IDIAP RESEARCH REPORT

Visual Focus of Attention Estimation from Head Pose Posterior Probability Distributions

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

IDIAP–RR 07-75

12-2007

SUBMITTED FOR PUBLICATION

[a] {Sileye.Ba,odobez}@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 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 07-75

Visual Focus of Attention Estimation from Head Pose Posterior Probability Distributions

Sileye O. Ba and Jean-Marc Odobez

12-2007

submitted for publication

Contents

1 Abstract

We address the problem of recognizing the visual focus of attention (VFOA) of meeting participants from their head pose and contextual cues. The main contribution of the paper is the use of a head pose posterior distribution as a representation of the head pose information contained in the image data. This posterior encodes the probabilities of the different head poses given the image data, and constitute therefore a richer representation of the data than the mean or the mode of this distribution, as done in all previous work. These observations are exploited in a joint interaction model of all meeting participants pose observations, VFOAs, speaking status and of environmental contextual cues. Numerical experiments on a public database of 4 meetings of 22min on average show that this change of representation allows for a 5.4% gain with respect to the standard approach using head pose as observation.

2 Introduction

Analyzing and understanding human-human interaction is one the main aim of social sciences. While in the past such analysis was relying on tedious manual annotations of few data, it is nowadays possible to study and model in a more systematic fashion these interactions through the instrumentation of rooms with microphones and videos. In particular, meetings are places where, in more or less formal ways, people are sharing ideas and information, discussing, taking decisions, allowing for a large range of human interactions to be expressed. These interactions occur through verbal or non verbal means. Among the latter ones, gaze, which defines a person's visual focus of attention (VFOA), conveys important information for understanding the ongoing interactions between people, as gaze is used to manifest interest or to regulate the discourse [3].

Estimating gaze is however a very difficult task, as it requires the tracking of pupils' motion within the eyes. Since this proves to be infeasible without high-resolution images, investigation has been conducted to recognize people's VFOA from their head pose and other cues [7, 5, 4]. To make the problem tractable, it has been assumed that people are mainly interested in a small set of focus (named the VFOA targets), a valid assumption in practice. Various meeting cases have been considered to study VFOA recognition. In the first case, the VFOA targets set is composed of the other meeting participants [7, 5]. In the second case, a larger set is considered that includes, in addition to the meeting participants, the projection screen and the table to study more general meeting situations [4].
However, increasing the number of VFOA targets makes the problem more challenging, as there is a high probability that similar head orientations are used to gaze at different focus targets. The resolution of these head pose ambiguity cases can be done by modeling the relationship between people's VFOAs to their speaking status, or other contextual cues related to the group activity [7, 5, 2].

Nevertheless, one of the main source of error when inferring the VFOA from videos is the uncertainty in estimating people's head poses. In [4], it was shown that using a state-of-the-art vision-based pose estimator rather than the pose measurements obtained from a magnetic sensor was leading to performance decrease of more than 15%. Hence, improvement in VFOA recognition be achieved with a better head pose estimation, and more generally, with a better modeling of the relationship between the head pose measurements and the VFOA targets that accounts for uncertainties in pose estimation.

In this paper, we address the VFOA recognition problem from head pose information. Rather than relying on an estimated head poses defined by a pan and tilt angle, as done in all previous studies on the topic [7, 5], we propose to rely on the posterior probability density function (pdf) of the different head poses given the data to represent the head pose information embedded in the image data. In this way, we obtain a richer representation than the mean or the dominant mode of this distribution which allows to better model the likelihood of the image data for a given VFOA target, and take into account measurement uncertainties. Numerical experiments on a significant and challenging database demonstrate the validity of this approach.

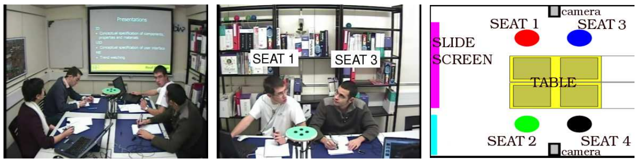


Figure 1: Evaluation setup. Central image: example of input video image. Seat numbers will be used to report VFOA results.

The remainder of this paper is organized as follows. Section 3 describes the task as well as the data used for evaluation. Section 4 presents the head pose algorithm, with an emphasis on the method for estimating the pose posterior pdf. Section 5 describes the architecture of the joint model of people VFOA, people head pose, people speaking status, and contextual cues. Section 6 presents experiments, and Section 7 gives conclusions.

