IDIAP RESEARCH REPORT

A Study on Visual Focus of Attention Recognition from Head Pose in a Meeting Room

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

IDIAP–RR 06-10

February 2006

published in
3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)

[a] IDIAP Research Institute

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-10

A Study on Visual Focus of Attention Recognition from Head Pose in a Meeting Room

Sileye O. Ba Jean-Marc odobez

February 2006

published in
3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)

Abstract. This paper presents a study on the recognition of the visual focus of attention (VFOA) of meeting participants based on their head pose. Contrarily to previous studies on the topic, in our set-up, the potential VFOA of people is not restricted to other meeting participants only, but includes environmental targets (table, slide screen). This has two consequences. Firstly, this increases the number of possible ambiguities in identifying the VFOA from the head pose. Secondly, due to our particular set-up, the identification of the VFOA from head pose can not rely on an incomplete representation of the pose (the pan), but requests the knowledge of the full head pointing information (pan and tilt). In this paper, using a corpus of 8 meetings of 8 minutes on average, featuring 4 persons involved in the discussion of statements projected on a slide screen, we analyze the above issues by evaluating, through numerical performance measures, the recognition of the VFOA from head pose information obtained either using a magnetic sensor device (the ground truth) or a vision based tracking system (head pose estimates). The results clearly show that in complex but realistic situations, it is quite optimistic to believe that the recognition of the VFOA can solely be based on the head pose, as some previous studies had suggested.

1 Introduction

An important aspect of human being daily life, as social being, is interaction with other humans. A topic of intense study in psychology is the ways these interactions happen in groups such as families or work teams [10]. Human interactions happen through speech or non verbal cues. One one hand, the use of verbal cues in groups is rather well defined because it is tightly connected to to the taught explicit rules of language (grammar, dialog acts). On the other hand, the usage of the non verbal cues is usually more implicit, which does not prevent it from following rules and exhibiting specific patterns in conversations. A person rising hand during a speech often means that he/she is requesting a speaking turn. In a face to face conversation a listener's head nod or shake can be interpreted as agreement or disagreement [6]. Besides hand and head gestures, the visual attention of people, as defined by the eye gaze, is another important cue of non verbal communication which contains information about the course of an interaction. For instance, gaze is often used as a mean to regulate the dialog. A speaker's gaze can be interpreted as a request of back-channels from an audience. Also, a listener can use his gaze to find good time window to request for speaking turns [4, 11]. Furthermore, studies have shown that a person's visual attention was influenced by the visual attention state of other people [7]. Thus, in brief, recognizing the visual attention pattern of group of people can reveal an important amount of information about the social nature of the occuring interactions and help us to better understand them. Due to its importance, after psychologists, computer vision researchers are currently investigating the identification and role of the gaze in social interaction in situations such as meetings in smarts rooms, making use of all the multi-modal abilities of such environments [17, 18].

As part of our effort to study human interaction, the goal of this paper is to analyze the correspondence between the head pose and the eye gaze of people. In other words we want to evaluate how well we can infer the visual focus of attention (VFOA) solely from the head pose. This study, conducted in the context of a smart meeting room environment, complements previous research in this domain. First our study generalizes to more complex situations similar works that have already been conducted in [12, 17]. Contrarily to these previous works, the scenario we consider involves people looking at slides or writing on the table. As a consequence, in our set-up, people have more potential visual focus of attention (6 instead of 3 in [12, 17]). Also, due to the physical spatial configuration of the VFOA targets, the identification of the VFOA can only be done using complete head pose representation (pan and tilt), instead of just the head pan as done previously. Finally, in our set-up, there were ambiguities between VFOA depending on people's sitting location in the room while in the previous work there were not. Thus our study reflects more complex, but realistic, meeting room situations in which people don't just focus their attention on the other people but also on other targets such as the table, the white board, the slide screen, or even sometimes are not focused on any predefined object.

In this paper, we propose to analyze the recognition of the VFOA of people from their head pose. In our experiments, the head poses are either obtain using a magnetic sensor (the ground truth) or a computer vision based probabilistic tracker, allowing to evaluate the degradation in VFOA recognition when going from true values to estimated ones. VFOA are then recognized using either the Maximum A Posteriori principle or an Hidden Markov Models (HMM) modeling, where in both cases the VFOAs are represented using Gaussian distributions. Our early experimental results, based on VFOA frame and event performance measures, show that in the case of more complex environments, previous work results were probably optimistic in concluding that it was possible to infer the VFOA from head pose alone, or, to a lower extent, to use tracking estimates instead of ground-truth head pose measurements.

