Recognizing Visual Focus of Attention from Head Pose in Natural Meetings

Sileye O. Ba, and Jean-Marc Odobez, *Member, IEEE*
IDIAP Research Institute, Rue Marconi 19, CP 592, CH-1920 Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
Email <first name>.<last name>@idiap.ch
URL www.idiap.ch

***Abstract*— We address the problem of recognizing the visual focus of attention (VFOA) of meeting participants based on their head pose. To this end, the head pose observations are modeled using a Gaussian Mixture Model (GMM) or a Hidden Markov Model (HMM) whose hidden states corresponds to the VFOA. The novelties of this work are threefold. First, contrary to previous studies on the topic, in our set-up, the potential VFOA of a person is not restricted to other participants only. It includes environmental targets as well (a table and a projection screen), which increases the complexity of the task, with more VFOA targets spread in the pan as well as tilt gaze space. Second, we propose a geometric model to set the GMM or HMM parameters by exploiting results from cognitive science on saccadic eye motion, which allows the prediction of the head pose given a gaze target. Third, an unsupervised parameter adaptation step not using any labeled data is proposed which accounts for the specific gazing behaviour of each participant. Using a publicly available corpus of 8 meetings featuring 4 persons, we analyze the above methods by evaluating, through objective performance measures, the recognition of the VFOA from head pose information obtained either using a magnetic sensor device or a vision based tracking system. The results clearly show that in such complex but realistic situations, the VFOA recognition performance is highly dependent on how well the visual targets are separated for a given meeting participant. In addition, the results show that the use of a geometric model with unsupervised adaptation achieves better results than the use of training data to set the HMM parameters.**

I. Introduction

Acknowledgments: This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the European union 6th FWP IST Integrated Project AMIDA (Augmented Multi-Party Interaction with Distance Access, FP6-0033812). This research was also funded by the U.S. Government VACE program. The authors also thank Dr. Daniel Gatica-Perez, Dr. Hayley Hung from IDIAP Research Institute for their helpful comments.

UNDERSTANDING the behavior or need of humans is a central issue in devising next-generation human computing systems that can emulate more human-like functions. At the heart of this issue lies, amongst others, the difficulty of measuring human behaviors in an accurate way, i.e. the challenge of developing algorithms that can reliably extract subtle human characteristics -e.g. body gestures, facial expressions, emotion- that allow a fine analysis of their behavior. One such characteristic of interest is the *gaze*, which indicates where and what a person is looking at. Or, in other words, the gaze indicates what the *visual focus of attention (VFOA)* of the person is. In the context of Human Computer Interface (HCI) applications, the development of gaze tracking systems has been the topic of many studies. Less research has been conducted for estimating and analyzing a person's gaze and VFOA in more open spaces, despite the fact that in many contexts, identifying the VFOA of a person conveys a wealth of information about that person: what is he interested in, what is he doing, how does he explore a new environment or react to different visual stimuli. Thus, tracking the VFOA of people could have important applications in the development of ambient intelligent systems.

For instance, in a public space, it could be useful to measure the degree of attraction of a given focus target such as advertisements or shop displays, as presented in [1]. In the context of meetings, tracking the gaze can be used in meeting digital assistants to analyze the social dynamics of non-verbal communication in the group and make the participants aware of the group dynamics. It has been shown that such a social awareness mechanism can affect people behavior and improve group cohesiveness [2]. The VFOA and group dynamics can also be used in remote meeting applications. It can provide to a remote participant a better understanding of what is happening in the environment where other meeting participants are co-located, thus increasing participant's satisfaction level. Needless to say, gaze plays an important role in

face-to-face conversations and more generally in group interaction, as it has been shown in a large body of social psychology studies [3]. Human interaction can be categorized as verbal (speech) or non-verbal (e.g. facial expressions). While the usage of the former is tightly connected to the explicit rules of language (grammar, dialog acts), the usage of non-verbal cues is usually more implicit. However, this does not prevent it from following rules and exhibiting specific patterns in conversations. For instance, in a meeting context, a person raising a hand usually means that he is requesting the floor. A listener's head nod or shake can be interpreted as agreement or disagreement [4]. Besides hand and head gestures, the VFOA is another important non-verbal communication cue with functions such as establishing relationships (through mutual gaze), regulating the course of interaction, expressing intimacy, and exercising social control [5], [6].

A speaker's gaze often correlates with his addressees, i.e. the intended recipients of the speech [7]. Also, for a listener, monitoring his own gaze in concordance with the speaker's gaze is a way to find appropriate time windows for speaker turn requests [8], [9]. Thus, recognizing the VFOA patterns of a group of people can reveal important knowledge about the participants' role and status [10], [6]. Following these studies in social psychology, computer vision researchers are showing more interest in the study of automatic gaze and VFOA recognition systems [1], [11], [12], as illustrated by some of the research tasks defined in several recent evaluation workshops [13], [14]. Since meetings are places where the multi-modal nature of human communication and interaction best occur, they are well suited to conduct such research studies.

In this context, the goal of this paper is to analyze the correspondence between the head pose of people and their gaze in more general meeting scenarios than those previously considered [11], [12] (see Fig. 1, and Fig. 9 for some results). In meeting rooms, where high resolution close-up views of the eyes typically required by HCI gaze estimation systems are not available in practice, it has been shown that head orientation can be reasonably utilized as an approximation of the gaze when VFOA targets are the other meeting participants [11]. In this paper, we investigate the estimation of VFOA from head pose in complex meeting situations. Firstly, unlike previous work ([11], [12]), the scenario we consider involves people looking at slides or writing on a sheet of paper on the table. As a consequence, people have more potential VFOA targets in our set-up (6 instead of 3 in the cited work), leading to more possible ambiguities between VFOA. Secondly, due to the physical

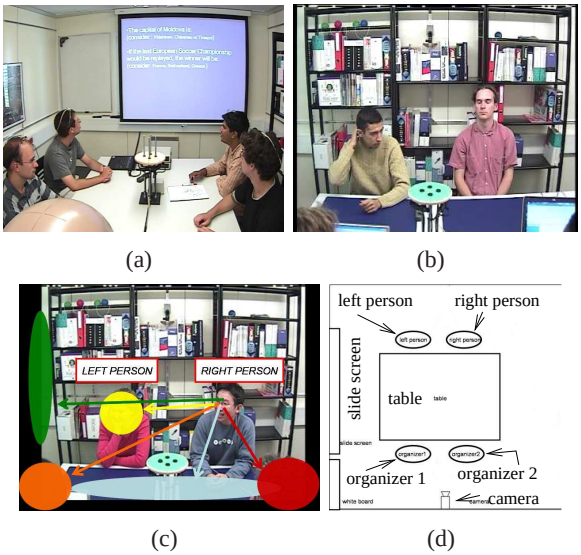


Fig. 1. Recognizing the VFOA of people. (a) the meeting room (b) a sample image of the dataset (c) the potential VFOA targets for the right person (d) the geometric configuration of the room.

placement of the VFOA targets, the identification of the VFOA can only be done using the complete head pose representation (pan and tilt), instead of just the head pan, as done previously. Thus, our work addresses general and challenging meeting room situations in which people do not just focus their attention on other people, but also on other room targets.

To recognize the VFOA of people from their head pose, we investigated two generative models. The first one is a Gaussian mixture model (GMM) that handle each frame separately. The second model is its natural extension to the temporal domain, namely a hidden Markov model (HMM), which segments pose observation sequences into VFOA temporal segments. In both cases, for each VFOA target, the head pose observations associated with each visual target are represented by Gaussian distributions. Alternative approaches were considered to set the model parameters. In one approach (referred to as the learning or training data approach), these were set using training data from other meetings. However, as collecting training data can be tedious, we used the results of studies on saccadic eye motion modeling [15], [16] and propose a novel approach (referred to as cognitive or geometric) that models the head pose of a person given his upper body pose and his effective gaze target. In this way, no training data is required to learn parameters, but some knowledge of the 3D room geometry is necessary. In addition, to account for the fact that people have their own head pose preferences for looking at the same given target, we adopted an unsupervised Maximum A Posteriori (MAP) scheme to adapt the parameters obtained from either the training data or the geometric approaches to unlabeled head pose

data of individual people in meetings.
To evaluate the different aspects of the VFOA modeling, we have conducted comparative and thorough experiments on a large and publicly available database. This database comprises 8 meetings of 10-minute average length for which both the head pose ground-truth and VFOA label ground truth are known. Therefore, we were able to differentiate between the two main error sources in VFOA recognition: (1) the use of head pose as a proxy for gaze, and (2) errors in the estimation of the head pose (e.g. using our vision-based head pose tracker [17]).
In summary, the contributions of this paper are the following:

- the development of a public database and a framework to evaluate the recognition of the VFOA solely from head pose;
- a novel geometric model to derive a person's head pose given his gaze target, which alleviates the need for training data;
- the use of an unsupervised MAP framework to adapt the VFOA model parameters to individual people;
- a thorough experimental study and analysis of the influence of several key aspects on the recognition performance (e.g. participant position, ground truth vs estimated head pose, correlation with tracking errors).

The remainder of this paper is organized as follows. Section II discusses the related work. Section III describes the task and the database that is used to evaluate the models we propose. Section IV provides an overview of our approach. Section V describes our algorithm for joint head tracking and pose estimation, along with its evaluation. Section VI describes the considered models for recognizing the VFOA from head pose. Section VII gives the unsupervised MAP framework used to adapt our VFOA model to unseen data. Section VIII describes our evaluation setup. We give experimental results in Section IX, and conclusions in Section X.

II. Related Work

We investigate the VFOA recognition from head pose in the context of meetings. Thus, we will analyze the related work along the following lines: gaze and VFOA tracking technologies, head pose estimation from vision sensors, and recognition of the VFOA from head pose.

The VFOA of a person is defined by his eye gaze, that is, the direction in which the eyes are pointing in the space. Many progresses in the design of gaze tracking technologies have been achieved. A review of such systems is presented in [18]. Gaze trackers are predominantly developed for HCI applications, where they are used for two main purposes: as an interactive tool, where the eyes are used as an input modality; or as a diagnostic tool, to provide evidence of a user's attention, such as in applications studying the visual exploration of images by people [19]. For this reason, these systems, while being accurate, are not appropriate for analyzing the VFOA of people in open spaces: they can be intrusive (user needs to wear special glasses) and require specific equipment (infrared light sources are often used to ease signal processing). More importantly, they are very constraining, as the head motion is limited to small position and angular variations (no more than 25cm and 20° [18]). In worst cases, chin rests or bite bars are required, but even eye-appearance vision-based gaze tracking systems restrict the mobility of the subject since their need of high resolution close-up eye images requires cameras with very narrow field-of-views. To alleviate this constraint, some papers [18], [20] propose using head pose tracking to localize eye corners and drive the acquisition of high resolution eye images using a pan-tilt-zoom (PTZ) camera. These systems, however, require very good calibration, and are still designed for near frontal head poses [20].

In spaces such as offices or meeting rooms, where the motion and head orientation of people are unconstrained, high resolution images of people's eye are not available. An alternative is to use the head pose as a surrogate for gaze, as proposed in [21]. Broadly speaking, head pose tracking algorithms can be divided into two groups: model based and appearance based approaches. In model based approaches, a set of facial features such as the eyes, the nose and the mouth are tracked. Then, knowing the relative positions of theses features, the head pose can be inferred using anthropometric information [22], [23]. The major drawback is that robust facial feature tracking is difficult unless high enough resolution images are used. Even in this later case, feature tracking can be problematic, especially when the pose reaches profile views. By modelling appearance of the whole head more robustness for low resolution images is obtained: [11] used neural network to model head appearance, [24], [25] developed the active appearance models based on principal component analysis, and [26], [27] used multidimensional Gaussian distribution to represent the head appearance likelihood.
From another perspective, head pose tracking algorithms differentiate themselves according to whether or not the tracking and the pose estimation are conducted jointly. Often, a generic tracker is used to locate the head, and then features extracted at this location are used to estimate the pose [11], [25], [26], [27]. Decoupling the tracking and the pose estimation results in a computational cost reduction. However, since head pose

estimation is very sensitive to head localization [27], head pose results are highly dependent on the tracking accuracy. To address this issue, [17], [24], [28] perform the head tracking and the pose estimation jointly.