3 Task And Dataset

Task: Our objective is to estimate people VFOA in meetings, and we assume that a person's VFOA can be any element of a finite set of visual targets that the person considers as interesting. In the scenario of our study, four people with different roles (meet around a table to discuss the design and creation of a new remote control. They take notes, use laptops, and display slides on a screen during presentations (see Fig. 1). Thus, the VFOA target set for a given participant seated at seat k comprises 6 VFOA targets: the 3 other participants, as well as the table, the slide screen, and an unfocused VFOA target. The later target (unfocused) is used when the person is not visually focusing on any of the previously cited targets (this case only represents 2% of our data).

Dataset description and analysis: Our dataset consists of 4 meetings of the AMI corpus[1], involving 4 people with real behaviors, according to the scenario description made above. Meeting duration ranged from 15min to 27min, for a total of 1h30min. Twelve different people participated in the meetings making the head pose tracking task challenging.

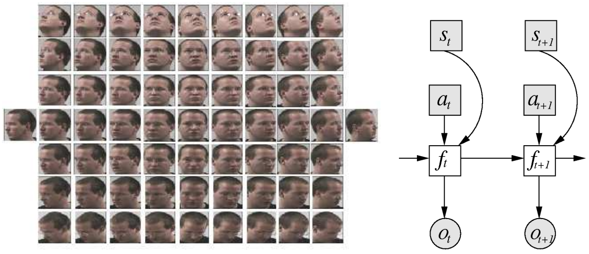
The meeting participants' VFOA were annotated based on the set of VFOA labels defined above. In this challenging scenario, in average, meeting participants looked at other people only 45% of the time, while looking at the table or at the slide screen represent respectively 30.8% and 21% of the data. These statistics are important as some targets are more difficult to recognize than others, and this will have effects on the overall performance. This is the case of looking at the table, which corresponds to two main situations: i) when people use their laptop, or ii) when people look downwards without actually changing their head pose, while still listening to a speaker, or while tending to disengage from the meeting.

4 Head Pose Tracking

To estimate the head pose, we used the computer vision tracker described in [1]. It relies on the Bayesian formulation of the tracking problem. Denoting the object configuration state at time t by X_t and the observations by Y_t, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of the state given the observation sequence $Y_{1:t} = (Y_1, \ldots, Y_t)$. In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters (PF), which consists of representing the filtering distribution using a set of N_s weighted samples (particles) $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrives. In [1], we applied such a framework to the joint tracking of the head and of its head pose.

More precisely, the state space contains both continuous variables L_t and a discrete variable θ_t. L_t

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



(a) (b)

Figure 2: Fig. 2(a):Discretized head pose space Θ. Fig. 2(b) IOHMM VFOA graphical model. Squares represent discrete variables, circles represent continuous variables. Unshaded variables are hidden, and shaded variables are observed.

represents the head location, vertical and horizontal scales, and an in-plane rotation that allows to localize the head in the image. The discrete index $\theta_t \in \Theta$ denotes an element of the discretized set of possible out-of-plane head poses shown in Fig. 2(a). As image observations, we used texture (output of Gaussian and Gabor filters) and skin color features at locations sampled from image patches extracted from the image and preprocessed by histogram equalization. For each element of discrete pose space Θ, a texture and color appearance model was learned from the Prima-Pointing database (www-prima.inrialpes.fr/Pointing04). These models were used to compute the likelihood of the observation given the state values.

One specificity of the approach in [1] was to use a Rao-Blackwellization approach to increase the sampling efficient, which results in a reduction of the number of samples for similar tracking performance. The Rao-Blackwellized particle filter (RBPF) consists of applying the standard PF algorithm to the continuous variables L while applying an exact filtering step over the exemplar variable θ, *given a sample of the tracking variables*. In this way, the likelihood of the state can be written as:

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

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

$$\{L^i_{1:t}, \pi^i_t, w^i_t\}^{N_s}_{i=1} \tag{2}$$

where $\pi^i_t(\theta) = p(\theta|L^i_{1:t}, Y_{1:t})$ is the pdf of the pose exemplars $\theta \in \Theta$ given a particle and a sequence of measurements, and $w^i_t \propto p(L^i_{1:t}|Y_{1:t})$ is the weight of the particle estimated through the PF approach.