The remaining of this paper is organized as follows. Section 2 discusses works related to ours. Section 3 indicates the way we obtain head pose measures using either a magnetic field location tracker, or using our probabilistic method for joint head tracking and pose estimation, and compares numerically the latter approaches with the former (ground truth) one. Section 4 describes the considered models for recognizing the VFOA from head pose. In Section 5, we give experimental results and Section 6 concludes the paper.

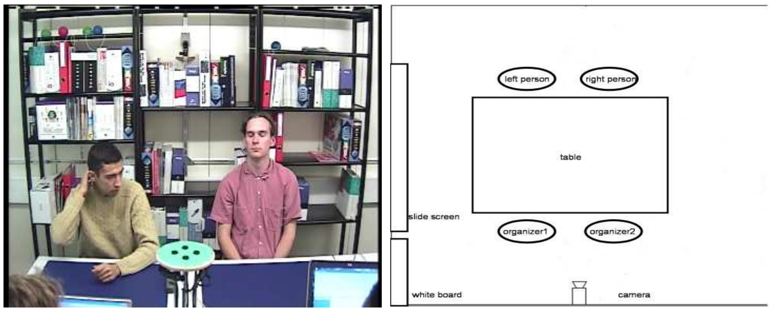

Figure 1: left: sample image of the dataset; right: visual focus of attention targets.

2 Related Work

The estimation of the VFOA of a person from visual input has been studied in the past using different approaches. For instance, in applications studying the visual exploration of images by people, wearable eye gaze trackers are often used. Wearable sensor for eye gaze tracking are infrared based systems. An infrared light is shined on the subject whose gaze is tracked, which creates a red eye effect. The difference of reflexion between the cornea and the pupil is then exploited to determine the gaze direction. As an example, [13] studied people's attentions and reactions to advertisement exposure using such technology, to understand where advertiser should put important information to capture clients' attention. However, besides the concerns over the safety of long exposure to infrared lights, wearable sensor can be used only in controlled experimental situations. In other situations, non invasive procedures to estimate the eye gaze are required. This is the case for instance of systems which aim at automatically detecting driver (loss of) attention. In such cases, computer vision technics can be used, given the availability of high resolution images, to estimate the gaze direction of a driver. In a representative example, [15], motion and skin color distribution are used to track facial features and from the eye balls location reconstruct the gaze direction. In the human computer interaction domain, [9], a similar approach has been used to estimate the gaze location of a worker siting in front of his computer in an office environment.

Although eye gaze tracking with computer vision technics is less invasive than eye gaze tracking with a wearable sensor, it is still relatively constraining, as the subject usually has to remain close to the camera because tracking eye features requires high resolution images. Thus, some people proposed to estimate the VFOA from head pose instead of eye gaze. As a good example, [17] showed that, in a 4 people meeting configuration, the hypothesis that the head is approximately oriented in the same direction as the gaze is a reasonable assumption. In this work, however, there was no ambiguity between the head poses which were defining people's different VFOA, because of their specific meeting physical set-up (four participants evenly space around a round table). Also, the head poses were assimilated to the head azimuth (head pan) only. Following [17], other researchers used the same assumption regarding head pose and eye gaze to model the VFOA of people. For instance, in a very interesting work, [12] makes use of both people's utterance information and head azimuth angle, obtained from a magnetic field head orientation sensor, to infer conversational models in a 4 persons conversation. In [3], the head pose was exploited to model visual attention in office from which workers social geometry was defined, where the social geometry defines when people are available or not for communication.

3 Head Pose Tracking

For our study we used a database comprising 8 meetings of 4 people (duration of meetings ranged from 7 to 14 minutes), recorded in IDIAP's smart meeting room. The scenario of the meeting was to discuss statements displayed on the projection screen. A sample image of the data is shown in Figure 1. Due to technological constraints[1], we were able to capture the head ground truth of only two participants (the left and right person in Fig. 1), using 3D magnetic sensors attached to the head. The head pose is defined as an instance of head rotation with respect to a reference configuration. In general, head poses are represented by three Euler angles (α, β, γ) which parameterize the decomposition of the rotation matrix of the head configuration with respect to the camera frame. Among the possible decompositions, we have selected the one whose rotation axes are rigidly attached to the head. With this choice, we have: α denotes the pan angle, a left/right head rotation; β denotes the tilt angle, an up/down head rotation; and finally, γ, the roll, represents a left/right "head on shoulder" head rotation (see Figure 2). In the following we describe the two alternative techniques that we used to extract the head pose information. The first one consisted of using orientation sensors, and the obtained head poses defined the ground truth values. The second approach provides head pose estimates based on a computer vision probabilistic tracking algorithm.

