Head pose tracking and focus of attention recognition algorithms in meeting rooms

Sileye O. Ba[1] and Jean-Marc Odobez[1]

IDIAP Research Institute, Martigny, Switzerland

Abstract. The paper presents an evaluation of both head pose and visual focus of attention (VFOA) estimation algorithms in a meeting room environment. Head orientation is estimated using a Rao-Blackwellized mixed state particle filter to achieve joint head localization and pose estimation. The output of this tracker is exploited in an Hidden Markov Model (HMM) to estimate people's VFOA. 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), which renders the task more difficult due to more ambiguity between VFOA target directions. By relying on a corpus of 8 meetings of 8 minutes on average featuring 4 persons involved in the discussion of statements projected on a slide screen, and for which head orientation ground truth was obtained using magnetic sensor devices, we thoroughly assess the performance of the above algorithms, demonstrating the validity of our approaches and pointing out to further research directions.

1 Introduction

The automatic analysis of human interaction constitutes a rich research field. In particular, meetings exemplify the multimodal nature of human communication and the complex patterns that emerge from the interaction between multiple people [6]. Besides, in view of the amount of relevant information in meetings suitable for automatic extraction, meeting analysis has attracted attention in fields spanning computer vision, speech processing, human-computer interaction, and information retrieval [13]. In this view, the tracking of people and of their activity is relevant for high-level multimodal tasks that relate to the communicative goal of meetings. Experimental evidence in social psychology has highlighted the role of non-verbal behavior (e.g. gaze and facial expressions) in interactions [9], and the power of speaker turn patterns to capture information about the behavior of a group and its members [6, 9]. Identifying such multimodal behaviors requires reliable people tracking.

In the present work, we investigate the estimation of head pose from video, and its use in the inference of the VFOA of people. To this end, we propose two algorithms to solve each of the task, and the objective is to evaluate how well they perform and how well we can infer the VFOA solely from the head pose.
Many methods have been proposed to solve the problem of head tracking and pose estimation. They can be grossly separated into two groups. The first group considers the problem of head tracking and pose estimation as two separate and independent problems: the head location is found, then processed for pose

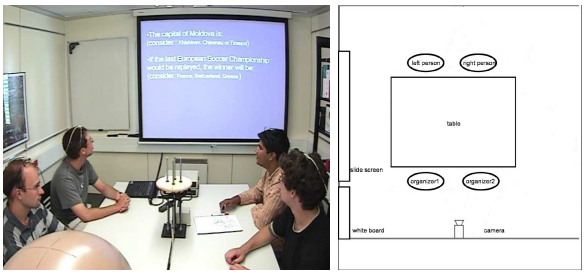

Fig. 1. Left: meeting room. Right: The set $\mathcal{F}$ of potential FOA comprises: other participants, the table, the slidescreen, the whiteboard, and an unfocus label when none of the previous applies.

estimation [2, 13, 10, 15, 17]. The main advantage is usually a fast processing, but then head pose estimation is highly dependent on the head tracking accuracy. Indeed, it has been shown that head pose estimation is very sensitive to head localization [2]. To address this issue, the second group of methods [3, 5, 14] considers jointly the head tracking and pose estimation problems, and we follow this approach.

In meeting data, it is often claimed that head pose can be reasonably used as a proxy for gaze (which usually calls for close views). In this paper, we evaluate tha validity of this assumption by generalizing to more complex situations similar works that have already been conducted in [8, 11]. 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 [8, 11]), leading to more ambiguities between VFOA, and 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. 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 room targets. In this work, we analyze the recognition of the VFOA of people from their head pose. VFOA are 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. In our experiments, the head poses are either obtain using a magnetic sensor or a computer vision based probabilistic tracker, allowing to evaluate the degradation in VFOA recognition when going from true values to estimated ones.

The remainder of this paper is organized as follows. Section 2 describes our database and the protocols used for evaluation. Section 3 and 4 respectively presents our head pose tracking and VFOA recognition algorithms. Results and analysis of the evaluation are reported in 5 and Section 6 concludes the paper.

2 Databases and Protocols

In this section, we describe the data and performance measures used to evaluate head pose estimation algorithms and VFOA recognition algorithms. In the latter case, the emphasis is on the recognition of a finite set $\mathcal{F}$ of specific FOA loci.

