Multi-party Focus of Attention Recognition in Meetings from Head Pose and Multimodal Contextual Cues

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

IDIAP–RR 07-50

May 2008

published in
IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)

[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 07-50

Multi-party Focus of Attention Recognition in Meetings from Head Pose and Multimodal Contextual Cues

Sileye O. Ba and Jean-Marc Odobez

May 2008

published in
IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)

Contents

1 Abstract

This paper presents investigations on visual focus of attention (VFOA) recognition in meetings from audio-visual perceptual cues. Rather than independently recognizing the VFOA of each participant from his own head pose, we propose to recognize participants' VFOA jointly in order to introduce context dependent interaction models that relates to group activity and the social dynamics of communication. To this end, we designed an input-output hidden Markov model (IOHMM), whose hidden states are the joint VFOA of all participants, and whose main observations are the head poses. Interaction models are introduced in the form of contextual cues that affect the temporal evolution of the joint VFOA sequence, allowing us to model group dynamics that accounts for people's tendency to share the same focus, or to have their VFOA driven by contextual cues such as slide activity or the participant speaking activity. The model is rigorously evaluated on a publicly available dataset of 4 real meetings of 23min on average, showing an overall 10% relative performance increase w.r.t. the independent recognition case.

2 Introduction

Meetings constitute an essential part of our working daily life, and due to world globalisation, remote meetings will be more and more frequent. In this domain, most of computer science research has focused on developing tools that support content management or transmission between remote meetings sites. Not much has been done on the automatic analysis of the meeting social dynamics, although studies have shown that real-time feedback of speaking or gaze activity statistics can positively affect participant behavior, improve group cohesiveness and participant satisfaction, which can lead to higher meeting efficiency [2]. In another example, when somebody is engaged in a distant meeting with a group of co-located people through an audio connection, he does not perceive the non-verbal communication signals indicating the reactions of meeting participants to propositions and comments. This lack of perception often leads to interrupting at the wrong time, or not answering promptly, and ultimately leads to a disengagement of the meeting. In both examples, gaze, which defines the VFOA of meeting participants, plays an important role as it is a major cue involved in determining the addressee or finding good moments to take speaking turns. Providing gaze and VFOA information could thus be useful to increase social awareness in meetings.

In this paper, we address the problem of recognizing the VFOA (which is defined by the eye gaze) from head pose and multi-modal contextual cues. Previous work in this domain have shown that in 4 persons short meeting situations where the VFOA targets of interest are only the other meeting participants, people's VFOA can be reliably recognized from their head pose [7, 5]. In previous work we conducted where other VFOA targets of interest also comprises a slide screen or the table, we showed that estimating VFOA only from head poses is a very challenging task [3], since the head pose can be ambiguous in determining the VFOA, i.e. due to the fact that we have no access on the eye orientation within the head, a given head pose can correspond to several VFOA targets. To reduce this ambiguity, we can perform the recognition of all participants VFOA jointly and user other cues such as speech. For instance, in [7], speech and head pose probabilities are combined in a linear fashion to recognize the VFOA, but no attempt was made to model the joint VFOA interactions. In [4], an interesting approach modeling conversational event interaction was proposed, but (like in [7]), only VFOA targets corresponding to meeting participants were considered, and their scenario did not consider other contextual cues related to group activity such as looking at slides, as done is this paper. In this paper, we propose to perform the recognition of all participants VFOA jointly by the use of interaction models that exploit contextual cues relating to group activity or group communication properties. A first property is that people naturally share VFOA targets, which means that for an external observer, knowing where 3 participants are looking at provide some hints on what or whom the 4th participant is looking at. A second property is that contextual information (e.g. is there a new slide ? who is speaking ?), which are unrelated to head pose, usually drive the attention of

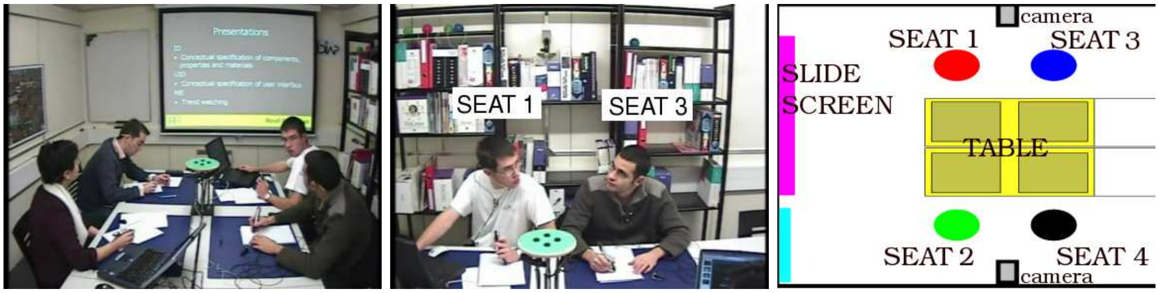

Figure 1: Evaluation data recording setup. Seat numbers will be used to report results for the VFOA recognition of people seated at these locations.

participants, and using these properties should thus provides hints on where people are looking. To account for these properties, we designed an IOHMM whose hidden states are the joint VFOA of all participants, and the main observation are people's head poses. The interaction models are introduced through the dynamic model of the joint VFOA states. This comprises a group prior which models person's tendency to share the same focus, and is influenced by priors driven by contextual cues in the meeting. This model was evaluated on a public database, significantly larger than in previous works (1h30min of meeting), and showed that the joint modeling helps to remove some of the head pose ambiguities.