3.1 Head Pose Ground Truth from Magnetic Sensors

The head pose ground truth was obtained using a 3D location and orientation magnetic sensor called flock of bird (FOB) [19]. The coordinate frame of this sensor was calibrated with respect to the camera frame. In each recording, the time delay between the FOB and the video was set by detecting the occurrence of the same events (peak oscillations) in both modalities.

3.2 Probabilistic Method for Head Pose Tracking

In this subsection, we summarize the Bayesian probabilistic approach described in [1] and which was used to track the head pose.
The probabilistic framework for tracking is well known. Denoting by X_t the hidden state representing the object configuration at time t and by Y_t an observation extracted from the image, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of X_t given all the observations $Y_{1:t}$. In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters. The idea behind particle filter consist in representing the filtering distribution using a set of weighted samples $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrive. Given the particle set of the previous time step, $\{X_{t-1}^n, w_{t-1}^n, n = 1, ..., N_s\}$, configurations of the current step are drawn from a proposal distribution $X_t \sim \sum_n w_{t-1}^n p(X|X_{t-1})$. The weights are then computed as $w_t \propto p(Y_t|X_t)$. Four elements are important in defining a particle filter: i) a state model defining the object we are interested in ii) dynamical model governing the temporal evolution of the state $p(X_t|X_{t-1})$ iii) a likelihood model measuring the adequacy of data given the proposed configuration of the tracked object and iv) a sampling mechanism which have to propose new configurations in high likelihood regions of the state space. These elements along with our model 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 represent the head location and size, θ represents the head in-plane rotation. Both S and θ parameterize a transform $\mathcal{T}_{S,\theta}$ defining the head spatial configuration. The variable l label an element of the discretized set of possible head poses.
Dynamical Model: The dynamical model governing the temporal evolution of the state is defined as

$$p(X_t|X_{1:t-1}) = p(\theta_t|\theta_{t-1}, l_t)p(l_t|l_{t-1}, S_t)p(S_t|S_{t-1}, S_{t-1}) \tag{1}$$

The dynamical model of the head in plane rotation variable θ_t and the discrete head pose variable l_t are learned using the head pose GT training data. While the head location and size variable dynamics is modelled as a second order auto-regressive process.
Observation Model: The observations $Y = (Y^{text}, Y^{col})$ are composed of textures and color observations. Assuming that, given the state value, texture and color observation are independent, the

[1]The magnetic field due to the setup caused distortions on the flock-of-birds readings.

pointing vector			pan			tilt			roll		
mean	std	med	mean	std	med	mean	std	med	mean	std	med
20.3	11.3	18.2	9.10	8.6	7.0	17.6	12.2	15.8	10.1	9.9	7.5

Table 1: pointing vector, pan, tilt and roll errors statistics.

observation likelihood was modeled as:

$$p(Y|X = (S, \theta, l)) = p_{text}(Y^{text}(S, \theta)|l)p_{col}(Y^{col}(S, \theta)|l) \quad (2)$$

where for each head pose variable l, the parameters of the texture likelihood p_{text} and the color likelihood p_{col} were learned from the Prima-Pointing database [5] containing head image appearances of the pose l. For a given hypothesized configuration X, the parameters (S, θ) allow to extract an image patch, on which the features are computed, while the examplar index l allows to select the appropriate appearence model.
Sampling Method: In this work, we use Rao-Blackwellization which consist in applying the standard PF algorithm over the tracking variables S and γ while applying an exact filtering step over 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 technique, the reader is referred to [1].

3.3 Head Pose Tracking Evaluation

For evaluation, we followed the protocol described below.