In contrast to head tracking algorithms, few works have investigated the recognition of the VFOA directly from head pose. Pionneering work from [11] used a GMM model, the parameters of which were learned on the test data after initialization from the output of a K-means clustering of the pose values. This approach was possible due to constraints on the physical set-up (four people evenly spaced around a round table) and by limiting the allowed VFOA targets to the other participants. These constraints allowed them to rely only on the pan angle to represent the head pose, and limited the possibility of ambiguities in the head pose. In addition, [11] showed that using other participants' speaking status could further increase the VFOA recognition. More recently, [12] used a dynamic Bayesian network to jointly recognize the VFOA of people, as well as different conversational events in a 4-person conversation, using on head pan and speaking status observations. Finally, [29] exploited the head pose extracted from an overhead camera tracking retro-reflective markers mounted on headsets to look for occurrences of shared mutual visual attention. This information was then exploited to derive the social geometry of co-workers within an office, and infer their availability status for communication.

III. Database and task

In this section, we describe the VFOA recognition task, and the data that is used to evaluate both our pose estimation and VFOA recognition algorithms.

A. The Task and VFOA Set

Our goal is to evaluate how well we can infer the VFOA state of a person using head pose in common meeting situations. Let us first note that while the VFOA is given by the eye gaze, psycho-visual studies have shown that people use other cues -e.g. head and body posture, speaking status- to recognize the VFOA state of another person [5]. Thus, one general objective of the current work is to see how well one can recognize the VFOA of people from these other cues in the absence of direct gazing measurements, a situation likely to occur in many applications of interest. An important issue is: what should be the definition of a person's VFOA state? At first thought, one can consider that each different gaze direction could correspond to a potential VFOA. However, studies on the VFOA in natural conditions [30] have shown that humans tend to look at targets, whether humans or objects, that are either relevant to the task they are solving or of immediate interest to them. Additionally, one interprets another person's gaze not as continuous 3D spatial locations, but as a gaze towards objects that have been identified as potential targets. This process is often called the shared-attentional mechanism [31], [5], and suggests that in general VFOA states correspond to a finite set of targets of interests.

Thus, in our meeting context the set of potential VFOA targets, denoted $\mathcal{F}$, has been defined as: the other participants, the slide-screen, and the table. When none of the previous applies (the person is distracted by some noise or visual stimuli and looks at another target) we use an additional label called (*unfocused*). As a result, for 'person left' in Fig. 1(c), we have: $\mathcal{F} = \{PR, O2, O1, SS, TB, U\}$ where PR stands for person right, $O1$ and $O2$ for organizer 1 and 2, SS for slide screen, TB for table and U for unfocused. For the person right, $\mathcal{F} = \{PL, O2, O1, SS, TB, U\}$, where PL stands for person left. Note that in practice, the *unfocused* label only represents a small percentage of our data (2%), while the other VFOA target represent 55%, 26% and 17% for the other participants, the slide screen, and the table, respectively.

B. The Database

Our experiments rely on the IDIAP Head Pose Database (IHPD) [1]. The video database was collected along with a head pose ground truth and each participant's discrete VFOA ground truth, as explained below.

Content description: the database is comprised of 8 meetings involving 4 people each, recorded in a meeting room (cf Fig. 1(a)). The meeting durations ranged from 7 to 14 minutes, which was long enough to realistically represent a general meeting scenario. In shorter recordings (less than 2-3 minutes), we found that participants tend to be more active resulting in moving their head more to focus on other people/objects. In our meetings or in longer situations, the attention of participants sometimes drops and people are less focused on the other meeting participants. Note, however, that the small group size encourages engagement of participants in the meeting, in contrast to meeting with larger groups. Meeting participants were instructed to write down their name on a sheet of paper, then discuss statements displayed on the projection screen. There were no restrictions placed on head motion or head pose.

Head pose annotation: the head poses of two persons were continuously annotated (person left and right in Fig. 1(c)) using a magnetic field sensor called flock

[1] Available at *http://www.idiap.ch/HeadPoseDatabase/* (IHPD)

of birds (FOB) rigidly attached to the heads. It resulted in a video database of 16 different people. The coordinate frame of the magnetic sensors was calibrated with respect to the camera frame, allowing us to generate the head pose ground truth (denoted GT in the rest of the paper) with respect to the camera. The head pose is defined by three Euler angles (α, β, γ) that parametrize the decomposition of the rotation matrix of the head configuration with respect to the camera frame. To report our results, we have selected as Euler decomposition the one whose rotation axes are rigidly attached to the head (see Fig. 3(b)): α denotes the pan angle, a left/right head rotation; β denotes the tilt, an up/down head rotation; and finally, γ, the roll, represents a left/right "head on shoulder" head rotation. Because of our meeting scenario, people often have negative pan values corresponding to looking at the projection screen. Recorded pan values range from -70 to 60 degree. Tilt values range from -60 (when people are writing) to 15 degrees, and roll from -30 to 30 degrees.

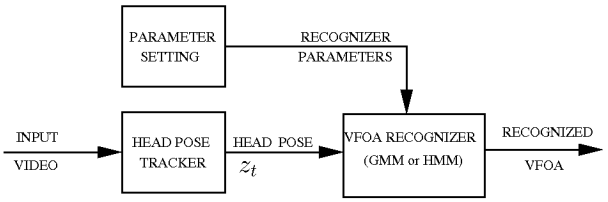
VFOA annotation: using the predefined discrete set of VFOA targets $\mathcal{F}$, the VFOA of each person (PL and PR) was manually annotated on the basis of their gaze direction by a single annotator using a multimedia interface. The annotator had access to all data streams, including the central camera view (Fig. 1(a)). Specific annotation guidance was defined in [32].

IV. Overview of the Proposed VFOA Recognition Methods

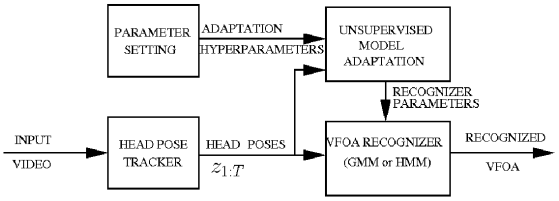
In this section, schematic representations of the components of the VFOA recognition methods proposed in this paper are provided in Fig. 2 to give a global view of the methods.

Fig. 2(a) presents the VFOA recognition method when no adaptation is used. The frames of an input video are sent to the head pose tracking algorithm (described in Section V) which outputs people's head poses. These poses are then processed by the VFOA recognizer module (described in Section VI-A), whose parameters are provided by a parameter setting module (Section VI-B).

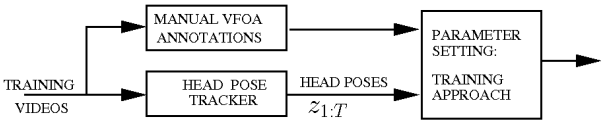
In Fig. 2(b), the use of unsupervised adaptation for VFOA recognition is sketched (described in Section VII). In this case, we employ a batch processing: the whole input video is processed by the head tracker to obtain the head poses of people over the entire meeting. Then, the adaptation module estimates in an unsupervised fashion (without using any annotated data) the VFOA recognizer parameters by fitting the recognizer model to the head poses while taking into account priors on these parameters. Some of the parameters of these priors are provided by the parameter setting



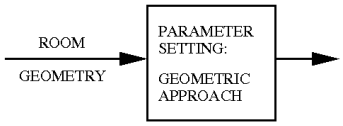
(a) VFOA recognition without adaptation.



(b) VFOA recognition with adaptation.



(c) VFOA parameter setting: training approach.



(d) VFOA parameter setting: geometric approach.

Fig. 2. Recognition approaches and modules overview.

module. Finally, the VFOA recognition module applies the parameters obtained through unsupervised adaptation to head poses to output the recognized VFOA.

Fig. 2(c) and 2(d) describes the two options that are used to define the parameter setting module involved in Fig. 2(a) and 2(b). The first option relies on training data: training videos are sent to the head pose tracking module whose output is used in conjunction with manual annotations of people's VFOA to learn the VFOA recognition parameters relating head pose to VFOA targets. The second option relies on a cognitive model of how people gaze at targets, and uses the location of people and object in the room as input. Section VI-B describes how the parameters are set in the two options and used when no adaptation is performed, while Section VII-C describes how the same parameters are used to define the hyper-parameters of the adaptation module.

V. Head Pose Tracking

In this section, we first summarize the computer vision probabilistic head tracker that we employed to estimate the pose. Then, the pose estimates provided by the tracker are compared with the pose GT (cf Section III) and analyzed in detail, ultimately giving us better insight into the VFOA recognition results of Section IX.

A. Probabilistic Method for Head Pose Tracking

The Bayesian formulation of the tracking problem is well known. Denoting the hidden state representing the object configuration at time t by X_t and the observation extracted from the image by Y_t, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of the state X_t given the sequence of all the observations $Y_{1:t} = (Y_1, \ldots, Y_t)$ up to the current time. Given standard assumptions, Bayesian tracking amounts to solving the following recursive equation:

$$p(X_t|Y_{1:t}) \propto p(Y_t|X_t) \times \int_{X_{t-1}} p(X_t|X_{t-1})p(X_{t-1}|Y_{1:t-1})\mathrm{d}X_{t-1} \quad (1)$$

In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters (PF). The idea behind PF consists in representing the filtering distribution using a set of N_s weighted samples (particles) $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrives. Given the particle set of the previous time step, configurations of the current step are drawn from a proposal distribution $X_t \sim \sum_n w_{t-1}^n p(X|X_{t-1}^n)$. The weights are then computed as $w_t \propto p(Y_t|X_t)$.

Four elements are important in defining a PF: i) a state model defining the object we are interested in; ii) a dynamical model $p(X_t|X_{t-1})$ governing the temporal evolution of the state; iii) a likelihood model measuring the adequacy of data given the proposed configuration of the tracked object; and iv) a sampling mechanism which has to propose new configurations in high likelihood regions of the state space. These elements are described in the next paragraphs.

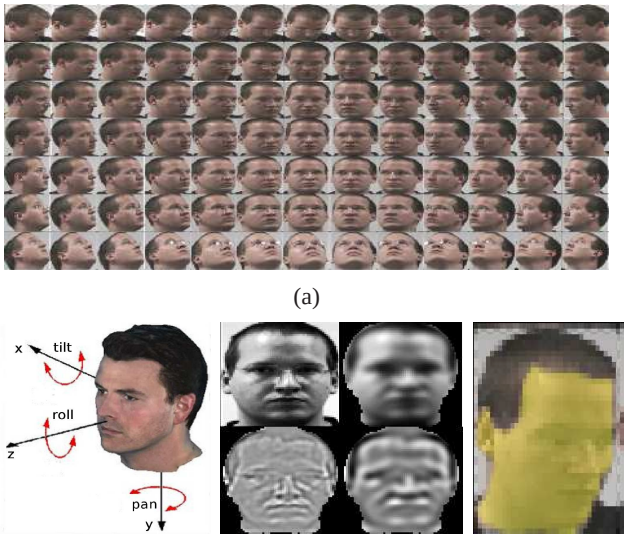
State Space: The state space contains both continuous and discrete variables. More precisely, the state is defined as $X = (S, \theta, l)$ where S represents the head location and size, and θ represents the in-plane head rotation. The variable l labels an element of the discretized set of possible out-of-plane head poses[2] (see Fig. 3a).