2.1 The database

Our evaluation exploits the IDIAP Head Pose Database[1]. In view of the limitations of visual inspection for evaluation, and the inaccuracy obtained by manually labeling head pose in real videos, we decided to record a video database with head pose ground truth produced by a flock-of-birds device. At the same time, as the database is also annotated with the discrete FOA of participants, we will be able to evaluate the impact of having the true vs an estimated head pose on the VFOA recognition.

Content description: the database comprises 8 meetings involving 4 people (duration ranged from 7 to 14 minutes), recorded in IDIAP's smart meeting room. The scenario was to discuss statements displayed on the projection screen. There were restrictions neither on head motions, nor on head poses. In each meeting, the head pose of two persons was continuously annotated ground truth of two participants (the left and right person in Fig. 1) using 3D magnetic sensors attached to the head, resulting in a video database of 16 different people.

Head pose annotation: the head pose configuration with respect to the camera was ground truthed. This pose is defined 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 to report and comment the results. 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.

VFOA set and annotation: for each of the two person ('left' and 'right' in Fig. 1), the set of potential focus is composed of the other participants, the slide-screen, the table, and an additional label (unfocused) when none of the previous could apply. As a person can not focus on himself/herself, the set of focus is thus different from person to person. For instance, for the left person, we have: $\mathcal{F} = \{right\_person, organizer1, organizer2, slide\_screen, table, unfocus\}$. The guidance for the annotation are given in [7].

2.2 Evaluation protocol

Head pose protocol: *Data and protocol.* Amongst the 16 recorded people, 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. In addition, from the 8 meetings of the test set, we selected 1 minute of recording (1500 video frames) for evaluation data. This decision was made to save machine computation time. Pan values range from -60 to 60 degrees (with a majority of negative values corresponding to looking at the projection screen). Tilt values range from -60 to 15 degrees (due to the camera looking down at people) and roll value from -30 to 30 degrees. *Performace measures:* four error measures are used. The three first measures are the errors in pan, tilt and roll angle, i.e. the absolute difference between the pan, tilt and roll of the ground truth (GT) and

[1] Available at http://mmm.idiap.ch/HeadPoseDatabase/

the tracker estimation. Also, the angle between the 3D pointing vector (the vector indicating where the head is pointing at (cf Figure 2) defined by the head pose GT and the pose estimated by the tracker was used as pose estimation error measure. This vector depends only on the head pan and tilt values (given the selected representation). For each error, mean, standard deviation and median (less sensitive to large errors due to erroneous tracking) values are reported.

VFOA protocol: *Data and protocol:* Experiments on FOA recognition are done separately for the left and right person (see Fig. 1). Thus, for each seating position, we have 8 sequences. We adopt a leave-one-out protocol, where for each sequence, the parameters of the recognizer that is applied to this sequence are learned on the 7 other sequences. *Performance measures:* two different types of measures are used.

- *frame-based recognition rate*: this corresponds to the percentage of frames in the video whose estimated FOA match the ground truth label. To avoid the emphasis on events that are long (i.e. when someone is continuously focused) we propose below alternative measures that may better reflect how well an algorithm is able at recognizing events, whether long or short, which might be more suited to understanding meeting dynamics and human interaction.
- *event-based recall/precision*: we are given two sequences of FOA events: the recognized sequence of FOA, $R = \{R_i\}_{i=1..N_R}$ and the ground truth sequence $G = \{G_j\}_{j=1..N_G}$. To compare the 2 sequences, we first apply an adapted string alignment procedure that account for time overlap to match events in the GT and R. Given this alignement, we can then compute for each event $l \in \mathcal{F}$, the *recall* ρ, *precision* π, and F measures of that event, defined as:

$$\forall l \in \mathcal{F}, \rho(l) = \frac{N_{mat}(l)}{N_G(l)}, \pi(l) = \frac{N_{mat}(l)}{N_R(l)} \text{ and } \frac{1}{F_{meas}(l)} = \frac{1}{2}(\frac{1}{\rho(l)} + \frac{1}{\pi(l)}) \quad (1)$$

  where $N_{mat}(l)$ represents the number of events l in the recognized sequence that match the same event type in the ground truth after the alignment, $N_R(l)$ denotes the number of occurence of event l in the recognition sequence, and $N_G(l)$ denotes the number of occurence of l in the ground truth. Qualitatively, the recall of l indicates the percentage of correctly recognized true looks at FOA l, while the precision indicates the percentage of looks at l that were recognized and indeed corresponds to the ground truth. The *F measure*, defined as the harmonic mean of the precision and recall, represents a composite value[2]. Finally, performance measures for the whole database are obtained through averaging of the recall, precision and F measures first over event types per person, then over individuals.