Data: we used the IHPD database[2]. Amongst the 16 recorded people with ground truth, we used half of the database (8 people) as training set to learn the pose dynamic model and the half remaining as test set to evaluate the tracking algorithms. Because of the scenario used to record data, people often have negative pan values corresponding to looking at the projection screen. The pan values range from -70 to 60 degree. Tilt values range from -60 (when people are writing) to 15 degrees, and roll value from -30 to 30 degrees.
Performance measures: four error measures are used. The three first measures are the errors in pan, tilt and roll angle, i.e. the average of the absolute difference between the pan, tilt and roll of the ground truth (GT) and the tracker estimation. The fourth error measure is defined by the average angular error between the 3D pointing vector (the unit vector of the z-axis of the basis attached to the head cf Figure 2) defined by the head pose GT and the pose estimated by the tracker, which can be used as pose estimation error measure. Note that this vector depends only on the head pan and tilt values (given the selected representation, cf first paragraph of Section 3).
Results: The statistics of the errors over the test set are showed in Table 1. Overall, given the small head size, and the fact that the appearence training set is composed of heads recorded in an external set up (and thus does not contain any individuals from our testing database), the results are quite good, with a majority of head pan errors smaller than 10 degrees. However, these results hide a large discrepancy between individuals. For instance, the average pan error ranges from 4 degrees to 15 degrees, and depends mainly on whether the tracked person resembles one of the person of the training set used to learn the appearance model. The table also shows that the errors in pan and roll are smaller than the errors in tilt. This is due to the fact that, even from a perceptive point of view, discriminating between head tilts is more difficult than discriminating between head pan or head roll [2]. With respect to other works, these results are good. For instance, in [17], a neural net 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 10 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

[2]http://www.idiap.ch/HeadPoseDatabase/

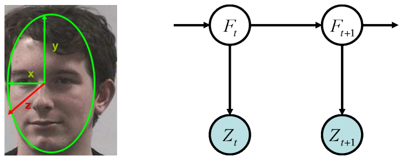

Figure 2: Left: basis attached to the head (head pointing vector in red). Right: visual focus of attention graphical model

rise to 20 degrees. This suggests that their appearance model is fitted to their set-up, to the contrary of our experiments, in which appearance models are trained from an external database [5].

4 Visual Focus of Attention Modeling

VFOA set: For these experiments, we considered only the two persons which have their head pose continuously annotated using FOB (left person and right person). For each one of these person, we identified 6 dominant VFOA of interest and defined the set of visual VFOA as $\mathcal{F} = \{f_1 = person,\ f_2 = organizer1,\ f_3 = organizer2,\ f_4 = table,\ f_5 = slidescreen,\ f_6 = unfocused\}$. For left person f_1 is person right and for right person, f_1 is person left. In the following we present our VFOA modeling approaches.

4.1 Modeling VFOA with a Gaussian Mixture Model (GMM)

Let us denote by $F_t \in \mathcal{F}$ and by Z_t the VFOA and the head pointing vector (defined by its pan and tilt angles) of a person at time instant t. Estimating the VFOA can be posed in a probabilistic framework as finding the label maximizing the a posteriori (MAP) probability:

$$\hat{F}_t = \arg\max_{F_t \in \mathcal{F}} p(F_t|Z_t) \text{ with } p(F_t|Z_t) = \frac{p(Z_t|F_t)p(F_t)}{p(Z_t)} \propto p(Z_t|F_t)p(F_t) \tag{3}$$

For each possible VFOA $f \in \mathcal{F}$ which is not $unfocused$, $p(Z_t|F_t)$ is modeled as a Gaussian distribution $\mathcal{N}(Z_t; \mu_f, \Sigma_f)$ with mean μ_f and full covariance matrix Σ_f. Besides, $p(Z_t|F_t = unfocused)$ is modeled as a uniform distribution. For the distribution over priors, two alternatives were considered. In the first case, no prior was used (i.e. the distribution was uniform), while in the second case, the priors were learned from the considered training data.

4.2 Modeling VFOA with a Hidden Markov Model (HMM)

Modeling the VFOA with a GMM does not account for the temporal dependencies between the VFOA events. As a model of these dependencies, we considered the classical graphical model shown in Figure 2. Given a sequence of VFOA $F_{0:T} = \{F_t, t = 0, ..., T\}$ and a sequence of observations $Z_{1:T}$, the joint posterior probability density function of the states and observation can be written:

$$p(F_{0:T}, Z_{1:T}) = p(F_0) \prod_{t=1}^{T} p(Z_t|F_t)p(F_t|F_{t-1}) \tag{4}$$

The emission probabilities were modeled as in the previous case (i.e. Gaussian distributions for regular VFOA, and uniform distribution for the *unfocused* label). The parameters of these models, along with the discrete transition matrix $p(F_t|F_{t-1})$ modeling the probability to transit from a VFOA to another were learned using standard techniques. In the testing phase, the estimation of the optimal sequence of states given a sequence of observations was conducted using Viterbi algorithm [14].

5 Experimental Results

5.1 Evaluation Set Up

To evaluate our modeling we annotated our IHPD database with the VFOA of person right and left by watching the videos. While there might be some issue about the feasibility of VFOA annotation based on eye gaze, experiments have shown that there might be more than 95% agreement among annotators for this task.