Dynamical Model: The dynamics $p(X_t|X_{1:t-1})$ governs the temporal evolution of the state, and is defined as

$$p(\theta_t|\theta_{t-1}, l_t)p(l_t|l_{t-1}, S_t)p(S_t|S_{t-1}, S_{t-2}). \quad (2)$$

The dynamics of the in-plane head rotation θ_t and discrete head pose l_t variables are learned using head pose GT training data. Head location and size dynamics are modeled as second order auto-regressive processes.

[2]Note that (θ, l) is another Euler decomposition (using different axis) of the head pose, which differs from the one described in Subsection III-B (cf Fig. 3a). Its main computational advantage is that one of the angles corresponds to the in-plane rotation. It is straightforward to transform from one decomposition to the other.



(a) (b) (c) (d)

Fig. 3. *(a) training head pose appearance range*. Pan and tilt angles range respectively from -90° to 90° and -60° to 60° by 15° steps. *(b) head pose Euler rotation angles.* Note that the z axis indicates the head pointing direction. *(c) and (d) tracking features*. Texture features from Gaussian and Gabor filters c) and skin color binary mask d).

Observation Model: The observation model $p(Y|X)$ measures the likelihood of the observation for a given state value. The observations $Y = (Y^{tex}, Y^{sk})$ are composed of texture and color observations (see Fig. 3 (c) and Fig. 3 (d)). Texture features are represented by the output of three filters (a Gaussian and two Gabor filters at different scales) applied at locations sampled from image patches extracted from the image and preprocessed by histogram equalization to reduce light variations effects. Color features are represented by a binary skin mask extracted using a temporally adapted skin color model. Assuming that, given the state value, texture and color observation are independent, the observation likelihood $p(Y|X = (S, \theta, l))$ is modeled as:

$$p_{tex}(Y^{tex}(S, \theta)|l)p_{sk}(Y^{sk}(S, \theta)|l) \quad (3)$$

where $p_{sk}(\cdot|l)$ and $p_{tex}(\cdot|l)$ are pose dependent models. For a given hypothesized configuration X, the parameters (S, θ) define an image patch on which the features are computed, while the exemplar index l selects the appropriate appearance model.

Sampling Method: In this work, we use Rao-Blackwellization, a process in which we apply the standard PF algorithm to the tracking variables S and θ while applying an exact filtering step to the exemplar variable l. The method theoretically results in a reduced estimation variance, as well as a reduction of the number of samples.

For more details about the models and algorithm, the reader is referred to [17]. Finally, in terms of complexity, the head tracker (in matlab) can process

TABLE I

PAN/TILT/ROLL ERROR STATISTICS FOR PERSON LEFT/RIGHT.

condition	right persons		left persons	
stat	mean	med	mean	med
pan	11.4	8.9	14.9	11.3
tilt	19.8	19.4	18.6	17.1
roll	14	13.2	10.3	8.7

TABLE II

PAN/TILT/ROLL ERROR STATISTICS FOR DIFFERENT CONFIGURATIONS OF THE TRUE HEAD POSE.

condition	pan near frontal ($\|\alpha\| < 45°$)		pan near profile ($\|\alpha\| > 45°$)		tilt near frontal ($\|\beta\| < 30°$)		tilt far from frontal ($\|\beta\| > 30°$)	
stat	mean	med	mean	med	mean	med	mean	med
pan	11.6	9.5	16.9	14.7	12.7	10	18.6	15.9
tilt	19.7	18.9	17.5	17.5	19	18.8	22.1	21.4
roll	10.1	8.8	18.3	18.1	11.7	10.8	18.1	16.8

around 1 frame per second.

B. Head Pose Tracking Evaluation

Protocol: We used a two-fold evaluation protocol, where for each fold, we used half (8 people) of our IHPD database (see Sec.III-B) as the training set to learn the pose dynamic model and the remaining half as the test set. Initialization was done automatically using a simple background subtraction technique, modeling the distribution of a pixel background color with one Gaussian, and the assumption that background image is available and that there was one face on the left and right half of the image (cf Fig. 1(c)).

It is important to note that the pose dependent appearance models were not learned using the same people or head images gathered in the same meeting room environment. We used the Prima-Pointing database [33], which contains 15 individuals recorded over 93 different poses (see Fig. 3(a)). However, when learning appearance models over whole head patches, as done in [17], we experienced tracking failures with 2 out of the 16 people of the IHPD database (see Section III) which had hair appearances not represented in the Prima-Pointing dataset (e.g. one of those two people was bald). As a remedy, we trained the appearance models on patches centered around the visible part *of the face, not the head*. With this modification, no failure was observed, but performance was slightly worse overall than those reported in [17].

Performance measures: three error measures are used. They are the average errors in pan, tilt and roll angles, i.e. the average of the absolute difference between the pan, tilt and roll of the ground truth (GT) and the tracker estimation. We also report the error median value, which should be less affected by very large errors due to erroneous tracking.

Results: The statistics of the errors are shown in Table I. Overall, given the small head size, and the fact that the appearance training set is composed of faces recorded in an external set up (different people, different viewing and illumination conditions), the results are quite good, with a majority of head pan errors smaller than 12° (see Fig. 4). However these results hide a large discrepancy between individuals: the average pan error ranges from 7° to 30°. It mainly depends on whether the tracked person's appearance is well represented by those of people in the training set. This was more the case for people seated on the right than on the left, as shown by Table I.

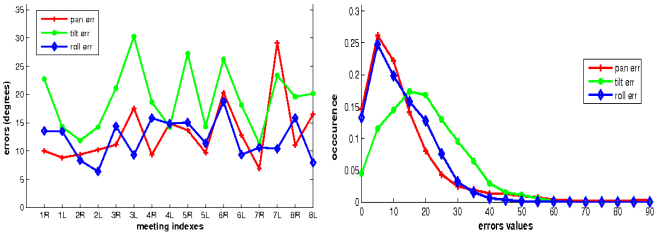


Fig. 4. *(a) and (b) pan, tilt and roll tracking errors* with a) average errors for each person (R for right and L for left person) and b) distribution of tracking errors over the whole dataset.

Table I also shows that overall the pan and roll tracking errors are smaller than the tilt errors. The main reason is that tilt estimation is more sensitive to the quality of the face localization than the pan, as pointed out by other researchers [27]. Indeed, even from a perceptive point of view, visually determining head tilt is more difficult than determining head pan or head roll.

Table II further details the errors depending on whether the true pose is near frontal or not. We can observe that, in the near frontal poses ($|\alpha| \leq 45°$ or $|\beta| \leq 30°$), the head pose tracking estimates are more accurate, in particular for the pan and roll value. This can be understood since for near profile poses, a variation in pan introduces much less appearance change than the same variation in a near frontal view. Similarly, for high tilt values, the face-image distortion introduced by perspective shortening affects the quality of the observations.

These results are comparable to those obtained by others in similar conditions. For instance, [26] achieved a pan estimation error of 16.9 degrees for poses near the frontal position, and 19.2 degrees for poses near profile ($|\alpha| > 45°$). In [11], a neural network is used to train a head pose classifier from data recorded directly in two meeting rooms. When using 15 people for training and 2 for testing, average errors of 5 degrees in pan and tilt are reported. However, when training the models in one room and testing on data from the other meeting room, the average errors rise to 10 degrees.

VI. Visual Focus of Attention Modeling

In this Section, we first describe the models used to recognize the VFOA from the head pose measurements, then the two alternatives we adopted to set the model parameters.

A. VFOA recognizer models

Modeling VFOA with a Gaussian Mixture Model: Let $s_t \in \mathcal{F}$ denote the VFOA state, and z_t the head pointing direction of a person at a given time instant t. The head pointing direction is defined by the head pan (α) and tilt (β) angles, i.e. $z_t = (\alpha_t, \beta_t)$, since the head roll ($\gamma$) has no effect on the head direction by definition (see Fig. 3(a)). Estimating the visual focus can be posed in a probabilistic framework as finding the VFOA state maximizing the a posteriori probability

$$\hat{s}_t = \arg\max_{s_t \in \mathcal{F}} p(s_t|z_t) \text{ with } p(s_t|z_t) \propto p(z_t|s_t)p(s_t) \tag{4}$$

For each VFOA $f_i \in \mathcal{F}$ which is not *unfocused*, $p(z_t|s_t = f_i)$, which expresses the likelihood of the pose observations for the VFOA state f_i is modeled as a Gaussian distribution $\mathcal{N}(z_t; \mu_i, \Sigma_i)$ with mean μ_i and full covariance matrix Σ_i. The unfocused state $p(z_t|s_t = unfocused) = u$ is modeled as a uniform distribution with $u = \frac{1}{180 \times 180}$, as the head pan and tilt angle can vary from -90° to 90°. In Eq. 4, $p(s_t = f_i) = \pi_i$ denotes the prior information we have on a VFOA target f_i. Thus, in this modeling, the total pose distribution is represented as a GMM (plus one uniform mixture), with the mixture index (i) denoting the focus target:

$$p(z_t|\lambda_G) = \sum_{i=1}^{K-1} \pi_i \mathcal{N}(z_t; \mu_i, \Sigma_i) + \pi_K u, \tag{5}$$

where $\lambda_G = \{\mu = (\mu_i)_{i=1:K-1}, \Sigma = (\Sigma_i)_{i=1:K-1}, \pi = (\pi_i)_{i=1:K}\}$ represents the parameter set of the GMM model. Fig. 12 illustrates how the pan-tilt space is split into different VFOA regions when applying the decision rule of Eq. 4 with the GMM modeling.

Modeling VFOA with a Hidden Markov Model: The GMM approach does not account for the temporal dependencies between VFOA events. A HMM is a natural extension to the GMM approach for modeling such temporal dependencies. Denoting the VFOA sequence by $s_{0:T}$ and the observation sequence by $z_{1:T}$, the joint posterior probability density function of states and observations can be written:

$$p(s_{0:T}, z_{1:T}) = p(s_0) \prod_{t=1}^{T} p(z_t|s_t)p(s_t|s_{t-1}). \tag{6}$$

In this equation, the emission probabilities $p(z_t|s_t = f_i)$ are modeled as in the previous case (i.e. Gaussian distributions for the regular focus targets, uniform distribution for the *unfocused* case). However, in the HMM modeling, the static prior distribution on VFOA targets is replaced by a discrete transition matrix $A = (a_{i,j})$, defined by $a_{i,j} = p(s_t = f_j|s_{t-1} = f_i)$, which models the probability of passing from the focus f_i to the focus f_j. Thus, the set of parameters of the HMM model is $\lambda_H = \{\mu, \Sigma, A = (a_{i,j})_{i,j=1:K}\}$. With this model, given the observation sequence, the VFOA recognition is performed by estimating the optimal sequence of focus targets which maximizes $p(s_{0:T}|z_{1:T})$. This optimization is efficiently conducted using the Viterbi algorithm [34][3].

B. VFOA Recognizer Parameter Setting

The parameters of our model can be set either using training data, or using a geometric model, as illustrated in Fig. 2(c) and 2(d), and explained in more details below.

Gaussian Parameter Setting using labeled Training Data: Since in many meeting settings, people are mostly static and seated at the same physical positions, we can set the model parameters using training data. Thus, given training data with VFOA annotations, and head pose measurements, we can readily estimate all the parameters of the GMM or HMM models. Parameters learned with this training approach will be denoted with a l superscript. Note that μ_i^l and Σ_i^l are learned by first computing the VFOA means and covariances per meeting and then averaging the results on the meetings belonging to the training set.