3 Head Pose Tracking

To address the tracking issue, we formulate the coupled problems of head tracking and head pose estimation in a Bayesian filtering framework, which is then solved through sampling techniques. In this paragraph, we expose the main points of our approach. More details can be found in references [1].

[2] Often, increasing the recall tends to decrease the precision, and vice-versa.

3.1 Head Pose Models

We use the Pointing'04 database to build our head pose model. Texture and color based head pose models are built from all the sample images available for each of the 93 discrete head poses $\theta \in \Theta = \{\theta_j = (\alpha_j, \beta_j, 0), j = 1, ..., N_\Theta\}$. In the Pointing database, there are 15 people per pose.

Head Pose Texture Model The head pose texture is represented by the output of three filters: a Gaussian at coarse scale and two Gabor filters at two different scales (finer to coarser). Training patch images are resized to the same reference size 64×64, preprocessed by histogram equalization to reduce light variations effects, then filtered by each of the above filters. The filter outputs at sample locations inside a head mask are concatenated into a single feature vector.
To model the texture of head poses, the feature vectors associated with each head pose $\theta \in \Theta$ are clustered into K=2 clusters using a kmeans algorithm. The cluster centers $e_k^\theta = (e_{k,i}^\theta)$ are taken to be the exemplars of the head pose θ. The diagonal covariance matrix of the features $\sigma_k^\theta = diag(\sigma_{k,i}^\theta)$ inside each cluster is also exploited to define the pose likelihood models. The likelihood of an input head image, characterized by its extracted features z^{text}, with respect to an exemplar k of a head pose θ is then defined by:

$$p_T(z|k,\theta) = \prod_i \frac{1}{\sigma_{k,i}^\theta} \max(\exp -\frac{1}{2}\left(\frac{z_i^{text} - e_{k,i}^\theta}{\sigma_{k,i}^\theta}\right)^2, T) \tag{2}$$

where $T = \exp -\frac{9}{2}$ is a lower threshold set to reduce the effects of outlier components of the feature vectors.

Head Pose Color Model To gain robustness to background clutter and help tracking, a skin color model M_k^θ is learned from the training images belonging to a each head pose exemplar e_k^θ. Training images are resized to 64×64, then their pixels are classified as skin (pixel value=1) or non skin(value=0). The mask M_k^θ is the average of training skin images. Additionally we model the distribution of skin pixel values with a Gaussian distribution [16] in the normalized RG space whose parameters are learned from the training images and continuously adapted during tracking.
The color likelihood of an input patch image at time t w.r.t. the k^{th} exemplar of a pose θ is obtained by detecting the skin pixels on the 64x64 grid, producing this way the skin color mask z_t^{col}, from which the color likelihood is defined as:

$$p_{col}(z|k,\theta) \propto \exp -\lambda ||z_t^{col} - M_k^\theta||_1 \tag{3}$$

where λ is a hyper parameter learned from training data, and $||.||_1$ denotes the L_1 norm.

3.2 Joint Head Tracking and Pose Estimation

The Bayesian formulation of the tracking problem is well known. Denoting by X_t the hidden state representing the object configuration at time t, and by z_t the observation extracted from the image, the objective is to estimate the filtering

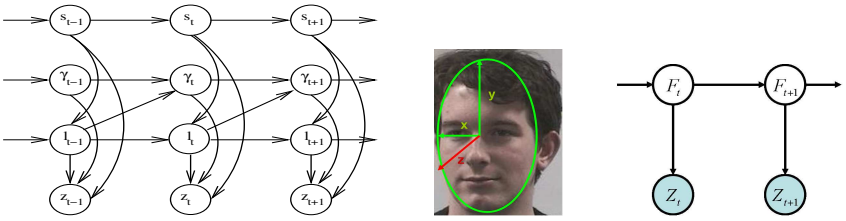

Fig. 2. Left: Mixed State Graphical Model. Middle: basis attached to the head (head pointing vector in red). Right: visual focus of attention graphical model

