge of these contexts can improve head pose interpretation in two ways. First, by setting priors on the VFOA targets to indicate which targets are more likely in a given context. For instance, humans tend to look more at a speaking person than a non-speaking one. And second, by automatically discovering which head orientations correspond to looking at some given targets. From a computational perspective, this means that we can reliably adapt a prior head pose-target correspondance model by observing for some time a person's behavior and the context.

In the past, only the conversation context has been investigated for VFOA recognition, mainly by exploiting the gaze/speaking turn relationships discussed above. For instance, Stiefelhagen *et al.* [14] directly used people's speaking statuses as conversation context, while Otsuka *et al.* [16] introduced the notion of conversation regimes driving jointly the sequence of utterances and VFOA patterns of people, and applied their model to short meetings involving only conversation.
However, there are other important events and situations which can significantly influence gaze and should be taken into account, and have never been considered so far. For instance, participants do not usually remain in their seats during an entire meeting. They can stand up to make presentations or use the white board. In such cases, introducing the location of people as a context is mandatory since it has a direct impact on the gazing direction of people. Furthermore, most meetings involve documents or laptops as well as projection screen for presentations. This calls for the exploitation of task-oriented contexts linked to the use of these artifacts for automatic inference of VFOA in groups, since the usage of artifacts significantly affects the person's gazing behavior and is more representative to a typical meeting scenario. This is the so called *situational attractor* hypothesis of Argyle [17]: objects involved in a task that people are solving attract their visual attention, thereby overruling the trends for eye gaze behavior observed in non-mediated human-human conversations. For instance, Turnhout *et al.* [18] reported that two people interacting with an information kiosk artificial agent looked much less at their partner when addressing him than in normal dyadic interactions. Similar impact of artifacts on gaze behavior was also reported in the analysis of task-based meeting data [19]. For instance, during presentation people usually look at slides not at speakers. Thus, generalizing state-of-the-art VFOA recognition systems designed for conversation-only meetings [14], [16], [20] to handle work meetings involving mobile targets and artifacts is not straightforward. The aim of this paper is to investigate new models towards such generalization.

Contributions. This paper addresses the joint recognition of people's VFOA in task-based meetings. This is achieved using meeting contextual models in a dynamic Bayesian network (DBN) framework, where the social context is defined by conversational events identifying the set of people holding the floor and the location of people in the room, and the task context is defined by a presentation activity variable. Our main contributions are:

- a model for which the influence of normal face-to-face conversation on gaze patterns is modulated by context to account for the change of persons' location in the room and the continuous change of gaze and conversational behavior during presentations.
- an explicit prior model on the joint VFOA which captures shared attention, and a fully automatic cognitive model to associate a person's head pose to a given VFOA target;
- the use of the model for the recognition of VFOA in task-based meetings, and its evaluation on a significant amount of data from a publicly available database; and
- a detailed analysis of the effects of the different contextual information and modeling steps on recognition performance, including the benefit of using context for model adaptation.

Experiments on five hours of meeting data from the AMI Corpus [21] indicate that the use of context increases the VFOA recognition rates by more than 15% with respect to using only the head pose for estimation, and by up to 5% compared to the sole use of the conversational context (i.e. with no presentation context and shared attention prior).

The paper is organized as follows. Section II discusses related work. Section III introduces the set-up, task, and data used for evaluation, while Section IV describes our DBN observations. Section V discusses our approach for context modeling, and gives an overview of our DBN model. Section VI details the components of this DBN model and our inference scheme. Section VII presents our results, and Section VIII concludes the paper.

II. Related Work

Multi-person VFOA and conversational event recognition relates to the automatic recognition of human interactions amongst small groups in face to face meetings. In the following, we review work along three different threads that researchers have taken to conduct such analysis. The first one relates to the temporal segmentation of meetings in different group activities. A second one relates to floor control and addressee recognition, which involves focus of attention as an important cue. A third important thread corresponds to approaches which directly investigate the recognition of the VFOA in meetings.

Turn taking patterns are one of the most important cues to analyze conversations in general, particularly in meetings. During interactions, conversations evolves through different communication phases characterizing the state and progress of a collaborative work. Based on this assumption, several researchers have investigated the automatic segmentation of meetings into different group activities such as monologue, discussion, presentations, etc, from audio-visual cues. For instance, McCowan *et al.* [5] explored the use of several DBNs to fuse the different data streams, where a stream corresponds to audio or visual features extracted from the group of meeting participants. Zhang *et al.* [22] improved the robustness of the model by using a layered approach, where the first layer modeled the activities of each participant and the second layer identifies the group activities. A major difference between these works and our approach is that group activities were modeled by global statistics over various feature streams and did not include explicit models of individual behavior interactions or context modeling. An exception is the work of Dai *et al* [23], who proposed to use a dynamic context for multi-party meeting analysis. Meetings events, including individual behaviors like speaking, were organized into a hierarchical structure. Higher level were defined based on lower level events and multi-modal features, while lower level events are detected using the higher level events as context. However, no results on individual behaviors were reported.

Floor control and addressee are two multi-modal interaction phenomenons involving gaze and speech that play a significant role in conversation analysis. When investigating audio-only features, a natural and common way to describe the conversation status is to define "floor holding" pattern variables, as done by Basu [24]. More generally, Chen *et al.* [19] investigated the combination of gaze, gesture, and speech for floor control modeling, i.e. the underlying mechanisms on how people compete for or cooperate to share the floor. By analyzing meetings of the VACE corpus, they showed that discourse markers occur frequently at the beginning of a turn, and mutual gaze occurs when a speaker is transmitting the floor to a listener. However, no automatic processing for floor estimation was reported.

Addressee detection (detecting to whom the speech is intended) has been studied in the context of artificial agents (information kiosk [18], robots [25]) interacting with multiple people, in order to differentiate between human-to-human and human-to-agent addressing. For instance, Katzenmeir *et al.* [25] used a Bayesian scheme to combine speech features and head pose to solve the task, but no attempt to model the dynamics of the interaction was done. Few researchers have investigated the identification of addressee in multi-party conversations. Jovanovic *et al.* [26] evaluated the combination of several manually extracted cues using a DBN classifier for addressee recognition on the task-based AMI corpus meetings. They reported that specific lexical features worked best to detect addressee, and that the speaker's gaze is correlated with his addressees, but not as strongly as in other works [20] due to the different seating arrangements and the presence of attentional 'distractors' (table, slide-screen) affecting the gaze behavior as an addressing cue [18]. No results with automatic feature extraction are reported.

Finally, more related to our research are the investigations about the recognition of the VFOA ("who looks at whom or what") from video (mainly head pose estimated from videos) [14], [15], [27] or audio-visual data [14], [20], [28]. Stiefelhagen *et al.* [14] proposed a method to recognize people's focus solely from their head pose using a Gaussian mixture model in which each mixture component corresponds to a given focus. A similar approach was used by Siracusa *et al.* [28] who investigated identifying speakers and people's VFOA in an information kiosk setting from a microphone array and a stereo rig. In our previous work [15], we proposed to use an HMM model along with an unsupervised parameter adaptation scheme, and a cognitive gaze model for parameter setting. Applied to task-based meetings, the results showed the benefit of the approach but were lower than those reported in other studies due to the data complexity. Very recently, Voit *et al.* [29] introduced a new dataset to investigate the recognition of the VFOA of one person seated in a lecture room and involving moving people. Initial results evaluated on a single person data were reported.

Contextual information has been exploited as well. In their work, Stiefelhagen *et al.* [14] proposed to linearly combine the VFOA recognition from head pose with a VFOA recognizer relying on the speaking statuses of all participants, but this led to only a 2% recognition rate increase, probably due to the simplicity of their setting (4 people equally spaced around a table). Otsuka *et al.* [16], [20] were the first to propose a joint modeling of people's VFOA. They introduced the notion of conversation regimes (e.g. a 'convergence' regime where people gaze at the speaker; or a 'dyadic' one where two people look at each other) modeling the interactions between the utterances and the VFOA patterns of all meeting participants. Regimes were considered as hidden variables along with people's VFOA, and were inferred from people's speaking statuses and head poses using a Markov chain Monte Carlo (MCMC) optimization scheme. However, no explicit prior on the joint focus was used, and like other previous works, the model could not handle moving people or exploit context other than speech cues. In addition, all models require the setting of head poses associated with gaze directions. Although these values, which can be numerous (we have 36 of these in our case), can have a high impact on performance, most works set these values manually [14], [16], [20] or semi-automatically [15]. Finally, both Otsuka [16] and Stiefelhagen [14] used a four-participants meeting setup with VFOA targets restricted to the other participants, which differs significantly from our complex task-based meetings involving artifacts, presentations and moving people. Thus, in addition to the modeling, our paper presents a thorough investigation of the VFOA contextual recognition in challenging, but real and natural meetings.