Gaussian Parameter Setting using a Geometric Model: The training approach to parameter learning is straightforward when annotated data is available. However, annotating the VFOA of people in video recording is tedious and time consuming, as training data needs to be gathered and annotated for each meeting setup. In the case of moving people, this is impossible. As an alternative, we propose a model that exploits the geometric and cognitive nature of the problem. The parameters set with this model will be denoted with a superscript g (e.g. μ_i^g).

We assume that we have a camera calibrated w.r.t. the room. Given a head location and a VFOA target location, it is possible to derive the Euler angles associated with the gaze direction. As gazing at a target is usually accomplished by rotating both the eyes ('eye-in-head' rotation) and the head in the same direction, the head is only partially oriented towards the gaze. In neurophysiology and

[3] In principle, such a decoding procedure is performed in batch. However, efficient online approximations are available.

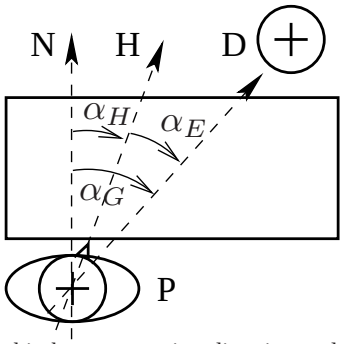


Fig. 5. Relationship between gazing direction and head orientation.

cognitive sciences, researchers studying the dynamics of the head/eye motions involved in saccadic gaze shifts have found that the relative contribution of the head and eyes towards a given gaze shift follows simple rules [15], [30]. While the experimental framework employed in these papers do not completely match the meeting room scenario, we have exploited these findings to propose a model for predicting a person's head pose given his gaze target.

The proposed geometric model is presented in Fig. 5. Given a person P whose reference head pose corresponds to looking straight ahead in the N direction, and given that he is gazing towards D, the head points in direction H according to:

$$\alpha_H = \kappa_\alpha \, \alpha_G \text{ if } |\alpha_G| > \xi_\alpha, \text{ and } 0 \text{ otherwise} \quad (7)$$

where α_G and α_H denotes the gaze and head pan angles respectively, both w.r.t. the reference direction N. The parameters of this model, κ_α and ξ_α, are constants independent of the gaze target, but usually depend on individuals [15]. While there is a consensus about the linearity aspect of the relation in Eq. 7, some researchers reported observing head movements for all gaze shift amplitudes (i.e. ξ_α=0), while others did not. In this paper, we will assume $\xi_\alpha = 0$. Besides, Eq. 7 is only valid if the contribution of the eyes to the gaze shift (given by $\alpha_E = \alpha_G - \alpha_H$) do not exceed a threshold, usually taken at $\sim 35^\circ$. Finally, in [15], it is shown that the tilt angle β follows a similar linearity rule (we denote by κ_β the corresponding proportionality factor). However, in this case, the contribution of the head to the gaze shift is usually lower than for the pan case. Typical values range from 0.2 to 0.5 for κ_β, and 0.5 to 0.8 for κ_α.

We assume we know the approximate positions of the people's heads, VFOA targets, and camera within the room[4]. The cognitive model can be used to predict the values of the mean μ of the Gaussian distributions associated with the VFOA targets. The reference direction N (Fig. 5) will be assumed to grossly correspond to the mean of all the gaze targets directions. For both person left and right, it corresponds to looking at O1 (cf Fig. 1(c)). The Gaussian covariances Σ were assumed to be diagonal, and were set by taking into account the physical target size, and the fact that VFOA targets corresponding to head poses in profile are associated with larger pan tracking errors. The specific values were: $\sigma_\alpha(O1, O2) = 12^\circ$, $\sigma_\alpha(Pr, PL, SS) = 15^\circ$, and $\sigma_\alpha(TB) = 17^\circ$ for the pan, and $\sigma_\beta(O1, O2, PR, PL, SS) = 12^\circ$, $\sigma_\beta(TB) = 15^\circ$ for the tilt.

Setting the Prior Distribution and Transition Matrix: When training data is available, one could learn these parameters. If the training meetings exhibit a specific structure, as is the case in our database, where the main and secondary organizers always occupy the same seats, the learned prior will have a beneficial effect on the recognition performances for similar unseen meetings. However, at the same time, this learned prior can considerably limit the generalization to other data sets, since by simply exchanging seats between participants, we obtain meeting sessions with different prior distributions. Thus, we investigated alternatives that avoided favoring any meeting structures. In the GMM case, this was done by considering a uniform distribution (denoted π^u) over the prior π. In the HMM case, transitions defining the probability of keeping the same focus were favored and transitions to other focuses were distributed uniformly according to: $a_{i,i} = \epsilon < 1$ (we used $\epsilon = 0.75$), and $a_{i,j} = \frac{1-\epsilon}{K-1}$ for $i \neq j$ where K is the number of VFOA targets. We denote as A^u the constructed transition matrix.

VII. VFOA Models Adaptation

The VFOA recognizers described in the previous section are generic and can be applied indifferently to any new person seated at the location corresponding to the defined model. In practice, however, we observed that people have personal ways of looking at targets. For example, some people use their eye-in-head rotation capabilities more and turn less their head towards the focused target than others (see Fig 6(a) and Fig 6(b)). In addition, our head pose tracking system is sensitive to the visual appearance of people, and can introduce a systematic bias in the estimated head pose for a given person. As a consequence, the parameters of the generic models might not be the best for a given person. As a remedy we propose to exploit the Maximum A Posteriori (MAP) estimation principle to adapt, in an unsupervised fashion, the generic VFOA models to the data of each

[4]The relation in Eq. 7 is valid in the person's head reference. The camera position is needed in order to transform the obtained pose values into head poses w.r.t. to the camera.

(a) (b)

Fig. 6. Examples of gaze behaviours. (a) and (b): in both images, the person on the right looks at the target $O1$. In (b), however, the head is used more rotated toward $O1$ than in (a).

new meeting, and thus produce models adapted to an individual's characteristics.

A. VFOA MAP Adaptation Principle

The MAP adaptation procedure we followed is a batch process, as explained in Section IV. Its principle is the following: Let $z = z_1, ..., z_T$ denotes the unlabeled sequence of head poses of one person, to which we want to adapt our model, and $\lambda \in \Lambda$ the parameter of the VFOA recognizer to be estimated from the head pose data. The MAP estimate $\hat{\lambda}$ is then defined as:

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} p(\lambda|z) = \arg\max_{\lambda \in \Lambda} p(z|\lambda)p(\lambda) \quad (8)$$

where $p(z|\lambda)$ is the data likelihood and $p(\lambda)$ is the prior on the parameters. The goal is thus to find the parameters that best fit the observed head pose distribution, while avoiding too large deviation from sensible values through the use of priors on the parameters. The choice of the prior distribution is crucial for the MAP estimation. In [35] it is shown that for GMMs and HMMs, by selecting the prior probability density function (pdf) on λ as the product of appropriate conjugate distributions of the data likelihood [5], the MAP estimation can be solved using the Expectation-Maximization (EM) algorithm, as detailed in the next two sub-sections.

B. VFOA GMM and HMM MAP Adaptation

GMM MAP Adaptation: In this case the data likelihood is $p(z|\lambda_G) = \prod_{t=1}^{T} p(z_t|\lambda_G)$, where $p(z_t|\lambda_G)$ is the mixture model given in Eq. 5, and λ_G are the parameters to be learnt.

For this model, it is possible to express the prior probability as a product of individual conjugate priors [35]. Accordingly, the conjugate prior of the multinomial mixture weights is the Dirichlet distribution $\mathcal{D}(\nu w_1, \ldots, \nu w_K)$ whose pdf is given by:

$$p^{\mathcal{D}}_{\nu w_1,\ldots,\nu w_K}(\pi_1, \ldots, \pi_K) \propto \prod_{k=1}^{K} \pi_i^{\nu w_i - 1} \quad (9)$$

Additionally, the conjugate prior for the Gaussian mean and the inverse covariance matrix of a given mixture is the Normal-Wishart distribution, $\mathcal{W}(\tau, m_i, d, V_i)$ ($i = 1, ..., K-1$), with pdf $p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1})$

$$\propto \Sigma_i^{-1}|^{\frac{d-p}{2}} \exp\left(-\frac{\tau}{2}(\mu_i - m_i)'\Sigma_i^{-1}(\mu_i - m_i)\right) \times \exp(-\frac{1}{2}tr(V_i\Sigma_i^{-1})),\ d > p \quad (10)$$

where tr denotes the trace operator, $(\mu_i - m_i)'$ denotes the transpose of $(\mu_i - m_i)$, and p denotes the observations' dimension. Thus the prior distribution on the set of all the parameters is defined as

$$p(\lambda_G) = p^{\mathcal{D}}_{\nu w_1,\ldots,\nu w_K}(\pi_1, \ldots, \pi_K) \prod_{i=1}^{K-1} p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1}). \quad (11)$$

The MAP estimate $\hat{\lambda_G}$ of the distribution $p(z|\lambda_G)p(\lambda_G)$ can thus be computed using the EM algorithm by recursively applying the following computations (see Fig. 7) [35]:

$$c_i = \sum_{t=1}^{T} c_{it}, \quad \text{with} \quad c_{it} = \frac{\hat{\pi}_i p(z_t|\hat{\mu}_i, \hat{\Sigma}_i)}{\sum_{j=1}^{K} \hat{\pi}_j p(z_t|\hat{\mu}_j, \hat{\Sigma}_j)} \quad (12)$$

$$S_i = \frac{1}{c_i}\sum_{t=1}^{T} c_{it}(z_t - \bar{z}_i)(z_t - \bar{z}_i)',\ \bar{z}_i = \frac{1}{c_i}\sum_{t=1}^{T} c_{it} z_t \quad (13)$$

where $\hat{\lambda_G} = (\hat{\pi}, (\hat{\mu}, \hat{\Sigma}))$ denotes the current parameter fit. Given these coefficients, the M step re-estimation formulas are given by:

$$\hat{\pi}_i = \frac{\nu w_i - 1 + c_i}{\nu - K + T}, \quad \hat{\mu}_i = \frac{\tau m_i + c_i \bar{z}_i}{\tau + c_i} \quad \text{and}$$
$$\hat{\Sigma}_i = \frac{V_i + c_i S_i + \frac{c_i \tau}{c_i + \tau}(m_i - \bar{z}_i)(m_i - \bar{z}_i)'}{d - p + c_i} \quad (14)$$

The setting of the hyper-parameters of the prior distribution $p(\lambda_G)$ in Eq. 11, which is discussed at the end of this section, is important as the adaptation is unsupervised, and thus only the prior prevents the adaptation process from deviating from meaningful VFOA distributions.

VFOA MAP HMM Adaptation: The VFOA HMM can also be adapted in an unsupervised way to new test data using the MAP framework [35]. The parameters to adapt in this case are the transition matrix and the parameters of the emission probabilities $\lambda_H = \{A, (\mu, \Sigma)\}$. Adaptation of the HMM parameters leads to a procedure similar to the GMM adaptation case. Indeed, the prior

[5] A prior distribution $g(\lambda)$ is the conjugate distribution of a likelihood function $f(z|\lambda)$ if the posterior $f(z|\lambda)g(\lambda)$ belongs to the same distribution family as g.