distribution $p(X_t|z_{1:t})$ of X_t given all the observations $z_{1:t} = (z_1 \dots z_t)$ up to the current time. This can be done through a recursive equation, which can be approximated through sampling techniques (or particle filters PF) in the case of non-linear and non-Gaussian models. The basic idea behind PF consists of representing the filtering distribution using a weighted set of samples $\{X_t^n, w_t^n\}_{n=1}^{N_s}$, and updating this representation as new data arrives. That is, given the particle set at the previous time step $\{X_{t-1}^n, w_{t-1}^n\}$, configurations at the current time step are drawn from a proposal distribution $q(X_t) = \sum_n w_{t-1}^n p(X_t|X_{t-1}^n)$. The weights are then computed as $w_t^n \propto p(z_t|X_t^n)$. Four elements are important in defining a PF: a state model, a dynamical model, an observation model, and a sampling mechanism. We now describe each of them.

State Model: The mixed state approach [12], allows to represent jointly in the same state variable discrete variables and continuous variables. In our specific case the state $X = (S, \gamma, l)$ is the conjunction of a discrete index $l = (\theta, k)$ which labels an element of the set of head pose models e_k^θ, while both the discrete variable γ and the continuous variable $S = (x, y, s^x, s^y)$ parameterize the transform $\mathcal{T}_{(S,\gamma)}$ defined by:

$$\mathcal{T}_{(S,\gamma)}u = \begin{pmatrix} s^x & 0 \\ 0 & s^y \end{pmatrix} \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix} u + \begin{pmatrix} x \\ y \end{pmatrix}. \tag{4}$$

which characterizes the image object configuration. (x, y) specifies the translation position of the object in the image plane, (s^x, s^y) denote the width and height scales of the object according to a reference size, and γ specifies the in-plane rotation of the object.

Dynamic Model: This model represents the temporal prior on the evolution of the state. Figure 2 describes the dependencies between our variables from which the equation of the process density can be defined:

$$P(X_t|X_{t-1}) = p(S_t|S_{t-1})p(l_t|l_{t-1}, S_t)p(\gamma_t|\gamma_{t-1}, l_{t-1}) \tag{5}$$

The dynamical model of the continuous variable S_t, $p(S_t|S_{t-1})$ is modeled as a classical first order auto regressive process. The other densities, learned from training sequences, allow to set some prior on the head eccentricity, as well as to model the head rotation dynamic, as detailed in [1].

The observation model $p(z_t|X_t)$ measures the adequacy between the observation and the state. This is an essential term, where data fusion occurs, and

whose modeling accuracy can greatly benefit from additional discrete variables in the state space. In our case, observations z are composed of texture and color observations (z^{text}, z^{col}), and the likelihood is defined as follows :

$$p\left(z|X=(S,\gamma,l)\right) = p_{text}(z^{text}(S,\gamma)|l)p_{col}(z^{col}(S,\gamma)|l), \tag{6}$$

where we have assumed that these observations were conditionally independent given the state. The texture likelihood p_{text} and the color likelihood p_{col} have been defined in Section 3.1.During tracking, the image patch associated with the image spatial configuration of the state space, (S,γ), is first cropped from the image according to $\mathcal{C}(S,\gamma) = \{\mathcal{T}_{(S,\gamma)}u,\ u \in \mathcal{C}\}$, where $\mathcal{C}$ corresponds to the set of 64x64 locations defined in a reference frame. Then, the texture and color observations are computed using the procedure described in sections 3.1.

Sampling mechanism: the Rao-Blackwellization. The sampling should places new samples as close as possible to regions of high likelihood. The plain particle filter (PF), denoted MSPF, described in the first paragraph of this subsection, can be employed. However, given that the exemplar label l is discrete, its filtering pdf can be exactly computed given the samples of the remaining variables. Thus we can apply the Rao-Blackwellization procedure, which is known to lead to more accurate estimates with a fewer number of particles [4].
Given the graphical model of our filter (Fig.2), the Rao-Blackwellized particle filter (RBPF) consists of applying the standard PF algorithm over the tracking variables S and γ while applying an exact filtering step over the exemplar variable l, *given a sample of the tracking variables*. In this way, computing the likelihood of the state can be done using:

$$p(S_{1:t},\gamma_{1:t},l_{1:t}|z_{1:t}) = p(l_{1:t}|S_{1:t},\gamma_{1:t},z_{1:t})p(S_{1:t},\gamma_{1:t}|z_{1:t}) \tag{7}$$