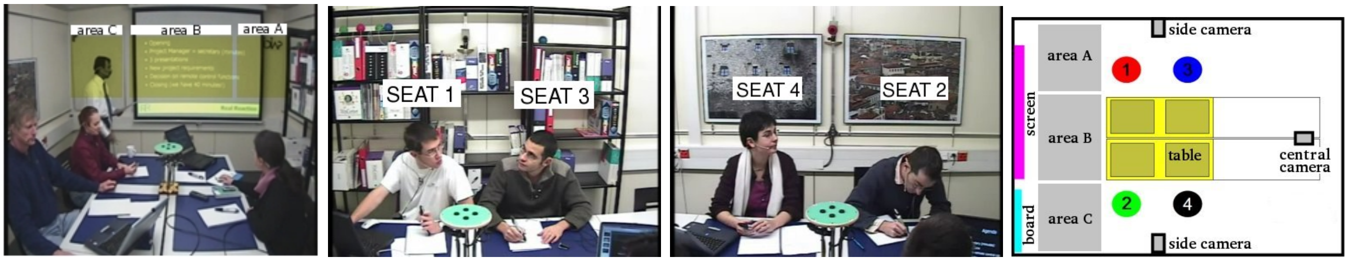

Fig. 1. Meeting recording setup. The first image shows the central view that is used for slide change detection. The second and the third image show the side cameras that are used for head pose tracking. The last image shows a top view of the meeting room to illustrate seats and focus target positions.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

III. Dataset and Task

A. Setup and Dataset description

Our data source is the Augmented Multi-party Interactions (AMI) corpus[1]. In this corpus, recorded meetings followed a general task-oriented scenario, in which four people with different roles (project manager, marketing expert, user interface and system designer) were involved in the creation and design of a television remote control. The phases of the design process were discussed in a series of four meetings, where the participants presented their contributions according to their expertise. During these meetings, people had natural behaviors, taking notes, using laptops, making presentations with a projection screen, and possibly manipulating a prototype of the remote control. The twelve meetings of this corpus for which VFOA annotation was available were used. Twenty different people were involved in the recordings, making the head pose tracking a challenging task. The meeting durations ranged from 15 minutes to 35 minutes, for a total of 5 hours. With respect to the dynamic aspect of the AMI Corpus meeting, 23% of the time there was a person standing to make a presentation.

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

B. VFOA recognition Task and Data analysis

Task: Our objective is to estimate the VFOA of seated people. Although we address the problem of looking at moving people, the reverse problem of estimating the VFOA of moving people has not been considered in this work. A main reason is that the head poses estimated from the central camera are too noisy to be exploited.

Although in principle the VFOA is defined by continuous 3D eye gaze directions, studies have shown that humans tend to look at specific targets (location/objects, humans) which are of immediate interest to them [30]. Thus, we have defined the VFOA of a person seated at seat k as an element belonging to a finite set of visual targets $\mathcal{F}_k$. This set is composed of the 3 other participants $\mathcal{P}_k$ as well a set of 4 other targets $\mathcal{O} = \{$Table, White Board, Slide Screen, Unfocused$\}$. The label 'Table' is used whenever people would look at the table and anything on it (e.g. their laptop), while the label 'Unfocused' is used when the person is not visually focusing on any of the other targets. It is important to note here that, while according to our VFOA definition the number of *semantic* targets a person can look at is 7, the number of *physical* locations he can look at is larger to account for the fact that standing people can occupy different places in the room. More precisely, when people stand-up to make a presentation, we assume that they are located in one of the 3 areas A, B or C, as shown in Fig. 1 (left).

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 view cameras. The VFOA annotations had high coding agreement among annotators (see [31] page 80). We computed the VFOA ground truth statistics, where we have grouped the VFOA labels corresponding to participants into a single label 'people'. Looking at people only represents 39% of the data, while looking at the table, slide screen, or white-board represents respectively 30% and 24%, and 2.7% of the data. These statistics show that the usual face to face conversation dynamics where people mainly look at the speakers did not hold. Artifacts such as the table and the projection screen play an important role that has to be taken into account to understand the conversation dynamics in our meeting scenario. This places our work in a different context as compared to previous studies which have investigated VFOA recognition in scenarios involving only short discussions and no contextual objects [14], [20]. Natural behaviors such as people looking downwards (at the Table) without changing their head pose while listening to a speaker, are less frequent in these shorter meetings than in our data, as we noticed when comparing with our previous study on 7 to 10 min long meetings [15]. In addition, in [14], [20], the only targets are 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 downward looking natural behavior described above, and to the fact that, in contrast to [14], [20], 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.

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

IV. Multi-modal features for VFOA modeling

The VFOA recognition and the identification of context relies on several observation cues: people speaking activities, head poses, and locations, and a presentation activity feature.

Audio Features $\tilde{s}_t^k$: The audio features are extracted from close-talk microphones attached to each of the meeting participant. At each time step t the speaking energy of participant k is thresholded to give the speaking status s_t^k which value is 1 if participant k is speaking and 0 otherwise. To obtain more stable features, we smoothed the instantaneous speaking status by averaging them over a temporal window of W frames. Our audio feature for person k is thus $\tilde{s}_t^k = \frac{1}{W}\sum_{l=-\frac{W-1}{2}}^{\frac{W-1}{2}} s_{t+l}^k$, that is the proportion of time person k is speaking during W frames.

The head poses o_t^k. To estimate people's head location and pose, we relied on an improved version of the method described in [32]. The approach is quite robust especially because head tracking and pose estimation are considered as two coupled problems in a Bayesian probabilistic framework solved through particle filtering techniques. We applied our method to automatically track people when they are visible in our mid-resolution side-view cameras (see center images of Fig.1). At each time t, the tracker outputs the head locations in the image plane and the head poses o_t^k (characterized by a pan and tilt angle) of each participant k visible in the side view cameras. Note that when a person is standing at the slide-screen or whiteboard, his head pose was not estimated (and his VFOA as well).

People location x_t^k: We assumed that the location x_t^k of a participant k is a discrete index which takes four values: seat k when he is seated, or the center of one of the three presentation areas A, B or C showed in Fig.1. This variable is extracted in the following simple way. People are tracked from the side view camera using our head pose tracking algorithm when they are seated, they stand-up, or they come back to their seat. When people are away from their seats to make presentations (when they are not visible on the side cameras anymore), they are assumed to stand in one of the area A, B or C. In this case, they are localized from the central camera view using area motion energy measures defined as the proportion of pixels in the image area whose absolute intensity differences between consecutive image is above a threshold. The motion energy measures are computed for each of the standing area and the standing person location is estimated as the area with the highest energy or his previous standing location when the energy measures of all standing areas are too small (the person stays still).

The projection screen activity a_t: As motivated in the next Section, we used the time that elapsed since the last slide change occurred[2] as slide activity feature a_t. Thus, we need to detect the instants of slide changes, from which deriving a_t is straightforward. Slide changes are detected by processing the image area containing the slide screen in the video from the central view camera (cf Fig. 1). We exploited a compressed domain method [33] which recovered 98% of the events with very few false alarms. 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. Slide transitions do not involve translational motion, thus inducing high residual coding bit-rate. Thus, we estimated slide changes by detecting, in the slide area, temporally localized peaks of residual coding bit rate. This allowed us to eliminate false alarms due to people passing in front of the projection screen. Note that similar results were achieved using motion energy measures as described for person localization. The main advantage of [33] is that it runs 50 times faster than real-time.

V. Context modeling

We use a Dynamic Bayesian Network (DBN) to address contextual VFOA recognition. From a qualitative point of view, a DBN expresses the probabilistic relationships between variables and reflects the assumptions we make on the way random variables are generated and their stochastic dependencies. Thus, the design of our DBN model requires us to define the variables representing the context and how they influence the VFOA recognition.

One typical approach for contextual event recognition is to use an Input-Output Hidden Markov Model (IO-HMM). In this model all observation variables related to the VFOA behavior (other than the head pose, which is the primary observable for VFOA recognition) are used to form the context. They influence the recognition of person k's VFOA at time t (denoted f_t^k) by allowing the modeling of a contextual dynamical process $p(f_t^k|f_{t-1}^k, a_t, x_t, s_t)$ and of a contextual observation model $p(o_t^k|f_t^k, a_t, x_t, s_t)$ relating the head pose of the person to his visual focus. Learning such contextual models is often difficult, especially if the variables defining the context are continuous values. In [34], we modeled separately the influence of each cue on their related VFOA targets (i.e. speaking cues would model how a person looks at other people and the slide activity variable would influence the gazing at the slide screen), but the interdependencies between cues were not taken into account. Also, while such an IO-HMM model can be computationally effective [34], its structure may not reflect our understanding of the 'logical' interactions between the VFOA and the context, and in several cases, it might be useful to introduce hidden variables to define the context. This happens when some context of the variable of interest is more conveniently represented by a semantic concept that can not be observed directly but whose impact on the VFOA events is more intuitive, or that represents a less noisy and discrete version of observations. The conversational event that we will introduce later is such an example. Note that the hidden context variables need to be estimated. While such estimation usually relies on appropriate contextual observations, it also depends on the VFOA variables themselves, reflecting that

[2] A slide change is defined as anything that modifies the slide-screen 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).

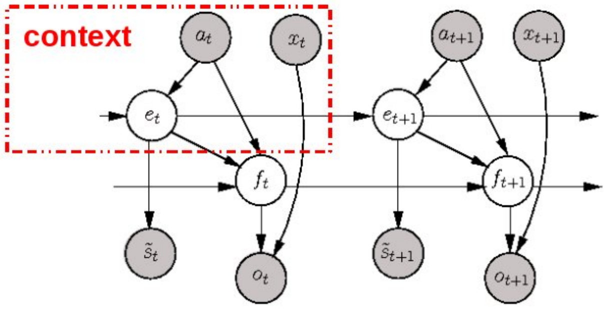

Fig. 2. Dynamic Bayesian network model. Observation are shaded, hidden variable are unshaded.

context and VFOA events are two interrelated concepts: on the one hand, context guides the estimation of the VFOA; on the other hand, people's VFOA can play the role of bottom-up observations to infer the context. For instance, when somebody is speaking, it is more probable that people will look at him; conversely, if a person is the visual focus of several other people, the chances that he is actually speaking are higher.
The DBN for VFOA recognition that we propose follows several of the above modeling considerations. It is given in Fig. 2, and qualitatively explained in the next subsection.