Evaluation protocol. For training and testing, we adopted the leave one out protocol. The data of 7 recordings were used to train the model parameters which are then used to test the recognition system on the remaining one. We defined two kinds of errors measure to evaluate the performances of our modeling.

Frame based recognition rate (FRR) which corresponds to the percentage of correctly estimated VFOA frames. This measure is interesting in term of pure recognition, indicating which proportion of the time the VFOA has been correctly identified. However, this rate can be dominated by long duration VFOA events (where a VFOA event is defined as a temporal segment with the same VFOA label). Since we are also interested in the patterns followed by the VFOA events, which contains information related to the 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 the VFOA annotations and the recognized sequence R obtained through the VFOA estimation process. The GT sequence is defined as $G = \{G_i = (f_i, b_i, e_i),\ i = 1, ..., N_G\}$ where N_G is the number of events, $f_i \in \mathcal{F}$ is the ith VFOA event label, b_i and e_i the beginning and end time of the event f_i. The recognized sequence R is defined similarly. The two sequences are aligned using an adaptive string alignment procedure that take into account the temporal extent of the events. Given this alignment we can compute for each event $f \in \mathcal{F}$, the recall $\rho(f)$ and precision $\pi(f)$ measures of that event defined as:

$$\forall f \in \mathcal{F},\ \rho(f) = \frac{N_{matched}(f)}{N_G(f)} \text{ and } \pi(f) = \frac{N_{matched}(f)}{N_R(f)} \tag{5}$$

where $N_{matched}(f)$ represents the number of events in the recognized sequence labeled f that match the same event in the GT after alignment. $N_G(f)$ (resp $N_R(f)$) denotes the number of occurrences of the event f in the ground truth (resp recognized) sequence. The recall measures the percentage of ground truth events that are correctly estimated while the precision measure the percentage of estimated events that are correct. Both precision and recall need to be high to denote good VFOA recognition performance. The F-measure defined by:

$$\phi(f) = \frac{2\rho(f)\pi(f)}{\rho(f) + \pi(f)} \tag{6}$$

reflects this requirement. Finally, the performance measures of a given person are computed through averaging:

$$\rho = \frac{\sum_{f\in\mathcal{F}} \rho(f)}{|\mathcal{F}|},\ \pi = \frac{\sum_{f\in\mathcal{F}} \pi(f)}{|\mathcal{F}|},\ \text{ and } \phi = \frac{2\rho\pi}{\rho + \pi} \tag{7}$$

The performance measure over the whole database is the average of the precision, recall and F-measure of the 8 individuals.

error measure	gt-ML	gt-gmm	gt-gmm-prior	gt-hmm	tr-ML	tr-gmm	tr-gmm-prior	tr-hmm
frame rr	62.1	53.6	60.7	53.9	42.8	38.2	46.6	38.4
event rec	65.7	57.3	52.3	50.6	54.5	51.5	38.1	34.8
event prec	43.6	43.6	47.3	52.2	18.5	17.1	19.4	40.6
event F-meas	52.1	47.2	48.4	50.4	29.5	25.3	24.9	36.9

Table 2: Average VFOA estimation results for right person using maximum likelihood estimation (ML), GMM modeling with and without prior term and HMM modeling (gt=using ground truth, tr=using tracking output)

error measure	gt-ML	gt-gmm	gt-gmm-prior	gt-hmm	tr-ML	tr-gmm	tr-gmm-prior	tr-hmm
frame rr	78.4	73	75.3	73	53.6	49.5	51	50.1
event rec	66.9	62	55.5	56.4	51.3	39.3	39	32.7
event prec	53.2	56.8	53.1	63.8	26.8	18.9	20.2	44.9
event F-meas	59	58.7	53.8	59.2	34.2	25.2	25.5	36.9

Table 3: Average VFOA estimation results for left person using maximum likelihood estimation (ML), GMM modeling with and without prior term and HMM modeling (gt=using ground truth, tr=using tracking output)

5.2 Experimental Results

5.2.1 Results exploiting the head pose ground truth

In this section we provide the VFOA estimation results when the input data to the algorithms of Section 4 are the head poses obtained from the flock-of-birds sensors.