In previous approaches, we were extracting the mean or the mode of this distribution. Here we propose to keep the whole distribution of head poses given the data as a representation of the image head pose information. It can be computed according to:

$$\begin{aligned}\pi_t(\theta) \quad &= \quad p(\theta|Y_{1:t}) = \int p(\theta, L_{1:t}|Y_{1:t}) &(3)\\ &= \quad \int p(\theta|L_{1:t}, Y_{1:t})p(L_{1:t}|Y_{1:t}) = \sum_{i=1}^{N_s} w^i_t \pi^i_t(\theta) &(4)\end{aligned}$$

This posterior pdf feature vector can then be used in the VFOA recognition model, as shown in the next Section.

5 VFOA Modeling with Hidden Markov Model

We developed an input-output hidden Markov model (IOHMM) that has as hidden state the VFOA of the meeting participants, as input variables meeting contextual cues, and as observation the meeting participants head pose pdf. The graphical model in Fig. 2(b) displays the relationship between our variables. Below we describe the main characteristics of our model.

5.1 Multi-Person VFOA Modeling with a HMM

The hidden state we are trying to estimate is $f_t = (f_t^1, f_t^2, f_t^3, f_t^4)$, the joint focus state of all participants (f_t^k denotes the VFOA of participant k at time t), which corresponds to all possible combinations of focus of the meeting participants. In addition to the head pose pdf of all participants ($o_t = (o_t^1, o_t^2, o_t^3, o_t^4)$), the observations comprise i) a slide-screen activity a_t variable, and ii) the speaking status of all participant $s_t = (s_t^1, s_t^2, s_t^3, s_t^4)$). In the HMM framework, estimating the multi-person VFOA can be posed as the maximization of the posterior probability density function (pdf) of the hidden states given the observations [6] which, according to the graphical model in Fig. 2(b), can be written as:

$$p(f_{1:T}|o_{1:T}, s_{1:T}, a_{1:T}) \propto p(f_0) \prod_{t=1}^{T} p(o_t|f_t) p(f_t|f_{t-1}, s_t, a_t) \tag{5}$$

This pdf is defined by the initial VFOA state distribution $p(f_0)$ (assumed to be uniform), the observation model $p(o_t|f_t)$ modeling the probability to measure an observation about people's head given their focus, and the state dynamic $p(f_t|f_{t-1}, s_t, a_t)$ modeling the probability of a group VFOA state given the past group VFOA state and the meeting context. We present below the observation models and state dynamics.

5.2 Observation Models

Assuming that given the VFOA state, people head poses information are independent of each other, the observation model can be factorized as $p(o_t|f_t) = \prod_{k=1}^{4} p(o_t^k|f_t^k)$. The individual terms were the modeled depending on the pose information exploited.

Using the estimated head pose: In this case, $o_t^k = (\alpha_t, \beta_t)$ is a head pose represented by a head pan angle α_t and a head tilt angle β_t. When VFOA is estimated from head pose, the cognitive model, presented [4], that relates people's gaze direction to their head pose can be used to predict the head pose corresponding to gazing at a given focus target. The cognitive VFOA model assumes that for a person k, the head pose θ corresponding to gazing at a target j can be modeled as a Gaussian distribution $\mathcal{N}(\theta, \mu_k^j, \Sigma_k^j)$ where the head pose center μ_k^j relates to the gaze direction $\mu_{k,j}^{gaze}$ through linear relation:

$$\mu_k^j = \kappa \mu_{k,j}^{gaze} \tag{6}$$

where κ the proportion of gaze rotation that can be attributed to the head rotation, and Σ_k^j represents the uncertainty in the head pose for person k gazing at target j depending on the target physical size and the distance between the observer and target. $p(o_t^k|f_t^k = \text{ unfocused }) = u$ is modelled as a uniform distribution. The choice of a Gaussian distribution to model the class conditional distributions modeling the head pose observation allows the use of an unsupervised MAP adaptation framework to adapt the observation model to input test data as presented in [4]. This property is used for the method estimating VFOA from head pose that we consider as our baseline model.

Using the head pose posterior distribution: In this case, we have $o_t^k = \pi_t^k$, and the VFOA conditional likelihood are modeled as exponential distributions, i.e.

$$p(\pi_t^k|f_t^k = j) = \lambda \exp -\lambda \rho(\pi_t^k, \Pi_k^j)^2 \tag{7}$$

where ρ is a distance on distribution, and Π_k^j is the distribution over the head poses representing people focusing at the target j. Assuming the Gaussian modeling presented above, this representative distribution Π_k^j is defined as:

$$\Pi_k^j(\theta = l) = \frac{\mathcal{N}(\theta = l, \mu_k^j, \Sigma_k^j)}{\sum_{l'} \mathcal{N}(l', \mu_k^j, \Sigma_k^j)} \tag{8}$$