Input : adaptation parameters $(\nu, \{w_i\})$ for the Dirichlet, $(\tau, d, \{m_i, V_i\})$ for the Wishart prior.
Output : estimated parameter $\hat{\lambda_G}$ of the recognizer model

- Initialization of $\hat{\lambda_G}$: $\hat{\pi}_i = w_i,\ \hat{\mu}_i = m_i,\ \hat{\Sigma}_i = V_i/(d-p)$
- EM: repeat until convergence:
  1) Expectation: compute c_{it}, $\bar{z}_i$ and S_i (Eq. 12 and 13) using the current parameter set $\hat{\lambda_G}$
  2) Maximization: update parameter set $\hat{\lambda_G}$ using the re-estimation formulas (Equations 14)

Fig. 7. GMM MAP adaptation procedure.

on the Gaussian parameters follows the same Normal-Wishart density (Eq. 10), and the Dirichlet prior on the static VFOA prior π is replaced by a Dirichlet prior on each row $p(.|s = f_i) = a_{i,.}$ of the transition matrix. Accordingly, the $p(\lambda_H)$ is proportional to:

$$\prod_{i=1}^{K} p^{\mathcal{D}}_{\nu b_{i,1},\ldots,\nu b_{i,K}}(a_{i,1},\ldots,a_{i,K}) \prod_{i=1}^{K-1} p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1}) \quad (15)$$

Then the EM algorithm to compute the MAP estimate can be conducted in the following manner. For a sequence of observations, $z = (z_1, ..., z_T)$, the hidden states are now composed of a corresponding state sequence $s_1, .., s_T$, which allows us to compute the joint state-observation density (cf Eq. 6). Thus, in the E step, one needs to compute $\xi_{i,j,t} = p(s_{t-1} = f_i, s_t = f_j|z, \hat{\lambda_H})$ and $c_{i,t} = p(s_t = f_i|z, \hat{\lambda_H})$, which respectively denote the joint probability of being in the state f_i and f_j at time $t-1$ and t, and the probability of being in state f_i at time t, given the current model $\hat{\lambda_H}$ and the observed sequence z. These values can be obtained using the Baum-Welch forward-backward algorithm [34]. Given these values, the re-estimation formulas for the mean and covariance matrices are the same as those in Eq. 14 and as follows for the transition matrix parameters:

$$\hat{a}_{i,j} = \frac{\nu b_{i,j} - 1 + \sum_{t=1}^{T-1} \xi_{i,j,t}}{\nu - K + \sum_{j=1}^{K} \sum_{t=1}^{T-1} \xi_{i,j,t}}. \quad (16)$$

C. Choice of Prior Distribution Parameters

In this section we discuss the impact of the hyper-parameter settings on the MAP estimates, through the analysis of the re-estimation formula (Eq. 14). Before going into details, recall that T denotes the size of the data set available for adaptation, and K is the number of VFOA targets.

Parameter values for the Dirichlet distribution: The Dirichlet distribution is defined by two kinds of parameters: a scale factor ν and the prior values on the mixture weights w_i (with $\sum_i w_i = 1$). The scale factor ν controls the balance between the prior distribution on the mixture weights w and the data. If ν is small (resp. large) with respect to $T-K$, the adaptation is dominated by the data (resp. by the prior, i.e. almost no adaptation occurs). When $\nu = T-K$, the data and prior contribute equally to the adaptation process. In our experiments, the hyper-parameter ν will be selected through cross-validation among the values in $C^\nu = \{\nu_1 = T-K,\ \nu_2 = 2(T-K)\ , \nu_3 = 3(T-K)\}$. The prior weights w_i, on the other hand, are defined according to the prior knowledge we have on the distribution of VFOA targets. Since we do not want to enforce any knowledge about the VFOA targets distribution, the w_i can be set uniformly equal to $\frac{1}{K}$.

Parameter values for the Normal-Wishart distribution: This distribution defines the prior on the mean μ_i and covariance Σ_i of one Gaussian. The adaptation of the mean is essentially controlled by two parameters (see Eq. 14): the prior value for the mean, m_i, which will be set to the value computed using either the learning ($m_i = \mu_i^l$) or the geometric approach ($m_i = \mu_i^g$) and a scalar τ, which linearly controls the contribution of the prior m_i to the estimated mean. As the average value for c_i, is $\frac{T}{K}$, in the experiments, we will select τ though cross-validation among the values in $C^\tau = \{\tau_1 = \frac{T}{2K},\ \tau_2 = \frac{T}{K},\ \tau_3 = \frac{2T}{K},\ \tau_4 = \frac{5T}{K}\}$. Thus, with the first value τ_1, the mean adaptation is on average dominated by the data. With τ_2, the adaptation is balanced between the data and prior distrubution on the means, and with the two last values, adaptation is dominated by the priors on the means.

The prior on the covariance is more difficult to set. It is defined by the Wishart distribution parameters, namely the prior covariance matrix V_i and the number of degrees of freedom d. From Eq. 14, we see that the data covariance and the deviation of the data mean from the mean prior also influence the MAP covariance estimate. As a prior Wishart covariance, we will take $V_i = (d-p)\tilde{V}_i$, where $\tilde{V}_i$ is either Σ_i^l or Σ_i^g, the covariance of target f_i set either using training data or the geometric model (Subsection VI-B) respectively. The weighting $(d-p)$ is important, as it allows V_i to be of the same order of magnitude than the data variance $c_i S_i$. In the experiments, we will use $d = \frac{5T}{K}$, which restricts the adaptation from deviating far from the covariance priors.

VIII. Evaluation Set Up

The evaluation of the VFOA models was conducted using the IHPD database (Section III). Below, we describe our performance measures and give details about the experimental protocol.

A. Performance Measures

We propose two kinds of error measures for performance evaluation.

The Frame based Recognition Rate (FRR) which corresponds to the percentage of frames, or equivalently, the proportion of time, during which the VFOA has been correctly recognized. This rate, however, can be dominated by VFOA events of long duration (a VFOA event is defined as a temporal segment with the same VFOA label). Since we are also interested in the dynamics of the VFOA, which contains information related to interaction, we also need a measure reflecting how well these events, short or long, are recognized.

Event based precision/recall, and F-measure. Let us consider two sequences of VFOA events: the GT sequence G obtained from human annotation, and the recognized sequence R obtained through VFOA estimation. The GT sequence is defined as $G = (G_i = (l_i, I_i = [b_i, e_i]))_{i=1,...N_G}$ where N_G is the number of events in the ground truth G, $l_i \in \mathcal{F}$ is the ith VFOA event label, and b_i and e_i are the beginning and end time instants of the event G_i. The recognized sequence R is defined similarly. To compute the performance measures, the two sequences are first aligned using a string alignment procedure that takes into account the temporal extent of the events. More precisely, the matching distance between two events G_i and R_j is defined as:

$$d(G_i, R_j) = \begin{cases} 1 - F_I & \text{if } l_i = l_j \text{ and } I_i \cap I_j \neq \emptyset \\ 2 & \text{otherwise}, \end{cases} \tag{17}$$

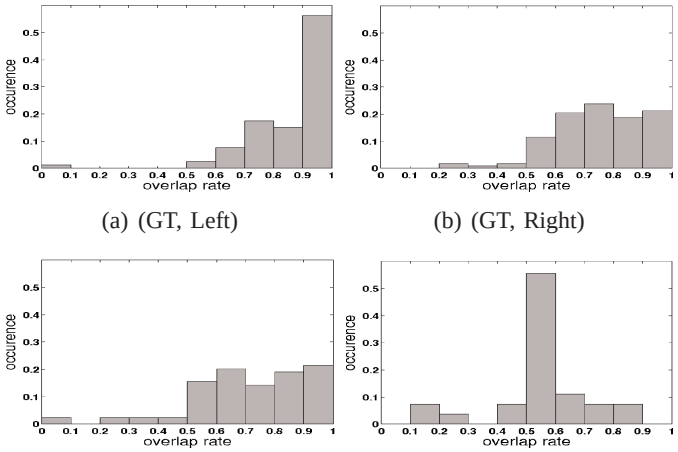
where $F_I = \frac{2\rho_I \pi_I}{\rho_I + \pi_I}$ with:

$$\rho_I = \frac{|I_i \cap I_j|}{|I_i|} \text{ and } \pi_I = \frac{|I_i \cap I_j|}{|I_j|} \tag{18}$$

where $|.|$ denotes the cardinality operator, and F_I measures the degree of overlap between two events. Then, given the alignment we can compute the recall ρ_E, the precision π_E, and the F-measure F_E for each person measuring the event recognition performance, defined as $F_E = \frac{2\rho_E \pi_E}{\rho_E + \pi_E}$ with

$$\rho_E = \frac{N_{matched}}{N_G} \text{ and } \pi_E = \frac{N_{matched}}{N_R}, \tag{19}$$

where $N_{matched}$ represents the number of events in the recognized sequence that match the same event in the GT after alignment. The recall measures the percentage of ground truth events that are correctly recognized while the precision measure the percentage of estimated events that are correct. Both precision and recall need to be high to characterize a good VFOA recognition performance.



(a) (GT, Left) (b) (GT, Right)

(c) (TR, Left) (d) (TR, Right)

Fig. 8. Distribution of overlap measures F_I between true and estimated matched events. The estimated events were obtained using the HMM approach. GT and TR respectively denote the use of GT head pose data and tracking estimates data. Left and Right denote person left and right respectively.

tags	description
gt	head pose measurements are from the magnetic field sensor
tr	head pose measurements are from the head tracking algorithm
gmm	the VFOA recognition model is a GMM
hmm	the VFOA recognition model is an HMM
ML	the same meeting is used fortraining and testing is used to
ge	Gaussian parameters are set using the geometric gaze approach
ad	VFOA model parameters were adapted

TABLE III

MODEL ACRONYMS (TAGS): ACRONYM COMBINATIONS DESCRIBE WHICH EXPERIMENTAL CONDITIONS ARE USED. FOR EXAMPLE, GT-HMM-GE INDICATES THAT THE HMM VFOA RECOGNIZER WITH PARAMETERS SET USING THE GEOMETRIC GAZE MODEL WERE APPLIED TO GROUND TRUTH POSE DATA.

The F-measure, defined as the harmonic mean of recall and precision, reflects this requirement. We report the average of the precision, recall and F-measure F_E of the 8 individuals over the whole database (and for each seat position). Note that according to Eq. 17, events are said to match whenever their common intersection is not empty (and labels match). One may think that the counted matches could be generated by spurious accidental matches due to a very small intersection. In practice, however, we observe that it is not the case: the vast majority of matched events have a significant degree of overlap F_I, as illustrated in Fig. 8, with 90% of the matches exhibiting an overlap higher than 50%, even using noisier tracking data.

B. Experimental Protocol

To study different modeling aspects, several experimental conditions have been defined. They are summarized in Table III along with the acronyms that identify them in the result tables. First, there are two alternatives

VFOA recognition without adaptation	
μ_i, Σ_i	Gaussian parameters - learned (μ_i^l, Σ_i^l) or given by geometric modeling (μ_i^g, Σ_i^g), cf Subsection VI-B.
π, A	GMM and HMM model priors - set to the values π^u, A^u, as described in Subsection VI-B.
VFOA recognition with adaptation	
μ_i, Σ_i, π, A	same description as above - set as the result of the adaptation process.
ν	scale factor of Dirichlet distribution - set through cross-validation.
$w_i, b_{i,j}$	Dirichlet prior values of π_i and $a_{i,j}$ - set to π_i^u and $a_{i,j}^u$.
τ	scale factor of Normal prior distribution on mean - set through cross-validation.
m_i	VFOA mean prior value of Normal prior distribution - set to either μ_i^l or μ_i^g.
d	scale factor of Wishart prior distribution on covariance matrix - set by hand (cf Sec. VII-C).
V_i	VFOA covariance matrices prior values in Wishart distribution - set to either $(d-2)\Sigma_i^l$ or $(d-2)\Sigma_i^g$.