VFOA and head pose correlation: Table 2 and 3 display the VFOA estimation results for the right and left person respectively. The first column of theses two tables give the results of VFOA maximum likelihood estimation (ML) using the head pose GT data, where the ML approach consists in estimating the VFOA model parameters using the data of a person and testing the model on the same data (when considering the GMM modeling). These results show in an optimistic case the performances our model can achieve, and illustrate somehow the correlation between a person's head poses and his VFOA. As can be seen, this correlation is quite high for the left person (close to 80% FRR), showing the good accordance between pose and VFOA. This correlation, however, drops to near 60% only for the right person. This can be explained by the fact that for person right, there is a strong ambiguity between looking at person left and at the slide screen (see Figure 1). More generaly, the range of azimuth values within which the three other participants and slide screen VFOA target lies has been divided by 2. The average angular distance between these targets is around 20 degrees for right person, a distance which can easily be compensated for using eye movements only (rather than head pose) to change focus. The values in the confusion matrices (Table 4 and 5) corroborate this analysis. The analysis of Tables 2 and 3 shows that this discrepancy holds for all experimental conditions and algorithms (when using ground-truth input), with a performance decrease of approx. 20% and 10% for FRR and event F-measure respectively.

VFOA Prediction: While the MLE is achieving the best results, its performances are not extremely out-performing the performances of the GMM modeling with or without a prior term and the HMM modeling using GT. The GMM and HMM modeling are showing the ability to predict a person VFOA from other persons. For both person right and left, the GMM modeling is achieving better performances in term of frame based recognition rate and event based recall while the HMM is giving better event based precision. This can be explained since the HMM approach is doing some data smoothing. As a results some events are missed (lower recall) but the precision increases due to the elimination of short spurious detections.

VFOA	person left	organizer1	organizer2	table	slide screen	unfocused
person left	54.1	7.9	2.29	0.55	35	0.13
organizer1	2.81	54.9	24.8	1.87	15.2	0.38
organizer2	0.18	8.16	86.4	4.81	0.034	0.38
table	4.13	21.8	31.4	31.7	1.31	9.71
slide screen	21.3	15.2	1.62	0.29	59.8	1.83
unfocused	3.44	15.6	62.5	2.12	2.31	14.1

Table 4: Person right HMM VFOA Estimation frame based confusion matrix using head pose GT

VFOA	person right	organizer1	organizer2	table	slide screen	unfocused
person right	73.8	0.51	25.2	0.36	0.157	0
organizer1	0.91	83.5	12.1	1.45	2.03	0.07
organizer2	12.7	34.2	49.6	3.41	0.01	0
table	14.8	29.1	14.1	38.7	0.61	2.71
slide screen	0.00	4.06	0.09	1.31	93.7	0.79
unfocused	6.52	59.1	17.8	9.71	2.59	4.25

Table 5: Person left HMM VFOA Estimation frame based confusion matrix using head pose GT

VFOA Confusions: Tables 4 and 5 give the VFOA frame recognition confusion matrixes for the person right and left using an HMM modeling. The confusion matrix for person right show that for this person , the VFOA $personleft$ is sometimes confused with the $slidescreen$, $organizer1$ is sometimes confused with $organizer2$ and the $table$ is confused with $organizer1$ and $organizer2$. The confusion for the person left shows that $personright$ is confused with $organizer2$, $organizer1$ is sometimes confused with $organizer2$ and the $table$ with $personright$, $organizer1$ and $organizer2$. These VFOA confusions for person right and left can be explained by the geometry of the meeting room and the fact that people do not need to turn their head to focus on a specific VFOA, they can modify their gaze without modifying their head pose.

Influence of Priors : Figure 3 shows the effect of the prior on the VFOA distribution for person right. the VFOA $organizer2$, represented in this figure by the green area, have its surface reduced while the VFOA $personleft$ (black area), $organizer1$ (blue area), $slidescreen$ (yellow area), $table$ (red area) have their surface extended because these events are more likely. The prior forces the model to concentrate on most likely events while almost removing less likely events. The use of these priors could clearly be a problem when using the VFOA recognition system on other meetings.

We can compare these results to other state of the art VFOA estimation algorithms based on head pose data. [12] is an interesting example. In this paper, the task, amongst others, consists of estimating the VFOA of four people engaged in conversation, using people's speaking status and head pose measured with magnetic sensors. For each person, the potential VFOA were the three other participants. They obtained an average frame based recognition rate of 67.9 %. Despite the lower number of target VFOA, their result is similar to ours (we obtained around 60% for person right and 75% for person left).