In practice, only the sufficient statistics $p(l_t|S_{1:t},\gamma_{1:t},z_{1:t})$ of the first term in the right hand side is computed and is involved in the PF steps of the second term. Thus, in the RBPF modeling, the pdf in Equation 7 is represented by a set of particles

$$\{S^i_{1:t},\gamma^i_{1:t},\pi^i_t(l_t),w^i_t\}_{i=1}^{N_s} \tag{8}$$

where $\pi^i_t(l_t) = p(l_t|S^i_{1:t},\gamma^i_{1:t},z_{1:t})$ is the pdf of the exemplars given a particle and a sequence of measurements, and $w^i_t \propto p(S^i_{1:t},\gamma^i_{1:t}|z_{1:t})$ is the weight of the tracking state particle. Figure 3 summarizes the steps of the RBPF algorithm with the additional resample step to avoid sampling degeneracy. In the following, we detail the methodology to derive the exact steps to compute $\pi^i_t(l_t)$ and the PF steps to compute w^i_t.

Deriving the Exact Step: The goal here is do derive $p(l_t|S_{1:t},\gamma_{1:t},z_{1:t})$. As l_t is discrete, this can be done using prediction and update steps similar to those involved in Hidden Markov Model (HMM), and generates as intermediate results $Z_1(S_t,\gamma_t) = p(S_t,\gamma_t|S_{1:t-1},\gamma_{1:t-1},z_{1:t-1})$ and $Z_2 = p(z_t|S_{1:t},\gamma_{1:t},z_{1:t-1})$.

Deriving the PF steps: The pdf $p(S_{1:t},\gamma_{1:t}|z_{1:t})$ is approximated using particles whose weight are recursively computed using the standard PF approach.

1. **initialization step:** $\forall\, i$ sample (S_0^i, γ_0^i) from $p(S_0, \gamma_0)$, and set $\pi_0^i(.)$ uniform and $t = 1$
2. **prediction of new head location configurations:** sample $\tilde{S}_t^i$ and $\tilde{\gamma}_t^i$ from the mixture $(\tilde{S}_t^i, \tilde{\gamma}_t^i) \sim p(S_t|S_{t-1}^i) \sum_{l_{t-1}} \pi_{t-1}^i(l_{t-1}) p(\gamma_t|\gamma_{t-1}^i, l_{t-1})$
3. **head poses distribution of the particles:** compute the exact step $\tilde{\pi}_t^i(l_t) = p(l_t|S_{1:t}^i, \gamma_{1:t}^i, z_{1:t})$ for all i and l_t
4. **particles weights:** for all i compute the weights $w_t^i = p(z_t|S_{1:t}^i, \gamma_{1:t}^i, z_{1:t-1})$
5. **selection step:** resample N_s particle $\{S_t^i, \gamma_t^i, \pi_t^i(.), w_t^i = \frac{1}{N_s}\}$ from the set $\{\tilde{S}_t^i, \tilde{\gamma}_t^i, \tilde{\pi}_t^i(.), \tilde{w}_t^i\}$, set $t = t + 1$ go to step 2

Fig. 3. RBPF Algorithm.

Using the discrete approximation of the pdf at time $t-1$ with the set of particles and weight, the current pdf $p(S_{1:t}, \gamma_{1:t}|z_{1:t})$ can be approximated (up to the proportionality constant $p(z_t|z_{1:t-1})$) by:

$$p(z_t|S_{1:t}, \gamma_{1:t}, z_{1:t-1}) \sum_{i=1}^{N_s} w_{t-1}^i p(S_t, \gamma_t|S_{1:t-1}^i, \gamma_{1:t-1}^i, z_{1:t-1}) \tag{9}$$

to which the standard PF steps can be applied. Indeed, the mixture in the the second part of Equation 9 can be rewritten as:

$$\sum_{i=1}^{N_s} w_{t-1}^i p(S_t|S_{t-1}^i) \sum_{l_{t-1}} \pi_{t-1}^i p(\gamma_t|\gamma_{t-1}, l_{t-1}) \tag{10}$$

which embeds the temporal evolution of the head configurations and allows to draw new (S_t, γ_t) samples. Similarly, the weight of this new samples, defined by the observation likelihood $p(z_t|S_{1:t}, \gamma_{1:t}, z_{1:t-1})$ can be readily obtained from the exact steps computation (cf the computation of the Z_2 constant).