TABLE IV

VFOA MODELING PARAMETERS: DESCRIPTION AND SETTING. THE GAZE FACTORS $\kappa_\alpha, \kappa_\beta$ WERE SET BY HAND.

regarding the head pose measurements: the ground truth *gt* case, where the data is obtained using the FOB magnetic sensor, and the *tr* case, which relies on the estimates obtained with the video tracking system. Secondly, there are the two VFOA recognizer models, *gmm* and *hmm*. Regarding the approach relying on training data for parameter setting, the default protocol is the leave-one-out approach: each meeting recording is in turn left aside for testing, while the data of the 7 other recordings are used for parameter learning, including hyper-parameter selection in the adaptation case (denoted *ad*). The maximum likelihood case *ML* is an exception, in which the training data for a given meeting recording is composed of the same single recording. The *ge* acronym denotes the case where the VFOA Gaussian means and covariances were set according to the geometric model instead of being learned from training data. Finally, the adaptation hyper-parameter pair (ν, τ) was selected (in the cartesian set $C^\nu \times C^\tau$) by cross-validation over the training data, using F_E as performance measure to maximize. A summary of all parameters involved in the modeling and the way they were set depending on whether there was adaptation or not is displayed in Table IV.

IX. EXPERIMENTAL RESULTS

This section provides results under the various experimental conditions. We first analyze the results obtained on the GT head pose data, and then compare them with those obtained using the tracking estimates instead. In both cases, we discuss the effectiveness of the modeling w.r.t. different issues: (i) relevance of head pose to model VFOA gaze targets, (ii) predictability of VFOA head pose parameters, (iii) impact of the person's position in the room. Then, we comment on the results of the adaptation scheme. Note that although these first sets of results are only shown with the parameter setting using training data, the conclusions that are made are also valid for the geometric parameter setting. In Section IX-D, we compare in details the results obtained with the geometric parameter setting and those obtained with the training parameter setting. In all cases, results are given separately for the left and right persons (see Fig. 1). Some result illustrations are provided in Fig.9.

data	ground truth (gt)			tracking estimates (tr)		
modeling	ML	gmm	hmm	ML	gmm	hmm
FRR	79.7	72.3	72.3	57.4	47.3	47.4
recall	79.6	72.6	65.5	66.4	49.1	38.4
precision	51.2	55.1	66.7	28.9	30	59.3
F-measure F_E	62	62.4	65.8	38.2	34.8	45.2

TABLE V

VFOA RECOGNITION RESULTS FOR PERSON LEFT UNDER DIFFERENT EXPERIMENTAL CONDITIONS (SEE TABLE III).

data	ground truth (gt)			tracking estimates (tr)		
modeling	ML	gmm	hmm	ML	gmm	hmm
FRR	68.9	56.8	57.3	43.6	38.1	38
recall	72.9	66.6	58.4	65.6	55.9	37.3
precision	47.4	49.9	63.5	24.1	26.8	55.1
F-measure F_E	56.9	54.4	59.5	34.8	35.6	43.8

TABLE VI

VFOA RECOGNITION RESULTS FOR PERSON RIGHT UNDER DIFFERENT EXPERIMENTAL CONDITIONS (SEE TABLE III).

A. Results on GT head pose data

VFOA and head pose correlation: Table V and VI display the VFOA recognition results for person left and right respectively. The first column of these two tables gives the results of the ML estimation (see Tab. III) with a GMM. These results show, in an optimistic case, the performances our model can achieve, and illustrate the correlation between a person's head poses and his VFOA. As can be seen, this correlation is quite high for

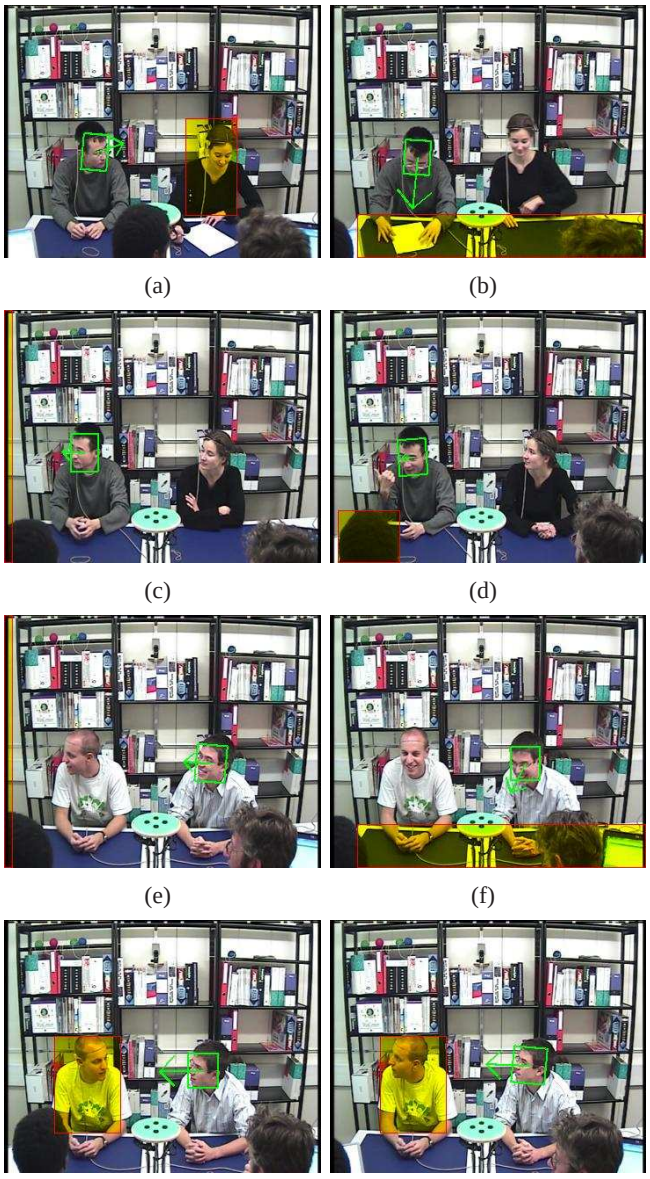


Fig. 9. Example of results and focus ambiguity. In green, tracking result and head pointing direction. In yellow, recognized focus (hmm-ad condition). Images (g) and (h): despite the high visual similarity of the head pose, the true focus differ (in (g): PL; in h: SS). Resolving such cases can only be done by using context (speaking status, other's people gaze, slide activity etc).

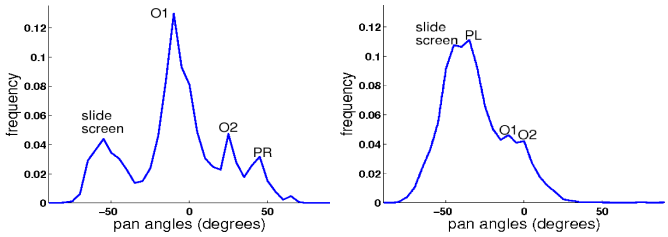
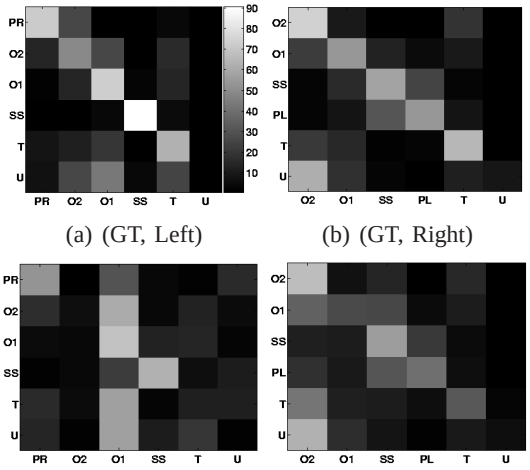


Fig. 10. Empirical distribution of the GT head pose pan angle computed over the database for PL (left image) and PR. For PL, the people and slide screen VFOA targets can still be identified through the pan modes. For PR, the degree of overlap is quite significant.

PL (almost 80% FRR), showing the good concordance between head pose and VFOA. This correlation, however, drops to near 69% for PR. This can be explained by the fact that for the person on the right (PR), there is a strong ambiguity between looking at PL or SS, as illustrated by the empirical distributions of the pan angle in Fig. 10. Indeed, the range of pan values within which the three other meeting participants and the slide screen VFOA targets lies is half the pan range of the person sitting to the left (PL). The average angular distance between these targets is around 20° for PR, a distance which can easily be covered using only eye movements rather than rotating the head. The values of the confusion matrices, displayed in Fig. 11, corroborate this analysis. The analysis of Tables V and VI shows that this discrepancy between the results for PL and PR holds for all experimental conditions and algorithms, with a performance decrease from PL to PR of approximately 10-13% and 6%, for the FRR and event F-measure respectively.

VFOA Prediction: In the ML condition, very good results were achieved but they were biased because the test data was used to set the Gaussian parameters. On the contrary, the GMM and HMM results in Table V and VI, for which the VFOA parameters were learned from other persons' data, highlights the generalization property of the modeling. We can observe that the GMM and HMM methods produce results close to the ML case. For both PL and PR, the GMM approach achieves better frame recognition and event recall performance while the HMM is giving better event precision and F_E results. This can be explained since the HMM approach is effectively denoising the event sequence. As a result some events are missed (lower recall) but the precision increases due to the elimination of short spurious detections.

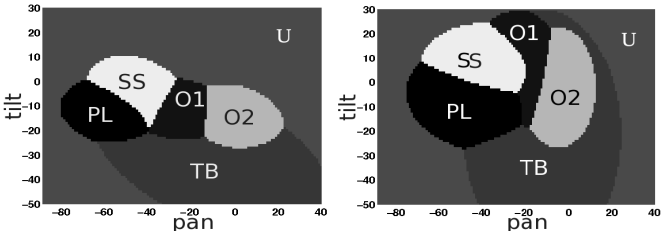
VFOA Confusions: Figure 11(a) and 11(b) display as images the confusion matrices for PL and PR obtained with the VFOA FRR performance measure and a HMM. They clearly exhibit confusion between VFOA targets which are proximate in the head pose space. For instance, for PL, $O2$ is sometimes confused with PR or $O1$. For PR, the main source of confusion is between PL and SS, as already mentioned. In addition, the table, TB, can be confused with $O1$ and $O2$, as can be expected since these targets share more or less the same pan values with TB. Thus, most of the confusion can be explained by the geometry of the room and the fact that people can modify their gaze without adjusting their head pose, and therefore do not always need to turn their heads to focus on a specific VFOA target.



(a) (GT, Left) (b) (GT, Right)

(c) (TR, Left) (d) (TR, Right)

Fig. 11. Frame-based recognition confusion matrices obtained with the HMM modeling (gt-hmm and tr-hmm conditions). VFOA targets 1 to 4 have been ordered according to their pan proximity: PR: person right - PL: person left - O1 and O2: organizer 1 and 2 - SS: slide screen - TB: table - U: unfocused. Columns represent the recognized VFOA.



(a) (b)

Fig. 12. Pan-tilt space VFOA decision maps for person right built from all meetings, in the GMM case (cf Eq. 4), using GT (a) or tracking head pose data (b). The areas corresponding to the VFOA targets are specified by their acronyms (PL, SS, $O1$, $O2$, TB, U).

B. Results on Head Pose Estimates data

Table V and VI provide the results obtained using the head pose tracking estimates, under the same experimental conditions as those used for the GT head pose data. As can be seen, substantial performance degradation is observed. In the ML case, the decrease in FRR and F-measure ranges from 22% to 26% for both PL and PR. These degradations are mainly due to small pose estimation errors and also, sometimes, large errors due to short periods when the tracker locks on a sub-part of the face. Fig. 12 illustrates the effect of pose estimation errors on the VFOA distributions. Shape changes in the VFOA decision maps when moving from GT pose data to pose estimates convey the increase of pose variance measured for each VFOA target. The increase is moderate for the pan angle, but quite important for the tilt angle.