5.2.2 Results with Head Pose Estimates:

Table 2 and 3 show also the results for VFOA estimation using tracking results, which exhibit performances degradation w.r.t. GT data. These degradation are mainly due to tracking errors (short periods when the tracker locks on a subpart of the face, tilt uncertainty) and the different (but individually consistent) head pose estimation tracker response to input with similar poses but different

VFOA	person left	organizer1	organizer2	table	slide screen	unfocused
person left	42.5	7.67	2.4	0.69	45.4	1.32
organizer1	12.5	28.6	17.3	2.22	32.9	6.5
organizer2	2.72	37.1	32.3	1.03	11.3	15.5
table	21.4	18.5	19.7	4.03	19	17.4
slide screen	21.4	5.56	2.88	0.32	66.2	3.71
unfocused	1.52	31.4	41.1	0.31	9.59	16

Table 6: Person right HMM VFOA Estimation frame based confusion matrix using head pose tracking estimates

VFOA	person right	organizer1	organizer2	table	slide screen	unfocused
person right	41.5	36.3	9.89	0	0.69	11.7
organizer1	1.49	79	9.51	3.95	4.74	1.29
organizer2	7.13	62.3	22.9	1.44	2.01	4.27
table	5.42	61.6	18.9	0.69	7.28	6.08
slide screen	0.98	23.6	7.62	7.32	56	4.42
unfocused	8.83	76.3	4.71	2.96	6.5	0.71

Table 7: Person left HMM VFOA Estimation frame based confusion matrix using head pose tracking estimates

appearences. Figure 1 shows the effect of tilt estimation errors on the VFOA distributions. While the VFOA distributions in pan of the GT (first row) and the tracking estimates (second row) are similar, the VFOA distributions in tilt are wider for the tracking estimates. The tables also show that, while, when using GT head pose data, the HMM modeling did not have much impact on performances w.r.t. the GMM case, we observe from the reported event F-measure that in presence of noisier data, its smoothing effect is quite beneficial. Tables 6 and 7 display, for respectively the right and left persons, the confusion matrices on the VFOA frame when using an HMM recognizer applied to head pose estimates. The same confusion than using the GT head pose are exhibited, but more pronounced because of the tracking errors (see above) and tilt estimation uncertainties.
Our results -close de 50% frame recognition rate- is quite far from the 73% reported in [17]. Several factors may explain the difference. First, in [17], a 4 people meeting situation was considered and no other VFOA target appart from the other meeting participants was considered. Thus, the pan head angle was sufficient to differentiate VFOA targets. In addition, these participants were sitting at equally spaced angles around a round table, optmising the discrimination between VFOA targets. Finally, it seems from the paper that the head pose tracking algorithm was trained on the face images of the same people appearing in the test video, which resulted in smaller tracking errors.

6 Conclusion and Future Work

In this paper we presented a system to recognize the VFOA of meeting participants from their head pose, the latter being defined by its pan and tilt angles. Such head pose measurements were obtained either through magnetic field sensors or using a probabilistic based head pose tracking algorithm. The experiments showed that, depending on people's position in the meeting room and on the angular distribution of the FOA targets, *the eye gaze may or may not be highly correlated with the head pose.* In absence of such correlation, and if eye white/gaze tracking is unaccessible due to low resolution images, the only way to improve VFOA recognition may only come from the prior knowledge embedded in the cognitive and interactive aspects of human-to-human communication. Ambiguous situations

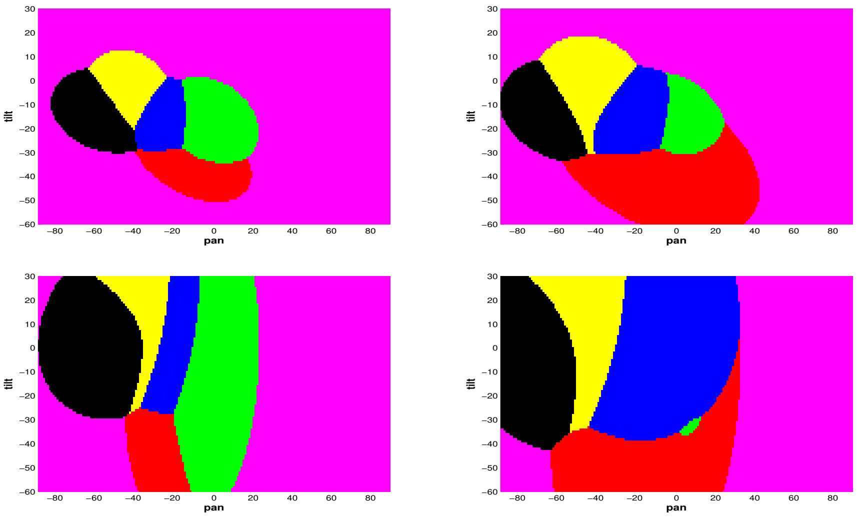

Figure 3: Pan-tilt space VFOA maps using GT (first row) and tracking estimates (second row) for person right. First column: without prior; second column: with priors. black=$personleft$, yellow=$slidescreen$, blue=$organizer1$, green=$organizer2$, red=$table$, magenta =$unfocused$

such as the ones illustrated in Figure 4, where the same head pose can correspond to two different VFOA targets, could be resolved by the joint modeling of the speaking and VFOA characteristics of all meeting participants. Such characteristics have been shown to exhibit specific patterns/statistics in the behavioral and cognitive literature, as already exploited in [12]. This will be the topic of future research.