A. *VFOA context modeling*

We defined two main types of VFOA context: the interaction context, and the task context. Below, we introduce the variables we used to define context and their role in the model.

Interaction context: The interaction context is defined by the understanding of the activities of others and their impact on the VFOA. In meetings, it is mainly dictated by the self-regulatory structure of the discourse, where the behavior of one person is constrained by the overall behavior of others. In our model, it is taken into account through the following elements.

Location context: Localizing people is essential in determining where a person needs to look in order to gaze at particular people. Knowing the number of participants influences the number of potential targets to take into account, while the location indicates how much the person has to turn his head. This context will be taken into account in the observation likelihood term $p(o_t^k|f_t^k, x_t)$ of the DBN.

Conversation context and conversational events: The behavior of people is characterized by two main variables: whether they speak, and where they look (their VFOA). Thus, all of these variables (the speaking status of all people, $(s_t^l)_{l=1..4}$ and the VFOA of others, $(f_t^l)_{l=1..4, l\neq k}$) could be used to define the conversational interaction context of person k's VFOA in a IO-HMM fashion, as introduced earlier. However, such contextual information is partly hidden (the VFOA state of others are unknown), and due to the large dimension of such context, the modeling of the statistical relationships between all these variables and a person's VFOA might not be intuitive and become challenging.
Thus, to condense and model the interactions between people's VFOA and their speaking statuses, we introduce as conversation context a discrete hidden variable $e_t \in \mathcal{E}$ called 'conversational event'. It provides the state of the current meeting conversation structure, defined by the set of people currently holding the floor. Since we are dealing with meetings of up to 4 people, the set $\mathcal{E}$ comprises 16 different events E_i. In practice, however, what matters and characterizes each event (e.g. to define the VFOA context models, see Section VI-B) are the conversation type $ty(E_i) \in \{silence, monologue, dialog, discussion\}$ and the size of the set $\mathcal{I}(E_i)$ of people actively involved in the event (e.g. the main speakers for a dialog).

The role of the conversational event variable can be deduced from our DBN structure (Fig. 2). The main assumption is that the conversational event sequence is the primary hidden process that governs the speaking patterns and influences the dynamics of gaze (i.e. people's utterances and gaze patterns are independent given the conversational events). Given a conversational event, the speaker (or speakers) is clearly defined, which allows a simple modeling of the speaking observation term $p(\tilde{s}_t|e_t)$. At the same time, through our modeling of the VFOA dynamics term $p(f_t|f_{t-1}, e_t, a_t)$, the conversational event will directly influence the estimation of a person's gaze by setting a high probability for looking at the people actively involved in the conversational event.

Task-based context: modeling presentation context for VFOA recognition. People do not always gaze at the current speaker due to the presence of visual attractors associated with the task being performed either by the individual or by the group [17]. Modeling such task context is thus important to overrule or modulate the VFOA trends associated with normal conversations. In work meetings, slide-based presentation is an important example of such a group task where people often look at the slide rather than the speaker.
One approach to model presentation context consists of using a binary variable to identify when people are involved in a presentation. This approach has been used in [5], [22], where the goal is to identify the different meeting group activities and use them as context to guide other processing [23]. However, although presentations may last for several tens of minutes, the influence of presentation context on the VFOA behavior manifests itself mainly when new information is displayed. Intuitively, right after a slide change people look at the slide and then progressively get back to normal conversation mode. This timing and progressive effect, which lasts for around 100 seconds after each slide change, is clearly visible in our data, as demonstrated by the graphs in Fig. 6. Thus, the use of a binary indicator variable as presentation context for VFOA recognition is too crude. Instead, we proposed to use the time a_t that elapsed since the last slide change, and one novelty of our approach is to model the impact of this timing information on the gaze pattern dynamics $p(f_t|f_{t-1}, e_t, a_t)$ of our DBN, as described in Section VI.

B. *System overview*

Our approach can also be defined from a system perspective, which introduces other aspects to the modeling (related to parameters setting and adaptation). A system representation is depicted in Fig. 3. The main part is the DBN model we have

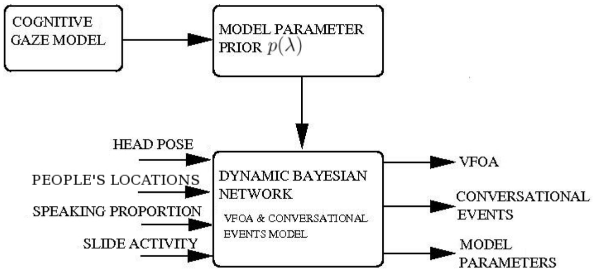

Fig. 3. Approach overview.

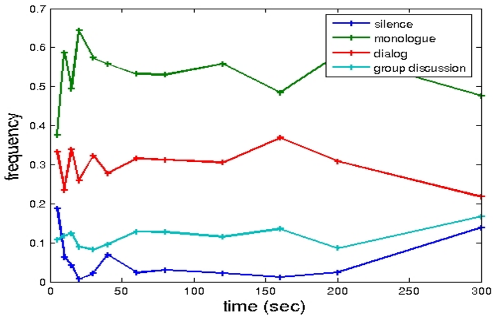

Fig. 4. Meeting event frequency given the time since the last slide change.

just described. This model takes as input a set of observation variables as well as the initial values for the model parameters λ, from which the set of the hidden variables are inferred, and an estimate of λ is given. The observations comprise the head pose of all participants $o_t = (o_t^1, o_t^2, o_t^3, o_t^4)$, their location x_t, their speaking time within a temporal sliding window $\tilde{s}_t = (\tilde{s}_t^1, \tilde{s}_t^2, \tilde{s}_t^3, \tilde{s}_t^4)$, and the presentation activity variable a_t. 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 and the conversational event $e_t \in \mathcal{E}$. Importantly, the graph in Fig. 3 enhances the fact that some prior probability $p(\lambda)$ on the DBN parameters has to be defined. In particular, defining a prior on the parameters relating the people's head pose observations to their gaze at different VFOA target appeared to be crucial to automatically adapt the DBN model to the specific head orientations of individuals in meeting. To set these prior values, we used an improved version of the cognitive gaze model we presented in [15], as described later in Section VI-D.

VI. Joint multi-party VFOA modeling with a DBN

To estimate people's joint focus, we rely on a DBN corresponding to the graphical model displayed in Fig. 2, and according to which $p(f_{1:T}, e_{1:T}, \lambda, o_{1:T}, \tilde{s}_{1:T}, a_{1:T}, x_{1:T})$ the joint distribution of all variables can be factorized up to a proportionality constant as:

$$p(\lambda) \prod_{t=1}^{T} p(o_t|f_t, x_t) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t) \tag{1}$$

where the different probabilistic terms are described in the following.

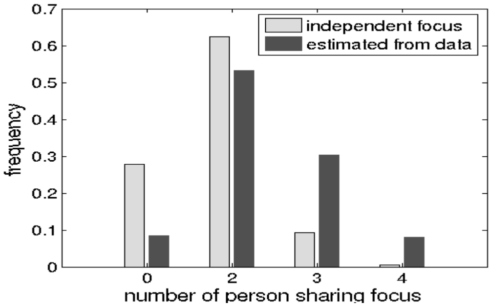

Fig. 5. Shared focus: distribution of frames where n persons are focused on the same VFOA target. Light bars: distribution c_n assuming people VFOA are independent. Dark bars: distribution d_n measured on the data.

A. *Conversational events dynamics*

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

$$p(e_t|e_{t-1}, a_t) = p(e_t|e_{t-1}) p(e_t|a_t). \tag{2}$$

The first term $p(e_t|e_{t-1})$ models the conversational event temporal transitions. It was defined to enforce temporal smoothness, by assuming that the probability to remain in the same event was 3 times higher than to transit to any other event. The second term models the prior probability of the conversational events given the slide context. To avoid over-fitting to specific meeting situations, we assumed that it only depends on the conversation event types. Fig. 4 gives the plot of the priors learned from our data. As can be seen, monologues and dialogs are more frequent than silences and group discussions. Also, silences are much more probable right after and a long time after a slide change, while monologues exhibit the reverse behavior.

B. *The VFOA dynamics*

We assume the following factorized form for the VFOA dynamics:

$$p(f_t|f_{t-1}, e_t, a_t) \propto \Phi(f_t) p(f_t|f_{t-1}) p(f_t|a_t, e_t) \tag{3}$$

where the different terms are described below.

1) The shared focus prior $\Phi(f_t)$: This term models people's inclination to share VFOA targets[3]. Fig. 5 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 more often share the same focus than if their focus was considered independent. Thus, we have set $\Phi(f_t)$ as:

$$\Phi(f_t) = \Phi(SF(f_t) = n) \propto \frac{d_n}{c_n} \tag{4}$$

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. 5. Qualitatively, the shared focus prior will favor states with shared focus according to the distribution observed on training data.

[3] The factorization made in Eq.3 assumes that $p(f_t|f_{t-1})$ and $p(f_t|a_t, e_t)$ model the effect of the conditional variables on the current focus, and the group prior $\Phi(f_t)$ models all the other dependencies between the VFOA of the meeting participants.

2) The joint VFOA temporal transition $p(f_t|f_{t-1})$: It is modeled assuming the independence of people's VFOA states given their previous focus, i.e. $p(f_t|f_{t-1}) = \prod_{k=1}^{4} p(f_t^k|f_{t-1}^k)$. The VFOA dynamics of 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 with prior values for these transitions defined to enforce temporal smoothness (cf Section VI-D).