Filter output: As the set of particles defines a pdf over the state space, we can use as output the expectation value of this pdf, obtained by standard averaging over the particle set. Note that usually, with mixed-state particle filters, averaging over discrete variable is not possible (e.g. if a discrete index represents a person identity). However, in our case, there is no problem since our discrete indices correspond to real Euler angles which can be combined.

4 Visual Focus of Attention Tracking

Modelling 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

	pan			tilt			roll			pointing vector		
	mean	std	med	mean	std	med	mean	std	med	mean	std	med
MSPF	10.0	9.6	7.8	19.4	12.7	17.5	11.5	9.9	8.8	22.5	12.5	20.1
RBPF	9.10	8.6	7.0	17.6	12.2	15.8	10.1	9.9	7.5	20.3	11.3	18.2

Table 1. Mean, standard deviation and median of errors on the different angles.

posteriori (MAP) probability:

$$\hat{F}_t = \arg\max_{F_t \in \mathcal{F}} p(F_t|Z_t) \text{ with } p(F_t|Z_t) = \frac{p(Z_t|F_t)p(F_t)}{p(Z_t)} \propto p(Z_t|F_t)p(F_t) \quad (11)$$

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 $p(F_t)$, we indeed used no prior (i.e. the distribution was uniform), in order to obtain a more general model of FOA and avoid overfitting to the considered specific scenario with roles (organizers, participants) that we considered.

Modeling VFOA with a Hidden Markov Model (HMM) The GMM modelling 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 1. 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:

$$p(F_{0:T}, Z_{1:T}) = p(F_0) \prod_{t=1}^{T} p(Z_t|F_t)p(F_t|F_{t-1}) \quad (12)$$

The emission probabilities were modeled as in the previous case (i.e. Gaussian distributions for regular VFOA, and uniform distribution for the *unfocused* label). Their parameters, along with the 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.

5 Results

5.1 Head pose evaluation

Experiments following the protocol described in Section 2.2 were conducted to compare head pose estimation based on the MSPF and the RBPF tracker. The MSPF tracker was run with 200 hundred particles and the RBPF with 100 particles. Except this difference, all the other models/parameters involved in the algorithm were the same (remember that both approaches are based on the same graphical model and involve the setting/learning of the same pdf).

Table 1 shows the pose errors for the two methods over the test set. Overall, given the small head size, and the fact that none of the head in the test set were used for appearence training, the results are quite good, with a majority of head

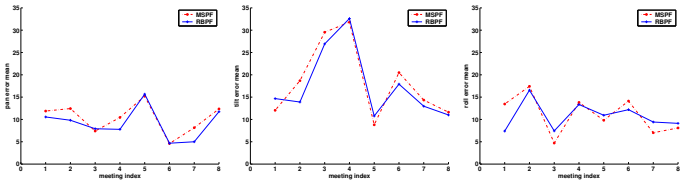

Fig. 4. Pan, tilt, and roll errors over individual participants.

Fig. 5. Sample of tracking failure for MSPF. First row : MSPF; Second row: RBPF.

pan errors smaller than 10 degrees. Also, 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]. Besides, as can be seen, the errors are smaller for the RBPF than for the MSPF approach. This improvement is mainly due to a better exploration of the configuration space of the head poses with the RBPF, as illustrated in Figure 5 which displays sample tracking results of one person of the test set. Because of a sudden head turn, the MSPF lags behind in the exploration of the head pose configuration space, to the contrary of the RBPF approach which nicely follows the head pose. The above results, however, hide a large discrepancy between individuals, as the mean errors for each person of the test set show (Fig. 4). This variance depends mainly on whether the tracked person resembles one of the person of the training set used to learn the appearance model. It is worth noticing in this figure that the improvements due to the Rao-Blackwellisation are more consistent on the marginalized variables (pan and tilt) than on the sampled one (the roll).

5.2 Focus of attention recognition evaluation

Table 2 and 3 display the VFOA estimation results for the right and left person respectively. VFOA and head pose correlation: The ML results corresponds to the maximum likelihood estimation (ML) of the VFOA, which consists in estimating the VFOA model parameters using the data of a person and testing the model on the same data (with a GMM model). These results show in an optimistic case the performances our model can achieve, and illustrate somehow

error measure	gt-ML	gt-gmm	gt-hmm	tr-ML	tr-gmm	tr-hmm
frame rr (FRR)	62.1	53.6	53.9	42.8	38.2	38.4
event rec	65.7	57.3	50.6	54.5	51.5	34.8
event prec	43.6	43.6	52.2	18.5	17.1	40.6
event F-meas	52.1	47.2	50.4	29.5	25.3	36.9