Besides, as shown by the experiments, there indeed exists some correlation between head pose tracking errors and VFOA recognition results. Improving the tracking algorithms, e.g. using multiple cameras, higher resolution images or adaptive appearence modeling techniques, would thus improve the VFOA results. 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. This has been one topic of our recent research [16].

References

[1] Ba, S.O., Odobez, J.-M.: A Rao-Blackwellized Mixed State Particle Filter for Head Pose Tracking. in Meetings. In Proc. of ACM ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy October 7, 2005

[2] Brown, L., Tian, Y.: A study of Coarse Head Pose Estimation. In Proc. of IEEE Workshop on Motion and Video Computing,Orlando Florida, Dec 2002

[3] Danninger, M., Vertegaal, R., Siewiorek, D., P., Mamuji, A.: Using Social geometry to manage interruptions and co-worker attention in office environments. In Proc. of Conference on Graphics Interface, Victoria, British Columbia, 2005

[4] Duncan Jr., S.: some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology, 23(2),pp283-292, 1972

Figure 4: Ambiguity in focus: despite the high visual similarity of the head pose of the right person, the two focus are different (left image: left person: right image: slide screen). Resolving such cases can only be done by using context (speaking status, other's people gaze, slide activity etc).

[5] Gourier, N., Hall, D., CrowleyJ., L.: Estimating face orientation from robust detection of salient facial features. in Proc. of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK

[6] Heylen, D.: Challenges ahead head movements and other social acts in conversation. in Proc. of The Joint Symposium on Virtual Social Agent, 2005

[7] Langton,S.R.H., Watt, R. J., Bruce, V.: Do the eyes have it? Cues to the direction of social attention. Trends in Cognitive Sciences, vol 4, no2, pp50-58, February 2000

[8] Langton,S.R.H.: The mutual influence of gaze and head orientation in the analysis of social attention direction. The Quarterly Journal of Experimental Psychology, 53A(3),825-845, 2000

[9] Matsumoto, Y., Ogasawara, T., Zelinsky, A: Behavior recognition based on head pose and gaze direction measurement. In Proc. of Conference on Intelligent Robots and Systems, 2002

[10] MacGrath, J. E.: Groups: Interaction and performances Prentice-Hall, Inc., Englewoods Cliffs, N.J., 07632, 1984

[11] Novick, D., Hansen, B., Ward, K.: Coordinating turn taking with gaze. In Proc. of International Conference on Spoken Language Processing, october 1996

[12] Otsuka, K., Takemae, Y., Yamato, J., Murase, H.: A probabilistic inference of multi party-conversation structure based on Markov switching models of gaze patterns, head direction and utterance. In Proc. of International Conference On Multi-modal and Interfaces, Trento, 2005

[13] Pieters, R. G. M., Rosbergen, E., Hartog., M.: Visual attention to advertising: The impact of motivation and repetition. In Proc. of Conference on Advances in Consumer Research, 1995.

[14] Rabiner, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. Readings in Speech Recognition, pp. 267–296, 1990

[15] Smith,P., Shah, M., da Vitoria Lobo, N.: Determining driver visuual attention with one camera. IEEE Transaction on Intelligent Transportation Systems, Vol 4.No4, pp. 205-218, December 2004

[16] Smith K., Ba S., Gatica-Perez D. and Odobez J.M.: Multi-Person Wandering-Focus-of-Attention Tracking. IDIAP Research Report 80, Nov., 2005.

[17] Stiefelhagen, R., Yang, J., Waibel, A.: Estimating focus of attention based on gaze and sound. In Proc. of Workshop on Perceptive User Interface (PUI), pp1-9, Florida USA, Nov 2001

[18] Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., Lathoud, G.: Modeling individual and group action in meetings: a two-layer hmm framework. In Proc. of IEEE CVPR Workshop on Event Mining in Video, Washington DC, July 2004

[19] Flock of Birds: http://www.ascension-tech.com/products/flockofbirds.php