3) The VFOA meeting contextual priors $p(f_t|a_t, e_t)$: The introduction of this term in the modeling is one of the main contribution of the paper. It models the prior on focusing at VFOA targets given the meeting context (conversational event, slide activity). To define this term, we first assume the conditional independence of people's VFOA given the context:

$$p(f_t|a_t, e_t) \propto \prod_{k=1}^{4} p(f_t^k|a_t, e_t). \tag{5}$$

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) more than at listeners, (ii) during presentations, people tend to look at the projection screen; (iii) the gazing behavior of the persons directly involved in a conversation event (for example two speakers in a dialog) might be different than that of the other people. Thus, what matters for the learning is the event type, and whether the subject (whose focus we model) and his human visual targets are involved or not in the event. We thus 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 $\mathcal{FT} = \{slide, table, unfocused, inv, not\_inv\}$, as a function of the event e. Defining i is straightforward: $i(l, e) = l$ if $l \in \{slide, table, unfocused\}$, and $i(l, e) = ki(l, e)$ if l is a human visual target. Then, we learn from training data the table $T(i, ki, ty, a) = p(i|ty, a, ki)$ providing the probability for a participant whose involvement status in a conversational event of type ty is ki, of looking at a VFOA target type i (either an environmental target or a participant involved in the event), given the event type and the time a since the last slide change. Then, the prior model is simply defined as: $p(f_t^k = l|a_t = a, e_t = e) \propto T(i(l, e), ki(k, e), ty(e), a)$.
The table $T(i, ki, ty, a)$ can be learned directly by accumulating the corresponding configuration occurrences in the training set and normalizing appropriately. Then, to obtain smoother versions of the contextual priors, we fitted by gradient descent functions of the form

$$T(i, ki, ty, a) \approx g_{i,ki,ty}(a) = \vartheta_1 e^{-\vartheta_2 a} + \vartheta_3 \tag{6}$$

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, ty)) provided a good approximation of the observed exponential increase or decrease of probability as a function of the elapsed time a. Fig. 6a gives an example of the fit.

Fig. 6b,c give interesting examples of the fitted priors[4] when the conversational event is of type $ty = dialog$. For the target 'Table', we can see that its probability is always high and not very 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 behavior depending on whether the person is involved in the dialog or not: people involved in the dialog focus sometimes at the side participants whereas a side participant seldom looks at the other side participants.

C. Observation models

They correspond to the terms $p(\tilde{s}_t|e_t)$ and $p(o_t|f_t)$, and are described below.

1) The speaking proportion observation model: This audio model is defined as:

$$p(\tilde{s}_t|e_t = E_j) = \prod_{k=1}^{4} \mathcal{B}(\tilde{s}_t^k; L\eta_{j,k}, L(1 - \eta_{j,k})) \tag{7}$$

were we have assumed that people's speaking proportion $\tilde{s}_t^k$ are independent given the event E_j, and $\mathcal{B}(x, p, q)$ denotes a Beta distribution defined as $\mathcal{B}(x, p, q) \propto x^{p-1}(1 - x)^{q-1}$. The parameter $\eta_{j,k}$ is the expected probability that person k speaks during the W frames defining the event E_j, while L controls the skewness of the distribution. The values of $\eta_{j,k}$ were set by assuming that the speakers in an event would equally hold the floor around 95 % of the time while non speakers would provide back-channels 5% of the time. For instance, for a dialog between person 2 and 3 (event E_8), we have $\eta_8 = [0.05, 0.475, 0.475, 0.05]$. We set $L = 10$ to have not too peaky conversational event likelihoods.

2) The head pose observation model: This term $p(o_t|f_t, x_l)$ is the most important term for gaze estimation, since people's VFOA is primarily defined by their head pose. Assuming that given their VFOA state, people's head poses (defined by pan and tilt angles) are independent, this term can be factorized as $p(o_t|f_t, x_t) = \prod_{k=1}^{4} p(o_t^k|f_t^k, x_t)$. To define $p(o_t^k|f_t^k, x_t)$, the observation model of person k, notice that we have to model what should be the head pose when person k is looking in the 9 directions corresponding to: the 3 objects (table, slide screen, white board); a person in any of the 3 other seats, and a person in any of the 3 standing locations. However, at a given time instant, from the 6 latter directions, only the ones occupied by people represent potential directions. Thus, in order to distinguish between looking at semantic targets and looking in a given direction, we introduce the function $d(i, x)$ that indicates which direction corresponds to looking at target

[4] This example was learned from data involving only seated people, where there is no gazing at the white-board.

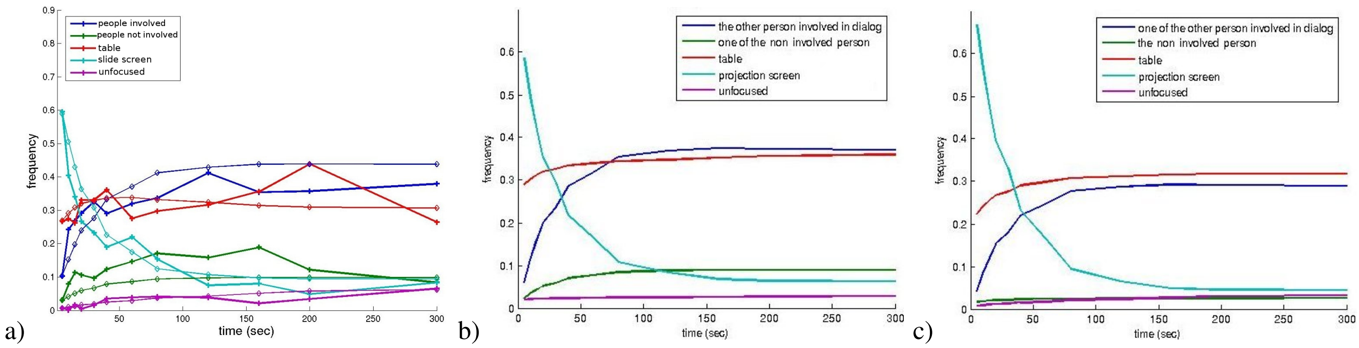

Fig. 6. Fitted contextual prior probabilities of focusing to a given target, in function of time a_t since last slide change. a) Measured probabilities and fit when the conversational event e_t is a monologue and the subject is not involved (i.e. is a listener). b,c) Fit when the conversational event is a dialog for b) a person involved in the dialog and c) a person not involved.

i given the location context x. For static objects, there is a unique direction. For a person i, this direction depends on his location x^i. Assuming that for a given direction the spread of head poses can be modeled by a Gaussian, we have:

$$p(o_t^k|f_t^k = i, x_t) = \mathcal{N}(o_t^k; \mu_{k,d(i,x_t)}, \Sigma_{k,d(i,x_t)}), \qquad (8)$$

where $\mu_{k,d(i,x_t)}$ represents the mean head pose when the person at seat k looks in the direction indexed by $d(i, x_t)$, and $\Sigma_{k,d(i,x_t)}$ is the Gaussian covariance. Alternatively, when the person is unfocused, we model the pose distribution as a uniform distribution $p(o_t^k|f_t^k = \text{ unfocused }) = u$.

D. Priors on the model parameters

We can define some prior knowledge about the model by specifying a distribution $p(\lambda)$ on the model parameters. In this paper, we are only interested in the estimation of the parameters involving the VFOA state variable, and all other parameters were set or learned a priori (as described in previous subsections). This is motivated by the fact that the head poses defining gazing behaviors is more subject to variations due to people's personal way of gazing at visual targets [15]. Thus, the parameters to be updated during inference were defined as $\lambda = (\lambda_k)_{k=1,\dots,4}$, with $\lambda_k = (B_k, \mu_k, \Sigma_k)$, i.e. for each person k, the VFOA dynamics B_k, and the means $\mu_k = (\mu_{k,j})_{j=1\dots9}$ and covariance matrices $\Sigma_k = (\Sigma_{k,j})_{j=1\dots9}$ defining the head pose observation model. We describe below some elements on the definition of the prior model. More details can be found in the supplementary material.

VFOA transition probability prior. We use a Dirichlet distribution $\mathcal{D}$ on each row of the transition matrices, and defined the prior values of the transition parameters $b_{k,i,j}$ to enforce smoothness in the VFOA sequences, i.e. we set a high probability (0.9) to remain in the same state and spread the remaining probability on the transitions to the other VFOA targets.

Head pose prior. The prior for the Gaussian mean $\mu_{k,j}$ and covariance matrix $\Sigma_{k,j}$ is the Normal-Wishart distribution. The parameters for this distribution were set as in [15], and we only detail the setting of $m_{k,j}$, the prior values of the means $\mu_{k,j}$.

Defining the $m_{k,j}$ using a Cognitive model. Due to the complexity of our task, there are 36 prior (pan,tilt) head pose values $m_{k,j}$ (4 persons times 9 potential gaze directions) that we need to set and which play a crucial role in the model. Hence, setting the parameters manually as in Otsuka *et al* [16] becomes tedious and prone to errors. In this paper, we employ a fully automatic approach. It builds upon our previous model ([15], [35]) which relies on studies in cognitive sciences about gaze shift dynamics [30], [36]. 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, as illustrated in Fig. 8. Thus, denoting by m_k^{ref} the reference direction of person k, by $m_{k,j}^{gaze}$ the gaze direction associated with looking at place j, the corresponding head pose mean $m_{k,j}$ is defined by:

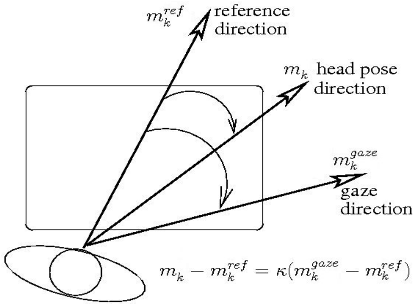
m_k^{ref} reference direction
m_k head pose direction
m_k^{gaze} gaze direction
$m_k - m_k^{ref} = \kappa(m_k^{gaze} - m_k^{ref})$

Fig. 8. Relationship between the gaze direction associated with a VFOA target and the head pose of a participant. Assuming that the participant has a preferred reference gazing direction (e.g. in front of him), the gaze rotation towards the VFOA target is made partly by the head and partly by the eye. Following [15], we used $\kappa = 0.5$ for the pan and $\kappa = 0.4$ for the tilt.

$$m_{k,j} - m_k^{ref} = \kappa(m_{k,j}^{gaze} - m_k^{ref}). \qquad (12)$$

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.