Table 2. Average VFOA estimation results for right person using (ML), GMM, and HMM modeling, and either gt (ground truth) or tr (pose tracking output) observations.

error measure	gt-ML	gt-gmm	gt-hmm	tr-ML	tr-gmm	tr-hmm
frame rr (FRR)	78.4	73	73	53.6	49.5	50.1
event rec	66.9	62	56.4	51.3	39.3	32.7
event prec	53.2	56.8	63.8	26.8	18.9	44.9
event F-meas	59	58.7	59.2	34.2	25.2	36.9

Table 3. Average VFOA estimation results for left person using (ML), GMM, and HMM modeling, and either gt (ground truth) or tr (pose tracking output) observations.

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. However, it drops to near 60% only for the right person, mainly due to the stronger ambiguity between looking at person left, slide screen and to a smaller extent, left organizer. VFOA Prediction: While ML is achieving the best results, its performances are not extremely outperforming the performances of the GMM and HMM modeling using GT data, which show the ability to learn a VFOA model applicable to new data. For both person right and left, the GMM modeling is achieving better frame recognition rate and event recall performance while the HMM is giving better event 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. Overall, our results are comparable to other state of the art VFOA estimation using sensor input. For instance, [8] with a VFOA target set composed of 3 people obtained an average frame recognition rate of 68%, similar to our results. Head pose estimates: As tables 2 and 3 show, we observe a degradation in performance when using head pose estimates. This degradation are 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 appearences. While the HMM modeling had only a small impact on performance when using GT data, we observe from the event F-measure that in presence of noisier data, its smoothing effect is quite beneficial.

6 Conclusion

We have presented a system for the recognition of the VFOA of people in meetings. The method relies on the estimation of the head orientation of people, from

which the VFOA is deduced. We obtained an average error of around 10 degrees in pan angle, and 18 degrees in tilt angle in pose estimation, with fluctuations due to variation in people's appearence. With respect to VFOA recognition, the obtained results are encouraging, but additional work is needed. A first direction is the use of individualized VFOA models obtained through unsupervised adaptation. Early results along this line exhibit an absolute increase of performance of around 8%. The second research line addresses the ambiguity issues by modeling the interaction between people and different cues (e.g. speaking status, slide activity).

References

1. S. O. Ba and J. M. Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM-ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy*, pages 9–16, 2005.
2. L. Brown and Y. Tian. A study of coarse head pose estimation. *IEEE Workshop on Motion and Video Computing*, Dec 2002.
3. T. Cootes and P. Kittipanya-ngam. Comparing variations on the active appearance model algorithm. *BMVC*, 2002.
4. A. Doucet, S. Godsill, and C. andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 2000.
5. L. Lu, Z. Zhang, H. Shum, Z. Liu, and H. Chen. Model and exemplar-based robust head pose tracking under occlusion and varying expression. *CVPR*, Dec 2001.
6. J. McGrath. Groups: Interaction and performance. *Prentice-Hall*, 1984.
7. J.-M. Odobez. Focus of attention coding guidelines. IDIAP-COM 2 , Jan. 2006.
8. K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. ICMI*, Trento, Italy, Oct. 2005.
9. K. Parker. Speaking turns in small group interaction: a context sensitive event sequence model. *Journal of Personality and Social Psychology*, 1988.
10. R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Trans. on Neural Network*, March 1998.
11. R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. *Conf. on Human Factors in Computing Systems, Minneapolis, Minnesota, USA*, 2002.
12. K. Toyama and A. Blake. Probabilistic tracking in metric space. *ICCV*, Dec 2001.
13. A. Waibel, M. Bett, F. Metze, K. Ries, T.Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. *Proc. ICASSP*, May 2001.
14. P. Wang and Q. Ji. Multi-view face tracking with factorial and switching hmm. *WACV/MOTION'05 Workshops, Breckenridge, Colorado*, 2005.
15. Y. Wu and K. Toyama. Wide range illumination insensitive head orientation estimation. *IEEE Conf. on Automatic Face and Gesture Recognition*, Apr 2001.
16. J. Yang, W. Lu, and A. Weibel. Skin color modeling and adaptation. *ACCV*, 1998.
17. L. Zhao, G. Pingali, and I. Carlbom. Real-time head orientation estimation using neural networks. *Proc. of ICIP*, Sept 2002.