A more detailed analysis of Table V and VI shows that the performance decreases (from GT to tracking data) in the GMM condition follows the ML case, while the deterioration in the HMM case is smaller, in particular for F_E. This demonstrates that, in contrast with what was observed with the clean GT pose data, in presence of noisy data, the HMM smoothing effect is quite beneficial. Also, the HMM performance decrease is smaller for PR (19% and 15% for respectively FRR and F_E) than for PL (25% and 20%). This can be due to the better tracking performance -in particular regarding the pan angle- achieved on people seated at the position PR (as reported in Table I). Fig. 13 presents the plot of the VFOA FRR versus the pan angle tracking error for each meeting participant, when using GT head pose data (i.e. with no tracking error) or pose estimates. It shows that for PL, there is a strong correlation between tracking errors and VFOA performances, which can be due to the fact that higher tracking errors directly generate larger overlaps between the VFOA class-conditional pose distributions (cf Fig. 10, left). For PR, this correlation is weaker, as the same good tracking performance results in very different VFOA recognition results. In this case, the increase of ambiguities between several VFOA targets (e.g. SS and PL) may play a larger role.

Finally, Fig. 11(c) and Fig. 11(d) display the confusion matrices when using the HMM and the head pose estimates. In this case, the confusion matrices are very similar to the case using GT. However more confusion is observed due to the tracking errors and the uncertainties in the tilt estimation (see Fig 13).

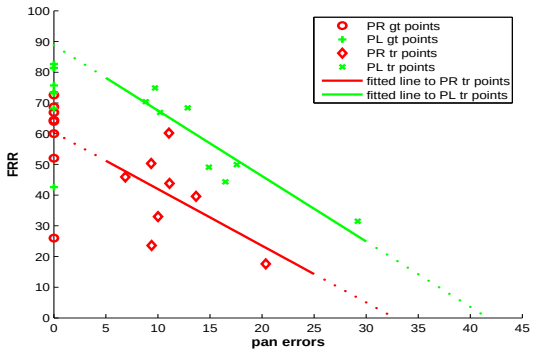


Fig. 13. VFOA frame based recognition rate vs head pose tracking errors (for the pan angle), plotted per meeting. The VFOA recognizer is the HMM modeling after adaptation.

C. Results with Model Adaptation

Table VII displays the recognition performance obtained with the adaptation framework described in Section VII[6]. For PL, one can observe no improvement when using GT data and a large improvement when using the tracking estimates (e.g. around 10% and 8% for resp. FRR and F_E with the GMM model). In this

[6]In the tables, we recall the values without adaptation for ease of comparison.

person	measure	gt-gmm	gt-gmm-ad	gt-hmm	gt-hmm-ad
L	FRR	72.3	72.3	72.3	72.7
	F_E	62.4	61.2	65.8	66.2
R	FRR	56.8	59.3	57.3	62
	F_E	54.4	56.4	59.5	62.7
person	measure	tr-gmm	tr-gmm-ad	tr-hmm	tr-hmm-ad
L	FRR	47.3	57.1	47.4	53.1
	F_E	34.8	42.8	45.2	47.9
R	FRR	38.1	39.3	38	41.8
	F_E	35.6	37.3	43.8	48.8

TABLE VII
VFOA RECOGNITION RESULTS FOR PERSON LEFT (L) AND RIGHT (R), BEFORE AND AFTER ADAPTATION.

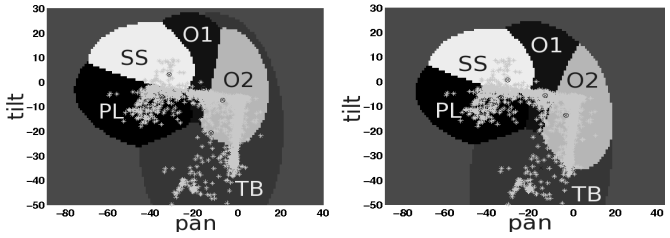


Fig. 14. VFOA decision map example before adaptation (Left) and after adaptation (right). After adaptation, the VFOA of $O1$ and $O2$ correspond to lower tilt values. The cloud of stars represents the tracking head pose estimates used for adaptation.

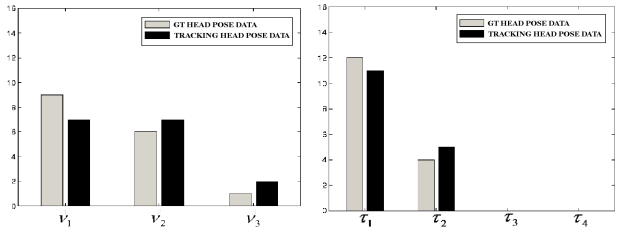
situation, the adaptation is able to cope with the tracking errors and the variability in looking at a given target. For PR, we notice an improvement with both the GT and tracking head pose data. For instance, with the HMM model and tracking data, the improvement is 3.8% and 5% for FRR and F_E. Again, in this situation adaptation can cope with an individual way of looking at the targets, such as correcting the bias in the estimated head tilt , as illustrated in Fig. 14.

When exploring the optimal adaptation parameters estimated through cross-validation, one obtains the histograms of Fig. 15. As can be seen, regardless of the kind of input pose data (GT or estimates), they correspond to configurations giving approximately equal balance to the data and prior w.r.t. the adaptation of the HMM transition matrices (ν_1 and ν_2), and configurations for which the data are driving the adaptation process of the mean pose values (τ_1 and τ_2).

D. Results with the Geometrical VFOA Modeling

Here we report the results obtained when setting the model parameters by exploiting the meeting room geometry (cf Subsection VI-B). This possibility for setting parameters is interesting as it removes the need for data annotation each time a new focus target is considered (a 5th person introduced around the table).

Fig. 16 shows the geometric VFOA Gaussian parameters (mean and covariance) generated by the model when using $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$. As can be seen, the VFOA



(a) (b)

Fig. 15. Histogram of the optimal scale adaptation factor of the HMM prior (a) and HMM VFOA mean (b), selected though cross-validation on the training set, and when working with GT head pose data.

pose values predicted by the model are consistent with the average pose values computed for individuals using the GT data. This is showed in Table VIII, which provides the prediction errors in pan E_{pan} defined as:

$$E_{pan} = \frac{1}{8 \times (K-1)} \sum_{m=1}^{8} \sum_{f_i \in \mathcal{F}/\{U\}} |\bar{\alpha}_m(f_i) - \alpha_m^p(f_i)| \tag{20}$$

where $\bar{\alpha}_m(f_i)$ is the average pan value of the person in meeting m and for the VFOA f_i, and $\alpha_m^p(f_i)$ is the predicted value according to the chosen model (i.e. the pan component of $\mu_{f_i}^g$ or $\mu_{f_i}^l$ in the geometric or learning approaches respectively). The tilt prediction error E_{tilt} is obtained by replacing pan angles by tilt angles in Eq. 20. As can be seen, using cross-validated κ_α and κ_β values provides better results than setting these parameters to the constant values $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$ used in all the recognition experiments reported below. Also, we noticed that usually the κ_α values providing good prediction are lower when using tracking data than when using the ground truth head pose data. A likely explanation is that the head tracker under-estimates the pan angles. Thus, to account for this, a smaller κ_α has to be used to obtain better prediction. Interestingly enough, however, in practice we did not find any particular relationship between an optimal angular prediction (as measured by Eq. 20) and the VFOA recognition results, showing that the selection of these values is not critical. We thus relied on $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$ for all our experiments.

The recognition performance is presented in Table IX. These tables show that, when using GT head pose data, the results are slightly worse than with the training data approach, which is apparent in the similarity of the prediction errors. However, when using the pose estimates, the results are better. For instance, for PL, with adaptation, the FRR improvement is more than 6%. It is interesting and encouraging given that the modeling does not require any training data. Also, we notice that the adaptation always improves the recognition, sometimes quite significantly (see the GT data condition

Method	learned VFOA		geometric VFOA (cross-validation)		geometric VFOA ($\kappa_\alpha = \kappa_\beta = 0.5$)	
Error	E_{pan}	E_{tilt}	E_{pan}	E_{tilt}	E_{pan}	E_{tilt}
PL	6.4	5.1	5.5	6.4	5.8	6.4
PR	5.9	6.1	5.6	7.6	12.8	7.4

TABLE VIII

PREDICTION ERRORS (IN DEGREES) FOR LEARNED VFOA AND GEOMETRIC VFOA MODELS (WITH GT POSE DATA). IN THE GEOMETRIC CROSS-VALIDATED CASE, THE SAME METHODOLOGY THAN IN THE LEARNING CASE IS USED: FOR EACH MEETING THE EMPLOYED κ_α (OR κ_β) HAS BEEN LEARNED ON THE OTHER MEETINGS.

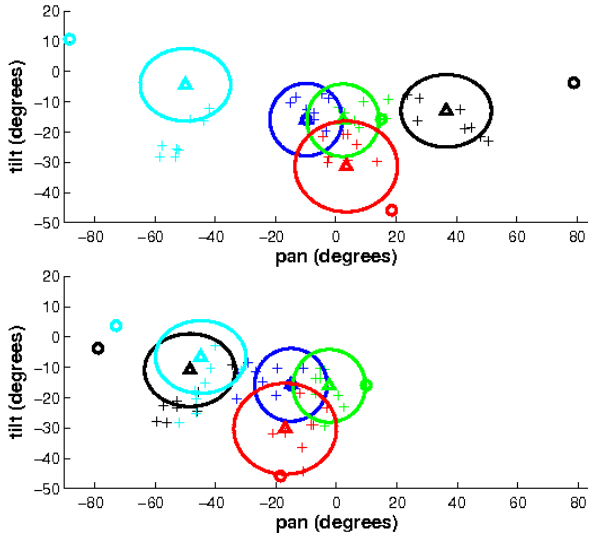


Fig. 16. Geometric VFOA Gaussian distributions for PR (top image) and PL (bottom): the figure displays the gaze target direction (◯), the corresponding head pose contribution according to the geometric model with values $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$ (△ symbols), and the average head pose (from GT pose data) of individual people (+). Ellipses display the standard deviations used in the geometric modeling. black=PL or PR, cyan=SS, blue=$O1$, green=$O2$, red=TB.

for PR, or the tracking data for PL).

Comparison with Stiefelhagen et al [11]: Our results seem quite far from the 73% reported by Stiefelhagen et al [11][7]. Several factors may explain the difference. First, in [11], meeting with 4 people were studied and no other target apart from the other meeting participants was considered. In addition, these participants were sitting at equally spaced positions around the table, optimizing the discriminability between VFOA targets. People were recorded from a camera placed directly in front of them. Hence, due to the table geometry, the majority of head pan lay between $[-45°, 45°]$, where the tracking errors are smaller (see Table I). Ultimately, our results are more

[7]Note that in [11], approaches to recognize the VFOA from audio, and a combination of audio and head pose are also provided. However, for the remainder of this paper, we compare our method with their approach on recognizing the VFOA solely from head pose, since this is the scope of our paper.

person	Measure	gt	gt-ge	gt-ad	gt-ge-ad
L	FRR	72.3	69.3	72.7	70.8
	F_E	65.8	65.2	66.2	65.3
R	FRR	57.3	51.8	62	58.5
	F_E	59.5	53	62.7	59.2
person	Measure	tr	tr-ge	tr-ad	tr-ge-ad
L	FRR	47.4	55.2	53.1	59.5
	F_E	45.2	48.2	47.9	50.1
R	FRR	38	41.1	41.8	42.7
	F_E	43.8	49.1	48.8	50.1

TABLE IX

VFOA RECOGNITION RESULTS FOR PL AND PR USING THE HMM MODEL WITH THE GEOMETRIC VFOA PARAMETER SETTING ($(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$), WITH/WITHOUT ADAPTATION. FOR EASE OF COMPARISON, WE RECALL THE RESULTS WITH THE TRAINING PARAMETER SETTING.

in accordance with the 52% FRR reported by the same authors [36] when using the same framework as in [11] but applied to a 5-person meeting, resulting in 4 possible VFOA targets.