Manual reference setting: the first alternative, used in [15], 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 splits a

```
1) Initialization: first estimation of the conversational events. $n = 0$:
       $\hat{e}_{1:T} = \arg\max_{e_{1:T}} \prod_{t=1}^{T} p(\tilde{s}_t|e_t)p(e_t|e_{t-1})p(e_t|a_t)$    (9)
2) Main loop: repeat a) and b) until $n = N_{iter}$
    a) estimate the VFOA state sequence and model parameters according to
       $(\hat{f}_{1:T}, \hat{\lambda}) = \arg\max_{(f_{1:T},\lambda=(\mu,\Sigma,B))} p(\lambda) \prod_{t=1}^{T} p(o_t|f_t, x_t, \mu, \Sigma)p(f_t|f_{t-1}, B)p(f_t|\hat{e}_t, a_t)$    (10)
    b) estimate the conversational event sequence as $\hat{e}_{1:T} = \arg\max_{e_{1:T}} \prod_{t=1}^{T} p(\tilde{s}_t|e_t)p(e_t|e_{t-1})p(e_t|a_t)p(\hat{f}_t|e_t, a_t)$    (11)
3) Final estimate: $(\hat{f}_{1:T}, \hat{e}_{1:T}, \hat{\lambda})$
```

Fig. 7. Approximate inference algorithm.

person's gazing space, but allows to adapt the reference to each individual. From another point of view, it can be seen as the head direction which minimizes the energy to rotate the head during the meeting.

E. Bayesian model inference

The estimation of the the conversational events $e_{1:T}$, the joint VFOA $f_{1:T}$, and the model parameters λ from the observations $(a_{1:T}, s_{1:T}, o_{1:T}, x_{1:T})$ is obtained by maximizing the posterior probability $p(f_{1:T}, e_{1:T}, \lambda|a_{1:T}, s_{1:T}, o_{1:T}, x_{1:T})$. Given our model, exact inference can not be conducted. Thus, we exploited the hierarchical structure of our model to define an approximate inference method consisting in the iteration of two main steps: estimation of the conversational events given the VFOA states, and conversely, estimation of the VFOA states and model parameters given the conversational events. This procedure guarantees at each time step to increase the joint posterior probability distribution. Similar inference procedures have been proposed in [37] for switching linear dynamic systems. Fig. 7 summarizes the different steps of our algorithm.

Conversational event estimation: It is conducted with a Viterbi algorithm at two places. First, in the initialization step: conversational events are inferred solely from the speaking variables and the slide context (see Eq.9). 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.11. VFOA states play the role of observations and the corresponding term $p(\hat{f}_t|e_t, a_t)$ allows us 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 step 2a, by maximizing the probability $p(f_{1:T}, \lambda|\hat{e}_{1:T}, s_{1:T}, o_{1:T}, a_{1:T}, x_{1:T})$. The inference is conducted by exploiting the maximum a posteriori (MAP) framework of [38], 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. Given our choice of prior, the MAP procedure relies on an EM algorithm similar to the one used in HMM model parameter learning. It is described in the supplementary material, where details are given on how to adapt the standard MAP procedure to take into account the prior on the joint VFOA state (i.e. the term $\Phi(f_t)$) and the presence of moving targets. Below, we enhance the important role of context in this adaptation procedure.

Typically, in the E-step of the MAP algorithm, we compute the expected value $\gamma_{i,t}^k$ for participant k to look at target i at time t given the observations and the current parameter values:

$$\gamma_{i,t}^k = p(f_t^k = i|o_{1:T}^k, \hat{e}_{1:T}, a_{1:T}, x_{1:T}, \hat{\lambda}_k) \tag{13}$$

and in the M-step, re-estimate parameters using these expectations. Qualitatively, in the re-estimation, prior values are combined with data measurements to obtain the new values. For instance, assuming there is no moving person in the meeting, we have:

$$\mu_{k,i} \quad = \quad \frac{\tau m_{k,i} + \sum_{t=1}^{T} \gamma_{i,t}^k o_t^k}{\tau + \sum_{t=1}^{T} \gamma_{i,t}^k} \tag{14}$$

i.e. $\mu_{k,i}$ is a linear combination of the prior values $m_{k,i}$ and of the pose observations o_t^k which most likely corresponds to the target i according to $\gamma_{i,t}^k$. With respect to our previous work [15], the only but crucial difference lies in the computation of the expectations $\gamma_{i,t}^k$. While in [15], this term was only depending on the pose values (i.e. $\gamma_{i,t}^k = p(f_t^k = i|o_{1:T}^k, \hat{\lambda}_k)$), here the expectation takes into account the contextual information (slide, conversation, location), see Eq.13, which increases the reliability that a given measurement is associated with the right target in the parameter adaptation process, leading to more accurate parameter estimates.

VII. Evaluation Setup and Experiments

This section describes our experiments. Results will be presented on two types of meetings: first the *static meetings* (4 recordings involving 12 different participants, one hour 30 minutes) in which people were remaining seated during the entire meeting; then the *dynamics meetings* (8 recordings with 20 different people, 3h30 min) involving moving people. This allows us to distinguish and enhance the differences between these two specific situations. Visual illustrations are provided at the end of the Section.

A. Experimental setup and protocol

Our experimental setup and the data used for evaluation were described in Section III-A.

Performance measure: The performances of our algorithms are measured in term of frame based VFOA recognition rate (FRR), i.e. the percentage of frames that are correctly classified. The FRR is computed as the average of meeting FRR. The evaluation protocol we followed is a leave-one-out: in turn one of the recordings is kept aside as evaluation data, while the other recordings are used to learn model parameters, mainly the contextual prior $p(f_t^k|a_t, e_t)$.

Algorithms: Several alternative algorithms have been evaluated and are presented in the next Section. Unless stated otherwise, the following components were used: the dynamics are given by Eq.3, *but the prior* Φ *on the joint VFOA was not used by default* (as discussed later); the cognitive model with automatic reference gaze setting (see Section VI-D); head pose adaptation was used, i.e. joint inference of VFOA states, conversational events, and head pose model parameter was conducted (cf Section VI-E); the model parameters values were the ones provided in Section VI.

Significance tests: We used the McNemar test to evaluate whether the difference between the recognition results of two algorithms is statistically significant. This standard test looks only at the samples where the two algorithms give different results [39]. It accounts for the fact that some data are easier or more difficult (e.g. when the head tracker provides erroneous results) to recognize than others, and allows us to check whether a given improvement is consistent or not. That is, does an algorithm provides almost systematically the same or a better answer. In practice, to ensure independence between VFOA samples, we extracted meetings chunks of 5 minutes separated by one minute intervals[5]. We then performed on these chunks a variant of the McNemar test that can account for correlated data (in the chunks) [39]. According to this test, all the differences between two parameterizations of the algorithm happen to be significant at a p value of 0.01.

B. Results on static meetings

We first provide overall results, and then discuss the impact of different aspects of the model on the results (explicit joint VFOA prior, context, adaptation, model parameters), including a discussion and comparison with the state-of-the-art.

1) Overall results: Table I presents the results obtained with our contextual model. As a comparison, it also provides the recognition results obtained when recognizing the VFOA independently for each person, using the head pose only (this method will be referred to as the 'head pose only' model) As can be seen, our task and data is quite challenging, due mainly to our scenario and the behavior of people, and to the noisiness of the head pose estimation (some heads are rather difficult to track given the image resolution). We can also notice that the task is more challenging for people at seat 3 and 4, with on average a FRR of 8 to 10% less than for seat 1 and 2. This is explained by the fact that the angular spread of the VFOA targets spans a pan angle of around only 90 degrees for seats 3-4 vs 180 degrees for seats 1-2, thus introducing more potential confusion in the pose defining the VFOA targets.

When using the contextual model, the recognition improves significantly, passing from 38.2% to 55.1%, with almost all participants having more than 50% recognition rate except for the person at seat 3 in the 3rd recording[6]. The influence of the speaking activities through the conversational events and of the slide context on the VFOA estimation (and adaptation process), is beneficial, esp. for seat 3 and 4, by helping to reduce some of the VFOA ambiguities.

Fig. 9 gives the average confusion matrices in the two cases for seat 1 and 3 for all recordings. They confirm the higher degree of VFOA confusion for seat 3 (and 4) w.r.t. 1 (and 2) for both the methods, and the general improvement when using context (brighter diagonals and darker off-diagonal terms overall).

2) Influence of modeling factors: We discuss here several elements of the model that can influence the results. The results of these variations are presented in Table II.