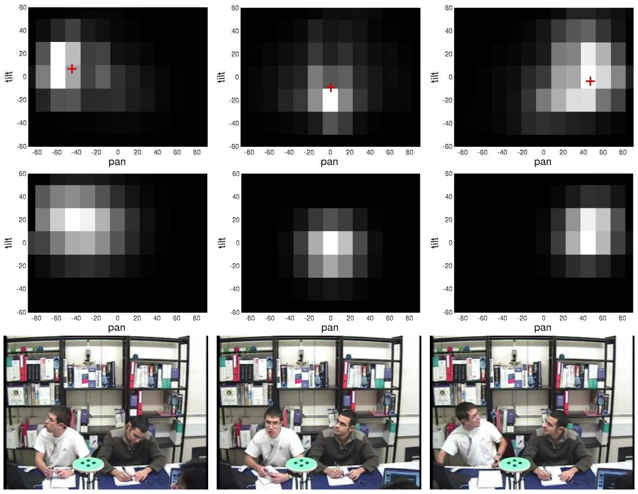


Figure 3: Head pose pdf observation and models. First row: head pose pdf obsevration and corresponding estimated head pose (red plus) for a person sitting a seat 1 and gazing at the slide screen (first column), at the person at seat 2 (second column), and seat 3 (3rd column). Note that the head pose pdfs are ordered as the discrete pose in Fig. 2(a). Second row: head pose pdf model for the seat 1 and third row typical image of a person sitting at seat 1 gazing at the slide screen, at the person at seat 2 and at the person at seat 3.

Fig. 3 shows illustrations of the head pose pdf models for a person sitting at the seat 1 focusing at the slide screen, the person sitting at seat 2, and the person sitting at seat 3. Note here that, since we do not perform adaptation, we do not need to rely on Gaussian distributions. Thus we could use more appropriate representative distributions, such as flatter distributions to better take into account the gaze spread of table or projection screen VFOA targets. For the 'Unfocused' VFOA label, we used a uniform distribution $\Pi_k^j(\theta) = u$.

5.3 State Dynamics

We define the state dynamics as follows:

$$p(f_t|f_{t-1}, a_t, s_t) \propto \Phi(f_t)p(f_t|f_{t-1})p(f_t|a_t)p(f_t|s_t) \tag{9}$$

where $\Phi(f_t)$ is a distribution modeling the prior probability of observing a given multi-person VFOA pattern, $p(f_t|f_{t-1})$ models the temporal transitions between VFOA states, $p(f_t|a_t)$ models the probability to observe a joint VFOA state given the slide activity, and $p(f_t|s_t)$ models the probability to observe a joint VFOA state given the speaking activities.

The multi-person VFOA prior $\Phi(f_t)$: This prior models people's inclination to share VFOA targets. Following the idea that in meeting people share more focus than if their focus were assumed independent, we have set $\Phi(f_t)$ as:

$$\Phi(f_t) = \Phi(SF(f_t) = n) \propto \frac{d_n}{c_n} \tag{10}$$

where $SF(f_t)$ denotes the number of people that share the same focus in the joint state f_t, and d_n is the frequency count of people sharing n focus learned from training data and c_n is the frequency count of people sharing n focus if people's focus were assumed independent.

VFOA temporal transitions: The role of the VFOA temporal transition is to enforce temporal smoothness on the state sequence. We modeled this term assuming that the individual transition probabilities of the different persons are independent given their previous focus:

$$p(f_t|f_{t-1}) = \prod_{k=1}^{4} p(f_t^k|f_{t-1}^k). \tag{11}$$

Person position	seat 1	seat 2	seat 3	seat 4	mean
pose	51.6	55.3	42.3	44.1	48.3
pose pdf L2 distance	51.3	52.3	35.6	47.6	46.7
pose pf Batt. distance	56.6	58.3	48.7	50.1	52.9
pose pdf KL distance	61.1	55.9	48.5	48.5	53.6

Table 1: FRR recognition rates per seating position for .

The individual VFOA dynamics $p(f_t^k|f_{t-1}^k)$ is modeled as a transition table with a high probability to remain in the same state and the remaining of the probability uniformly spread on the other states.