Nevertheless, as comparing algorithm results on different setups is quite difficult, we implemented the methodology proposed in [11], [36] to recognize the VFOA solely from head pose. This methodology consists of first clustering the head pose measurements of an individual person using the K-means algorithm, and then using the outcome to initialize the learning of a GMM similar to the one we presented. Finally, each component of the GMM mixture is associated with a target focus using a set of rules. This approach clearly has several issues, especially when the number of targets is large: how to initialize the K-means algorithm, and how to define the association rules. As no information was given in [11] w.r.t. K-means initialization, we experimented with different alternatives and report the best results, which were obtained using the gaze values predicted by the geometrical model (random initialization produced on average much worse results than those presented, around 10% less). Each component was associated with a focus by taking the mixture with the lowest mean tilt value as the table, and other mixtures were associated to the other VFOA targets based on their respective pan values. The comparative results are given in Table X. They show that our method leads to significant improvements in all conditions. Interestingly enough, the improvement is higher when using uncorrupted head pose measurements (i.e. the GT data). These improvements validate our use of the MAP adaptation framework. Indeed, while in [11] full freedom is given to the data to drive the adaptation process, our experiments show (cf Fig. 15) that the optimal adaptation parameters, selected by cross-validation, give equal importance to the data and the prior set on the GMM parameters to obtain better models.

Method	Stiefelhagen et al [11]				Our model (ge-ad)			
measure	gt-L	tr-L	gt-R	tr-R	gt-L	tr-L	gt-R	tr-R
FRR	61.9	55.7	53.1	39.6	70.8	59.5	58.5	42.7
F_E	53.8	35.1	43.8	34.7	65.3	50.1	59.2	50.1

TABLE X
COMPARISON OF OUR VFOA RECOGNITION APPROACH (HMM WITH GEOMETRIC MODEL AND ADAPTATION) AND [11] (SEE FOOTNOTE 7).

X. CONCLUSION AND FUTURE WORK

In this paper, we addressed the VFOA recognition of meeting participants from their head pose in complex meeting scenarios. Head pose measurements were obtained either through magnetic field sensors or using a head pose tracking algorithm. Several alternative models were studied. Thorough experiments on a large and challenging database made publicly available, gave the following outcome:

- **influence of the physical setup:** when using head pose tracking estimates, average recognition rates of 60% and and 42% were obtained for the left and right seat respectively. It shows that good VFOA recognition can only be achieved if the visual targets of a person are well separated in the head pose angular space, which mainly depends on the person's position in the meeting room.
- **head pose tracking:** accurate pose estimation is essential for good results. Around 11% and 16% error decreases were observed for the left and right seat respectively when using the pose estimates instead of the ground truth. In addition, experiments showed that there exists some correlation between head pose tracking errors and VFOA recognition results.
- **VFOA recognizer model:** the HMM method is performing better than that of the GMM. While this can not be observed with the standard Frame Recognition Rate measure, the newly introduced event-based measure F_E shows that the temporal smoothing introduced by the HMM removes spurious detections in the VFOA estimation.
- **training data vs geometric model:** to avoid the need for training data, we have proposed a novel cognitive model exploiting the room geometry to set the recognizer parameters which links the head pose measures to the VFOA targets. Compared with the standard approach based on training data, and with a state-of-the-art algorithm, the new approach was shown to provide much better results when using the head pose tracking estimates as input.
- **unsupervised adaptation:** results show that in all conditions, automatically adapting the VFOA recognition parameters using the *unlabeled* head pose measurements improves the recognition.

From the above, there are several ways to improve performance. The first one is to increase the separation between the visual targets. However, in practice, this is only feasible for applications which allows for the design of a specific set-up, e.g. a meeting room. Still, increasing the separation is limited by the number of people that we want to accommodate and the activities that people are allowed to perform. The second one is to improve the pose tracking algorithms. This can be achieved using multiple cameras, higher resolution images, or adaptive appearance modeling techniques, preferably in a supervised fashion, by setting up training session to acquire people's appearance at the beginning of a meeting.

A third way to improve VFOA recognition can only come from the prior knowledge embedded in the cognitive and interactive aspects of human-to-human communication. Ambiguous situations such as the one illustrated in Fig. 9(g) and Fig. 9(h), where the same head pose can correspond to two different VFOA targets, could be resolved by the joint modeling of the speaking status and VFOA of all meeting participants. The relationship between speech and VFOA, used for instance in [11], has been shown to exhibit specific patterns in the behavioral and cognitive literature, as already exploited by [12] to derive conversation structures.

Finally, in the case of meetings in which people are moving to the slide screen or white board for presentations, the development of a more general approach that models the VFOA of these moving people will be necessary.

REFERENCES

[1] K. Smith, S. Ba, D. Gatica-Perez, and J.-M. Odobez, "Multi-person wandering focus of attention tracking," in *International Conf. on Multimodal Interfaces*, Banff, Canada, Nov. 2006.

[2] O. Kulyk, J. Wang, and J. Terken, *Machine Learning for Multimodal Interaction*, ser. LNCS 3869. Springer Verlag, 2006, ch. Real-Time Feedback on Nonverbal Behaviour to Enhance Social Dynamics in Small Group Meetings.

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

[4] D. Heylen, "Challenges ahead head movements and other social acts in conversation," in *The Joint Symposium on Virtual Social Agent*, 2005.

[5] 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.

[6] J. N. Bailenson, A. Beal, J. Loomis, J. Blascovitch, and M. Turk, "Transformed social interaction, augmented gaze, and social influence in immersive virtual environments," *Human Comm. Research*, vol. 31, no. 4, pp. 511–537, Oct. 2005.

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

[8] 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.
[9] D. Novick, B. Hansen, and K. Ward, "Coordinating turn taking with gaze," in *Inter. Conf. on Spoken Lang. Processing*, 1996.
[10] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and Application of Influence Rankings in Small Group Meetings," in *International Conf. on Multimodal Interfaces*, Banff, Canada, Nov. 2006.
[11] 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.
[12] 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 *International Conf. on Multimodal Interface (ICMI'05)*, Trento, Italy, Oct. 2005, pp. 191–198.
[13] ICPR-POINTING, "ICPR POINTING'04: Visual observation of deictic gestures workshop," 2004.
[14] CLEAR, "CLEAR evaluation campaign and workshop," 2006.
[15] 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.
[16] I. Malinov, J. Epelboim, A. Herst, and R. Steinman, "Characteristics of saccades and vergence in two kinds of sequential looking tasks," *Vision Research*, 2000.
[17] S. O. Ba and J. M. Odobez, "A Rao-Blackwellized mixed state particle filter for head pose tracking," in *ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy*, 2005, pp. 9–16.
[18] C. Morimoto and M. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer Vision and Image Understanding*, vol. 98, pp. 4–24, 2005.
[19] R. Pieters, E. Rosbergen, and M. Hartog, "Visual attention to advertising: The impact of motivation and repetition," in *Conf. on Advances in Consumer Research*, 1995.
[20] J.-G. Wang and E. Sung, "Study on eye gaze estimation," *IEEE Trans. on Systems, Man and Cybernetics, Part B*, vol. 32, pp. 332–350, 2002.
[21] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *Conf. on Human Factors in Computing Systems*, 2002.
[22] A. Gee and R. Cipolla, "Estimating gaze from a single view of a face," in *International Conf. on Pattern Recognition*, 1994.
[23] T. Horprasert, Y. Yacoob, and L. Davis, "Computing 3d head orientation from a monocular image sequence," in *International Conf. on Automatic Face and Gesture Recognition*, 1996.
[24] T. Cootes and P. Kittipanya-ngam, "Comparing variations on the active appearance model algorithm," in *British Machine Vision Conf. (BMVC)*, 2002.
[25] S. Srinivasan and K. L. Boyer, "Head pose estimation using view based eigenspaces," in *International Conf. on Pattern Recognition*, 2002.
[26] Y. Wu and K. Toyama, "Wide range illumination insensitive head orientation estimation," in *Conf. on Automatic Face and Gesture Recognition*, 2001.
[27] L. Brown and Y. Tian, "A study of coarse head pose estimation," in *IEEE Work. on Motion and Video Computing*, 2002.
[28] L. Lu, Z. Zhang, H. Shum, Z. Liu, and H. Chen, "Model and exemplar-based robust head pose tracking under occlusion and varying expression," in *IEEE Workshop on Models versus Exemplars in Computer Vision (CVPR-MECV)*, Dec. 2001.
[29] M. Danninger, R. Vertegaal, D. Siewiorek, and A. Mamuji, "Using social geometry to manage interruptions and co-worker attention in office environments," in *Conf. on Graphics Interfaces*, Victoria, Canada, 2005, pp. 211–218.
[30] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *TRENDS in Cog. Sciences*, vol. 9(4), pp. 188–194, 2005.
[31] S. Baron-Cohen, "How to build a baby that can read minds: cognitive mechanisms in mindreading," *Cahier de psychologies Cognitive*, vol. 13, pp. 513–552, 1994.
[32] J.-M. Odobez, "Focus of attention coding guidelines," IDIAP Reasearch Institute, Tech. Rep. IDIAP-COM-2, 2006.
[33] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures*, 2004, pp. 183–191.
[34] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Readings in Speech Recognition*, vol. 53A(3), pp. 267–296, 1990.
[35] 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.
[36] R. Stiefelhagen, "Tracking and modeling focus of attention," Ph.D. dissertation, University of Karlsruhe, 2002.

Sileye O. Ba received an M.S. degree in applied mathematics and signal processing from Dakar University in 2000 and an M.S. degree in mathematics, computer vision and machine learning from Ecole Normale Superieure de Cachan, Paris, in 2002. In march 2007 he received his Ph.D. from the Ecole Polytechnique Fédérale de Lausanne (EPFL) that was prepared while he was working as a research assistant at the IDIAP Research Institute. His thesis dissertation covered vision based sequential Monte Carlo methods for head pose tracking and visual focus of attention (VFOA) recognition from video sequences. He is currently a postdoctoral researcher at IDIAP working on using audio-visual features for VFOA and conversational events recognition in meetings.

Dr. Jean-Marc Odobez graduated from the Ecole Nationale Supérieure de Télécommunications de Bretagne (ENSTBr) in 1990, and received his Ph.D degree in Signal Processing and Télécommunication from Rennes University, France in 1994. He performed his dissertation research at IRISA/INRIA Rennes on dynamic scene analysis (image stabilization, object detection and tracking, image sequence coding) using statistical models (robust estimation, 2D statistical labeling with Markov Random Field). He then spent one year as a post-doctoral fellow at the GRASP laboratory, University of Pennsylvania, USA, working on visually guided robotic navigation problems. From 1996 until september 2001, he was associate professor at the Universit du Maine, France. He his now a senior researcher at the IDIAP Research Institute in Martigny, Switzerland. His main area of research are computer vision and pattern recognition, where he is author and coauthor of more than 75 papers in international journals and conferences. He was an active contributor or principle investigator to several IST european projects. He is holding 2 patents on video motion analysis. He is the co-founder of the Swiss Klewel SA company active in the intelligent capture, indexing, and webcasting of conference and seminar events.