Group prior: A first novelty of this paper is to introduce an explicit prior $\Phi(f_t)$ on the joint focus of all meeting participants, which models the tendency for people to look at the same VFOA target. Table II displays the results obtained when adding such a prior to the models with and without context. When this term is used without context, it leads to an important 7.7% increase of the recognition result, demonstrating the validity of the group behavior assumption. However, used in conjunction with the context, we obtain a more modest 0.5% gain[7]. Indeed, the context we have defined accounts for the group effect (encouraging individual people to look at the current speaker or at a recent slide change) to a great extent. The other situations where people share the same visual target (e.g. when somebody refers to a specific object, like an object on the table or part of a slide which was displayed long ago) are probably too rare to lead to a substantial improvement. Since the use of this group prior has a high computational cost we did not use this term in the other experiments.

Influence of the different contexts: We evaluated the contribution of the conversation and presentation contexts to the recognition. When only the slide context is used, the conversational events and speaking nodes are removed from the DBN, and $p(f_t|e_t, a_t)$ reduces to $p(f_t|a_t)$, i.e. a prior on looking at VFOA targets as a function of a_t. When only speech is used, the presentation context is removed, and $p(f_t|e_t, a_t)$ reduces to $p(f_t|e_t)$, i.e. a prior on looking at the table, the slide, and at people in function of their involvement in the conversational event. As can be seen, using the slide context increases the recognition by around 5% (FRR=45.7%), while

[5] We showed with a Chi-square independence test that VFOA samples apart by more than one minute can be considered as independent.

[6] The low performance for this person can be explained by higher head pose tracking errors, as assessed visually, because this person appearance was not well represented in the training data of the head pose tracking appearance models.

[7] Note that this increase was statistically significant, as the group prior improved the recognition result of 14 out of 16 people.

	head-pose only (no contextual cues)					with contextual cues				
recordings	seat 1	seat 2	seat 3	seat 4	average	seat 1	seat 2	seat 3	seat 4	average
1	48.8	53.1	30.5	29.3	40.4	59.3	67.9	53.9	58.4	59.9
2	55.2	34.8	22.2	36	37	67.2	50.5	47.1	54.8	54.9
3	37.1	35.8	18.3	33.6	31.2	70.4	47.5	21.6	58.6	49.5
4	24	58.4	47.9	46.6	44.2	53	59.3	62.3	49.4	56
average	41.3	45.5	29.7	36.4	**38.2**	62.5	56.3	46.3	55.3	**55.1**

TABLE I

OVERALL VFOA RECOGNITION PERFORMANCE ON STATIC MEETINGS, WITHOUT (LEFT) AND WITH (RIGHT) CONTEXTUAL MODEL.

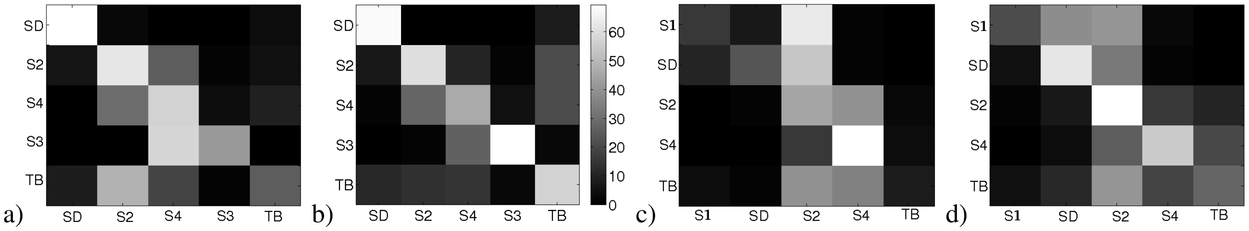

Fig. 9. VFOA recognition confusion matrices without context (a and c) and with context (b and d), for seat 1 (a and b) and 3 (c and d). For each seat, the VFOA target order in the matrix is organized by increasing pan values, and the table is put as the last VFOA target. $Si, i = 1, 2, 3, 4$ stand for seat 1,2,3,4; SD stands for slide screen, and TB stands for the table. More confusions are exhibited for seat 3, and context reduces confusion overall.

Model	seat 1	seat 2	seat 3	seat 4	average
Contextual with group prior Φ	61.7	56.1	48.5	55.9	55.6
Contextual	62.5	56.3	46.3	55.3	55.1
Head pose only with group prior Φ	47.1	50.0	40.1	46.6	45.9
Head pose only	41.3	45.5	29.7	36.4	38.2
Contextual: slide only	50.7	48.7	38.3	45.2	45.7
Contextual: speech/convers. events only	62.5	48.8	43.6	48.1	50.7
Contextual: no adaptation	53	52.5	44.3	53.9	50.9
Head pose only: no adaptation	48.8	44.5	32.9	38.8	40.0
Contextual: manual head pose reference	62.8	56.4	47.3	48.1	53.6
Contextual: W=1sec	62.3	56.2	46.2	56.5	55.3
Contextual: W=9sec	62.1	55.43	45.8	55.5	54.7

TABLE II

VFOA RECOGNITION RESULTS ON STATIC MEETINGS FOR DIFFERENT MODELING FACTORS. ONLY FACTORS WHICH DIFFER FROM THE DEFAULT ALGORITHM/PARAMETERS (DENOTED 'CONTEXTUAL' MODEL), SEE SECTION VII-A, ARE GIVEN.

the use of the conversation context increases the results by around 10% (FRR=50.7%) w.r.t. the case without context. One explanation for the difference in impact is that the slide context mainly increases the recognition of a single VFOA target, the slide screen (which is already well recognized for seats 1 and 2), while the conversational events improve pose parameter adaptation and set the priors on 3 targets, i.e. the 3 other participants, which represent 40% of the VFOA gaze data. This effect is illustrated in Fig. 10 which depicts the recognition rates per VFOA target category as a function of the context[8].

Importantly, our overall results (FRR=55.1%) shows that modulating the conversational prior by the slide context allows for 4.4% further improvements, justifying the need for the introduction of such group task context. Interestingly enough, we can notice in Fig. 10 that the joint use of conversational events and slide context further and significantly increases the slide screen recognition accuracy w.r.t. the sole use of the slide context. It shows that the model is not simply adding priors on the slide or people targets according to the slide and conversational variables, but that this is the temporal interplay between these variables, the VFOA, and the adaptation process (as shown below) which makes the strength of the model.

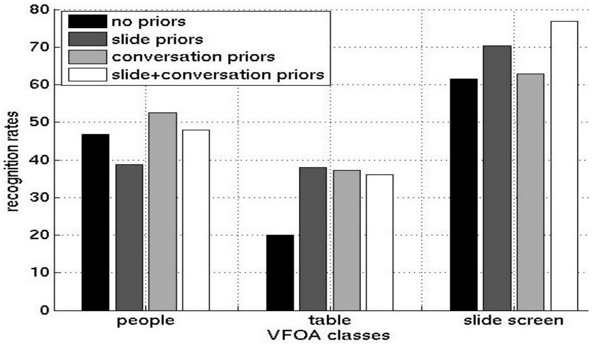

Fig. 10. Effects of different VFOA contexts on the recognition rates of the static meetings, per category.

[8] Notice that the recognition rate for the 'Table' label has greatly increased w.r.t. the no context (head pose only) case. This directly derives from the fact that, in the contextual case, the different models for $p(f_t|e_t, a_t)$ learned from the data include prior on the Table (see Fig. 6 for instance), which is not the case of the no context model shown here.

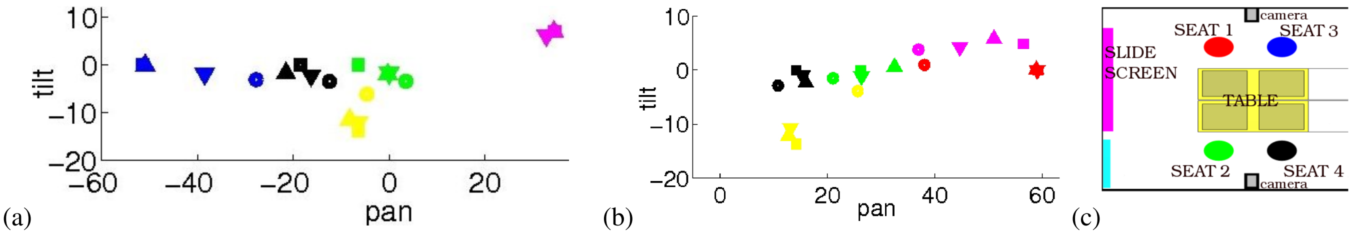

Fig. 11. Effect of adaptation on the Gaussian means defining the VFOA targets for (a) seat 1 in recording 3 (b) seat 3 in recording 2. Labels denote: (□) prior pose value $m_{k,i}$ from the cognitive model - (△) $\mu_{k,i}$ after adaptation without contextual priors - (▽) $\mu_{k,i}$ after adaptation with contextual priors - (◯) empirical mean pose from the person data. Notive that the mean values adapted with context are generally closer to the empirical means. (c) Meeting room configuration and VFOA color codes.

Model adaptation: Recognition improvements w.r.t. the head-pose only case can be mainly attributed to two related aspects of our model: first, context variables introduce direct prior on the VFOA targets when inferring the VFOA states; second, we perform the joint inference of the VFOA and conversational events along with the head pose model parameters (i.e. the unsupervised adaptation of these parameters, cf Section VI-D). To distinguish between the two, we removed the adaptation process, by setting the model parameters to their prior values.