Slide activity prior modeling: The slide variable a_t denotes the time that elapsed since the last slide change occurred. When the time elapsed since the last slide change a_t is small, it is more probable that people are looking at the slide screen than to other VFOA targets. We have modeled this term as:

$$p(f_t|a_t) \propto \prod_{k=1}^{4} p(f_t^k|a_t). \tag{12}$$

where we assumed that the individual person VFOA states were independently influenced by the slide activity variable. The prior of a person k observing the slide screen after a slide change is defined as:

$$p(f_t^k = \text{ slide screen } |a_t) = p_{ss} = \vartheta_1 e^{-\vartheta_2 a_t} + \vartheta_3$$

where $\{\vartheta_i\}_{i=1,2,3}$ are parameters learned from training data. The remaining of the probability mass of the slide change prior is uniformly spread among the other targets.

Speaking activity modeling: We model the speaking dependent term $p(f_t^k|s_t)$, following the idea that people in meetings are more likely to focus at speakers than non-speakers. Assuming that given the speaking status, people VFOA are independent, we have:

$$p(f_t|s_t) \propto \prod_{k=1}^{4} p(f_t^k|s_t) = \prod_{k=1}^{4} p(f_t^k|S_t^k) \tag{13}$$

where S_t^k denotes the set of speakers at time t which are not person k. The prior on the focus of person k given the speaker set $p(f_t^k|S_t^k)$ is learned from training data and favors people focusing at speakers.

6 Evaluation Setup and Experiments

Evaluation was conducted using the data described in Section 3, the frame recognition rate FRR (percentage of frame that are correctly classified) as a performance measure, and a leave one out protocol. More precisely, in turn, one meeting is left aside as test data, the remaining 3 meetings are used to train the models parameters.

We ran experiments to compare the effect of various distribution metrics ρ introduced in Eq. 7 to define the observation model, on the performance of VFOA recognition from head pose pdf. Experiments were also conducted to study the performance of VFOA recognition from people's head pose to VFOA recognition from people's head pose pdf. First we describe the effect of the distribution metric.

Distribution metrics: We considered three metrics to define the observation model in Eq. 7, the Euclidean distance applied to distribution, the Battacharya distance defined as

$$\rho(p,q) = \sqrt{1 - \sum_{\theta \in \Theta} \sqrt{p_\theta q_\theta}} \tag{14}$$

where p and q are two distribution on the pose space Θ and the Kullback-Liebler (KL) distance defined as

$$\rho(p, q) = \sum_{\theta \in \Theta} \left(p_\theta \log \frac{q_\theta}{p_\theta} + q_\theta \log \frac{p_\theta}{q_\theta} \right) \tag{15}$$

The second, third and fourth rows of Table 1 give the performance when using the three metrics. As can be seen in this table, for all the metrics, the recognition performances for seat 1 and 2 are better than the performances for seat 3 and 4. This is explained by the geometric configuration of the room. Because of the meeting room settings, the persons sitting at seats 3 and 4 have their head poses more subject to ambiguities in defining their VFOAs as shown in [4, 2]. Table 1 also shows that the KL distance and the Battacharya distance are performing better than the Euclidean distance. This is expected since the KL distance and the Battacharya distance are better suited for comparing probability distributions. The Euclidean distance is sensitive to outliers and tends to be unreliable in high dimension spaces.

Head pose versus head pose pdf: Table 1 gives also the performances for recognizing people's focus using their head pose instead of their head pose pdf. While when using the head pose as observation, the average performance is about 48.3%, when using the head pose pdf with a KL divergence the performance is about 53.6%. Using head pose pdfs leads to an absolute improvements of 5.4%. For each seat, the methods based on the head pose pdf with KL or Battacharya distance are consistently better than the method estimating VFOA from head pose. This shows that head pose pdf allows a more robust representation of people's head pose.

7 Conclusion

In this paper a method to jointly estimate the VFOA of a set of meeting participants from head pose pdf using meeting contextual cues was presented. To the best of our knowledge this is the first work using head pose pdf to estimate people's focus. Significant improvements were achieved with respect to a baseline model estimating people VFOA from their head poses. Our future research directions will be about a joint inference of the head pose pdf models and the parameters of the state dynamics. Investigations will also be conducted in order to introduce priors on higher level meeting contexts to account for the dependency of people's focus to the meeting events such as monologues, dialogs, and group discussions.

8 Acknowledgements

This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the European union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811). This research was also funded by the U.S. Government VACE program. The authors also thank Dr. Daniel gatica Perez, Dr. Hayley Hung, and Dinesh Jayagopi from IDIAP Research Institute for their helpfull comments.

References

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

[2] S. O. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proc. of ICASSP*, 2008.

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

[4] J-M. Odobez and S.O. Ba. A cognitive and unsupervised MAP adaptation approach to the recognition of focus of attention from head pose. In *Proc. of ICME*, 2007.

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

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

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