The results are given in Table II. As can be seen, unsupervised adaptation slightly degraded the results in the no context case, while it allowed a 4.2% gain in the contextual case. Indeed, in the no context case, where adaptation is solely based on the head pose observations, there is a risk that the model parameters drift to wrong values due to the complexity of our experimental setup with high potential ambiguities in the head poses defining the VFOA targets. In particular, due to these ambiguities, the mean pose value $\mu_{k,i}$ associated with the VFOA target i can drift to the mean of another target j, thus introducing important recognition errors. This is illustrated in Fig. 11, which compares the adapted mean pose values to the empirical means of the data. For instance, on the Fig. 11a example, the prior value for looking at seat 4 (black □) is wrongly adapted (black △) towards the measured pose for looking at seat 3 (blue ◯) rather than to its corresponding pose (black ◯), while the prior pose value for looking at seat 3 does not evolve (the blue □ and △ are at the same place). As a result, for the person of this example, looking at seat 3 will mainly be interpreted as looking at seat 4. A similar effect can be seen for the second example in Fig. 11b: the pose prior value □ for the green target adapts (green △) toward the data of the red target (red ◯).

When the contextual model is used, such situations are avoided, since during inference, head poses observations are better associated with the correct VFOA target thanks to the contextual prior, as explained in Section VI-E, resulting in better adapted means (see the ▽ symbols for the above cases in Fig. 11) and overall better results.

Cognitive model head pose reference: We introduced a fully automatic approach to define our gaze model and in particular to set the reference pose used to build the relationship between a person's gaze and his head pose (cf Section VI-D). As a comparison, Table II provides the results when using our previous manual approach [15]. As can be seen, the automatic approach performs 1.5% better than the manual one. Although significant, this difference is mainly due to the performance increase of one person, for whom the automatic method, being person dependent, corrects a bias introduced by the head pose tracker and produces better head pose prior values. Excluding this person, the performance increase is only of 0.2%. Nevertheless, this is an interesting result, as this shows the robustness of the automatic approach w.r.t. the variability of people gazing behavior and meeting content.

Window size for measuring speech activity: Results are reported in Table II for window sizes of 1s, 9s (default parameter is 5s). As can be seen, the rates are slowly decreasing as the window size increases. This can be explained by the fact that larger speaking conversational window sizes favors the recognition of more dialog and discussion events which i) spreads the focusing prior on more participants than in monologues, and ii) loses accurate information about when these participants speak. Overall, this slightly reduces the relevance of the events' influence on the VFOA. In other words, using smaller windows, the conversational events capture the floor holding and floor exchange events, whereas larger windows provide a higher level description of the ongoing conversation activity, as was targeted in the investigations about meeting action recognition [5], [40]. Interestingly, this suggests that participants' interactions and the overall meeting structure of turn taking and non-verbal patterns should be described within a framework involving different temporal scales.

Comparison with the state-of-the art: In our previous work [34], we used an IO-HMM structure to model the context. Instantaneous speaking statuses were used to model the tendency to look at speaking people, whereas the slide variable influenced the probability of looking at the slide. A result of 46.7% was achieved on our data. Several factors explain the difference: first, our IO-HMM did not model the interaction between contextual cues (speech and slide); secondly, as explained in Section V, the context in the IO-HMM directly imposes its prior for VFOA estimation, whereas with the DBN, the conversational event context and the VFOA are jointly decoded; lastly, a contextual adaptation process was not exploited. Comparing VFOA results with other works is quite difficult, since with the exception of our data, there is no public dataset. In addition, performance is highly dependent on many factors different from the model, such as the setup and placement of people, the type of meetings, the accuracy of head pose estimation, etc. For instance, recognition rates of 75% and 66% were reported in [14] and [16] respectively, but this was measured in 4 persons' short pure conversation meetings with only the 3 other participants as VFOA targets. Nevertheless, our results obtained using the conversation context only model can be considered as representative of state-of-the-art performance, since it models the integration between speech cues and VFOA, as done in [14] or Otsuka *et al* [16]. As we have seen before, the use of the slide-based presentation activity context, along with the addition of a group prior on focus, leads to a significant increase of 4.9%, definitively showing the interest of our new model.

C. Results on dynamic meetings

Our model can recognize the VFOA of (seated) people even when they look at people that move during the meeting. This allows us to handle meetings with less restrictions than what had been considered so far in the literature.

The model was tested on 8 recordings (for a total of 3h30 min), where to the contrary of the static meetings, participants were mostly standing up to make their slide-based presentations. Table III reports the main results obtained on these data. Despite the challenging conditions, as can be seen from the results obtained without the use of context, the performance are close to the static case (52.3% vs 55.1%). The conclusions drawn with the static case still apply here, although one can notice that overall the differences in performance between the different experimental conditions have shrunk. For instance, the addition of the slide context to the conversation context (i.e the full model) increased performance by 1.4% w.r.t. the conversation context alone (4.1% in the static case). We can also notice that the use of the group prior slightly degrades the overall performance. However, the details shows that results are similar when all people are seated, and that the degradation comes mainly from the standing presentation situations. This can be explained by the fact that we have one less measurement (from the presenter) in such cases, which weakens the reliability of the group prior. In addition, the issue of the focus ambiguity during presentation discussed in the next paragraph also applies for the group prior.

Table III further shows that the results are lower during presentation periods. Indeed, the person standing either in front of the white board or projection screen highly increases the confusion between the visual targets since i) the presenter and the slide screen are the two predominant VFOA targets; ii) they are located in almost the same gazing direction; iii) this direction corresponds to profile views of the seated persons, where head pose estimates are noisier (head pan variations induce only little visual change for profile views, see Fig. 12). Thus, context plays a major role to remove ambiguities. However, as the presentation or conversation context alone favors only one type of focus (either the slide screen or the presenter) to the detriment of the other main VFOA target, our full model is able to reach appropriate compromises between the two context types, leading to overall better results in these periods. Finally, results when nobody is standing (last column of Table III) show that, as expected, in the absence of presentations, conversation context is the main source of improvement.

Finally, notice that although this did not happen in our meetings, our approach can easily handle people returning to different seats, as long as the tracking is done properly. The main point to take into account in such cases would be to conduct the gaze model adaptation for each seat that a person occupies.

D. Qualitative discussion of the performance

Beyond the recognition accuracy, qualitative analysis of the results is often useful to understand the behavior of the algorithm. The supplementary material provides several videos and a plot depicting on a one minute segment the recognized conversational events along with the ground truth (GT) and recognized VFOA. It shows that the estimated VFOA sequences are smoother than in the GT: while the average duration between two changes of gaze is 3 seconds in the GT, it is 5.4 seconds in the recognized sequences. This is due to the VFOA dynamics which favor smooth VFOA sequences, in combination with the fact that most of the very brief gaze changes are the effect of glances done through eye-shift only, with no or very subtle head rotation.

Fig. 12 compares several examples of the results obtained with and without context. In the first row, person 3 is commenting a displayed slide; while the presentation context allows to correct the recognized VFOA of person 4, the conversation context simultaneously allows to correct the focus of person 1. The second row illustrates the positive effects of the conversation context on person 1, and of parameter adaptation (on person 3). It also illustrates the typical issue raised by the addition of the table as one VFOA target: due to a slightly over-estimated head tilt, and although the pan value is well estimated, the estimated VFOA for person 4 is 'Table'. Although the contextual model sometimes helps to correct this effect, improvement should come from a more accurate tilt estimate or the use of contextual cues related to the table activity (manipulation of an object or laptop). Note that the table was not used as a target in previous work [14], [20] and only the pan angle was used to represent the gaze.

Finally, the last row illustrates the result of our model in the dynamic situation.

VIII. Conclusion

We presented a DBN for the joint multi-party VFOA and conversational event recognition in task-oriented meetings, from head pose estimated from mid-resolution images and multi-modal contextual information. The model represents the social communication dynamics linking people's VFOA and their floor holding status, while taking into account the influence of the group activity (here slide-based presentations) on the conversation and gaze behaviors. In addition, the use of people's locations as context allowed us to handle looking at people moving to the white-board to make presentations. Context was shown to provide informative priors on people's focus, and favor a correct adaptation of the parameters relating people's head pose to their focus. Experiments on a 5 hour challenging dataset involving 20 different people demonstrated the benefits of taking into account all types of context (conversation, presentation activity, persons' locations) for the VFOA

Model	seat 1	seat 2	seat 3	seat 4	average	presentation	all seated
Contextual	55.9	58.2	48.3	46.6	52.3	47.3	54.5
Contextual with group prior Φ	54.6	56.2	48.0	49.1	52.0	46.4	54.4
Head pose only	37.9	40.6	31.2	31.9	35.4	34.2	35.9
Contextual: slide presentation only	53.1	54.7	44.7	42	48.6	44.3	51.0
Contextual: conversational events only	55.3	56.3	45.6	46.5	50.9	44.4	54.3
Contextual: no adaptation	56.8	54.2	49.1	40.7	50.2	43.4	53.0

TABLE III

VFOA RECOGNITION RESULTS FOR THE DYNAMIC MEETINGS AND DIFFERENT MODELING FACTORS. THE LAST TWO COLUMNS PROVIDE THE FRAME RECOGNITION RATES COMPUTED DURING PRESENTATIONS (ONE PERSON STANDING) AND DURING THE REST OF THE MEETINGS (ALL PEOPLE ARE SEATED). BY DEFAULT, THE GROUP PRIOR Φ IS NOT USED.

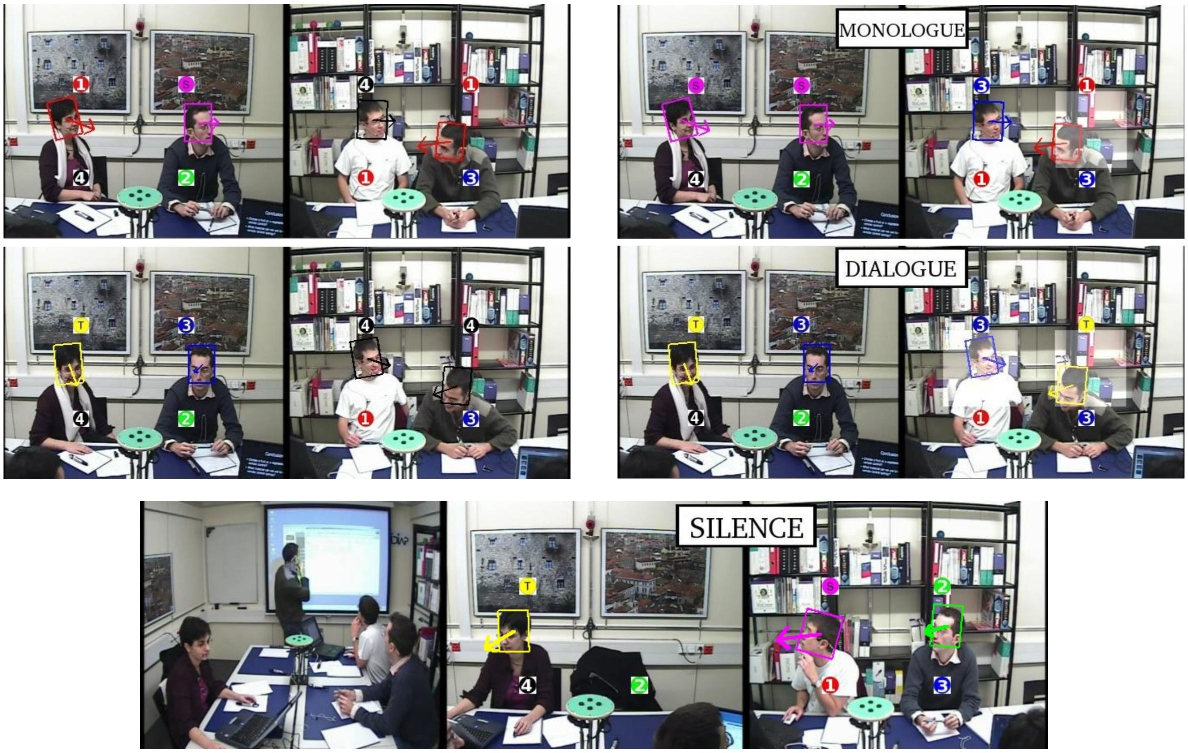

Fig. 12. Rows 1 and 2: effect of the contextual cues on the VFOA recognition. First column without context, second column with context. Last row: a result example on dynamic meetings. The box surrounding people's head and the arrows denotes the head location and pose estimated using our tracking system. Their color gives the estimated focus of the corresponding person VFOA (see color codes in Fig. 1), which is further stressed by the tag above the person's head. On the body of each person, a tag gives his seating location. A semi-transparent white square indicates when a person is speaking.

recognition, achieving a recognition rate of 55% on static meetings, and 52.3% on the dynamic ones.

Several research directions can be investigated to improve performance of our model. First, better head pose estimates -e.g. by processing higher resolution images- would lead to better performances, as we showed in [15]. Secondly, other contextual activity cues could easily be introduced in the model, such as for instance, when people manipulate objects, or, importantly in meetings nowadays, the use of laptops. Although their automatic extraction might not be trivial, they would deliver valuable information to disambiguate the visual target 'Table' from other focuses. Finally, investigations would be needed to see whether (and how) taking into account more subtle findings from social psychology related to gaze behaviors (e.g. looking at speakers is more important at the beginning or end of utterance/monologues) or people's personality (e.g. introversion) could have a significant impact on the VFOA recognition accuracy.

AKNOWLEDGEMENTS

This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the IST European projects FP6 AMIDA (Augmented Multi-Party Interaction with Distance Access). The authors thank Dr. Daniel Gatica-Perez, Dr. Hayley Hung, and Dinesh Jayagopi from the Idiap Research Institute for their helpfull discussions.

REFERENCES

[1] P. Wellner, M. Flynn, and M. Guillemot, "Browsing recorded meetings with Ferret," in *Works. on Machine Learning and Multimodal Interaction*, 2004.

[2] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta, "The AMIDA automatic content linking device: Just-in-time document retrieval in meetings," in *Works. on Machine Learning and Multimodal Interaction*, 2008.

[3] T. Kleinbauer, S. Becker, and T. Becker, "Combining multiple information layers for the automatic generation of indicative meeting abstracts," in *European Works. on Natural Language Generation*, 2007.

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

[5] 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*, vol. 27, pp. 305–317, 2005.

[6] S. Favre, H. Salamin, A. Vinciarelli, D. H. Tur, and N. P. Garg, “Role recognition for meeting participants: an approach based on lexical information and social network analysis,” in *ACM Int. Conf. on Multimedia*, 2008.

[7] A. Kendon, “Some functions of gaze-direction in social interaction,” *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[8] N. Jovanovic and H. Op den Akker, “Towards automatic addressee identification in multi-party dialogues,” in *5th SIGdial Works. on Discourse and Dialogue*, 2004.

[9] O. Kulyk, J. Wang, and J. Terken, “Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings,” in *Works. on Machine Learning for Multimodal Interaction*, 2006.

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

[11] T. Ohno, “Weak gaze awareness in video-mediated communication,” in *Conf. on Human Factors in Computing Systems*, 2005, pp. 1709–1712.

[12] C. Morimoto and M. Mimica, “Eye gaze tracking techniques for interactive applications,” *Computer Vision and Image Understanding*, vol. 98, pp. 4–24, 2005.

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

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

[15] S. Ba and J.-M. Odobez, “Recognizing human visual focus of attention from head pose in meetings,” *IEEE Transaction on Systems, Man, and Cybernetics.Part B*, vol. 39(1), pp 16–34, Feb 2009. 2008.

[16] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, “Conversation scene analysis with dynamic bayesian network based on visual head tracking,” in *Int. Conf. on Multimedia & Expo*, 2006.

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

[18] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen, “Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features,” in *Int. Conference on Multimodal Interfaces*, 2005.

[19] L. Chen, M. Harper, A. Franklin, T. Rose, and I. Kimbara, “A Multimodal Analysis of Floor Control in Meetings,” in *Workshop on Machine Learning for Multimodal Interaction*, 2005.

[20] 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 *Int. Conference on Multimodal Interfaces*, 2005, pp. 191–198.

[21] 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 *Works. on Machine Learning for Multimodal Interaction*, 2005.

[22] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, “Modeling individual and group actions in meetings with layered hmms,” *Transactions on Multimedia*, vol. 8(3), pp. 509–520, 2006.

[23] P. Dai, H. Di, L. Dong, L. Tao, and G. Xu, “Group interaction analysis in dynamics context,” *Transactions on Systems, Man, and Cybernetics-Part B*, vol. 39(1), pp. 34–43, 2009.

[24] S. Basu, “Conversational scene analysis,” Ph.D. dissertation, Massachusset Institute of Thechnology, 2002.

[25] M. Katzenmeir, R. Stiefelhagen, and T. Schultz, “Identifying the addressee in human-human-robot interactions based on head pose and speech,” in *Int. Conference on Multimodal Interfaces*, 2004.

[26] N. Jovanovic, R. op den Akker, and A. Nijholt, “Addressee identification in face-to-face meetings,” in *the 11th Conf. of the European Chapter of the Association for Computational Linguistics (ACL)*, 2006.

[27] K. Smith, S. Ba, D. Gatica-Perez, and J. Odobez, “Tracking attention for multiple people: Wandering visual focus of attention estimation,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(7), pp. 1212 – 1229, 2008.

[28] M. Siracusa, L. Morency, K. Wilson, J. Fisher, and T. Darrell, “A multimodal approach for determining speaker location and focus,” in *Int. Conf. on Multimodal Interfaces*, 2003.

[29] M. Voit and R. Stiefelhagen, “Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios,” in *Int. Conf. on Multimodal Interfaces*, 2008.

[30] M. Hayhoe and D. Ballard, “Eye movements in natural behavior,” *Trends in Cognitive Sciences*, vol. 9(4), pp. 188–194, 2005.

[31] N. Jovanovic, “To whom it may concern: adressee identification in face-to-face meetings,” *PhD thesis, Univ. of Twente, Netherlands*, 2007.

[32] S. O. Ba and J.-M. Odobez, “A Rao-Blackwellized mixed state particle filter for head pose tracking,” in *ICMI Workshop on Multimodal Multiparty Meeting Processing*, 2005, pp. 9–16.

[33] C. Yeo and K. Ramchandran, “Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection,” University of California, Berkeley, Tech. Rep. UCB/EECS-2008-79, June 2008.

[34] S. Ba and J. Odobez, “Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues,” in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[35] J.-M. Odobez and S. Ba, “A cognitive and unsupervised MAP Adaptation approach to the recognition of focus of attention from head pose,” in *Int. Conference on Multi-media & Expo*, 2007.

[36] E. G. Freedman and D. L. Sparks, “Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys,” *Journal of Neurophysiology*, vol. 77, pp. 2328–2348, 1997.

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

[38] J. Gauvain and C. H. Lee, “Bayesian learning for hidden Markov model with Gaussian mixture state observation densities,” *Speech Communication*, vol. 11, pp. 205–213, 1992.

[39] V. Durkalski, Y. Palesch, S. Lipsitz, and P. Rust, “Analysis of clustered matched-pair data,” *Statistical Medecine*, vol. 22, no. 15, Aug. 1975.

[40] A. Dielmann and S. Renals, “Automatic meeting segmentation using dynamic Bayesian networks,” *Transactions on Multimedia*, vol. 9(1), pp. 25–36, 2007.