The paper is organized as follows. Section 3 defines the addressed task and describes our evaluation data. Section 4 details the model we propose. Section 5 presents the evaluation protocol and the results. Section 6 concludes the paper.

3 Task and Dataset

Task: Our objective is to estimate the VFOA of people in a meeting scenario. A person's VFOA can be any element of a finite set of visual targets in the environment that the person considers as interesting. In the scenario of our study, four people with different roles (project manager, marketing expert,...) have a meeting around a table to discuss the design and creation of new remote control. They take notes, use laptops, and display slides on a screen during presentations (see Fig. 1). Thus, the set of interesting visual targets for a given participant seated at seat k, denoted $\mathcal{F}_k$, comprises 6 VFOA targets: the 3 other participants $\mathcal{P}_k$ (e.g. for seat 1, $\mathcal{P}_1 =$\{seat2,seat3,seat4\}), as well other targets $\mathcal{O} =$\{Table, Slide Screen, Unfocused \}. The later target (Unfocused) is used when the person is not visually focusing on any of the previously cited targets.

Dataset description and analysis: The dataset used for our study consisted in 4 meetings of the AMI corpus[1], involving 4 people with real behaviors, according to the scenario description made above. The duration meetings ranged from 15min to 27min, for a total of 1h30min. Twelve different people were involved 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. Table 1 gives the VFOA statistics, where we have grouped the VFOA labels corresponding to participants into a single label 'people'. Looking at people only represents 45% of the data, while looking at table or slide represents a significant proportion of the VFOA. The label 'Table' corresponds to two main situations i) when people use their laptop ii) when people look downwards without actually changing their head pose while listening to a speaker. In particular, situation ii) has been found to have increased w.r.t. our previous study on 7 to 10min long meetings [3]. These VFOA statistics contrast with other setups and places our work in a different context than studies investigating VFOA

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

VFOA	people	table	slide	unfocused
Percentage of data	44.9	30.8	21.5	2.5

Table 1: Distribution of VFOA labels.

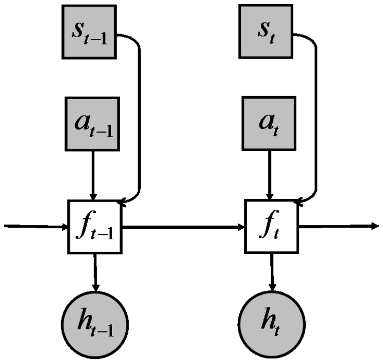

Figure 2: IOHMM VFOA graphical model. Squares represent discrete variables, circles represent continuous variables. Unshaded variables are hidden, and shaded variables are observed.

estimation when the targets are contrived to be only other meeting participants [7, 4]. Indeed some targets are more difficult to recognize than others, and this will have effects on the performance. This is the case for the label 'Table' due to the situation ii) described above, and to the fact that, in contrast to [7, 4], we can no longer rely only on the head pan, but also need to use the head tilt -which is known to be more difficult to estimate from images- to distinguish different VFOA targets.

4 Multi-Party VFOA modeling Using Head Pose and Contextual Cues

To account for group dynamics and contextual information, we have developed an IOHMM model whose graphical model is displayed in Fig. 2. Its main characteristics are described below.

4.1 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 of all participants ($h_t = (h_t^1, h_t^2, h_t^3, h_t^4)$), 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, can be written as:

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

This pdf is defined by the initial VFOA state distribution $p(f_0)$ (assumed to be uniform), the observation model $p(h_t|f_t)$ modeling the probability to observe people's head pose given their VFOA, 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

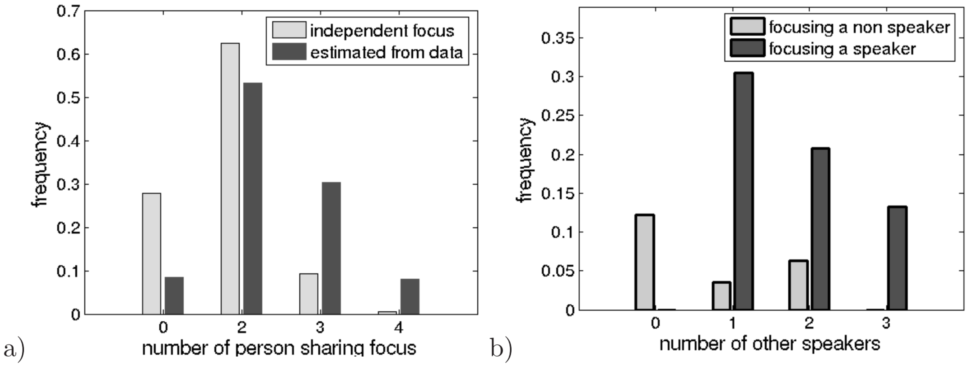

Figure 3: a) Shared focus. Distribution of frames where n persons (and no more) are focused on the same VFOA target. Light bars: distribution c_n assuming people VFOA are independent. Dark bars: distribution d_n measured on the data. b) Probability for a participant k to focus on a speaking (dark grey) or a non speaking (light grey) participant, given the number of other speakers ($|S^k|$).

group VFOA state and the meeting context. We present below the state dynamic and observation models.

4.2 State Dynamics

We define the state dynamics as follows:

$$p(f_t|f_{t-1}, a_t, s_t) \propto \Phi(f_t)p(f_t|f_{t-1})p(f_t|a_t)p(f_t|s_t) \tag{2}$$

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 factorization made in Eq. 2 is based on the assumption that the group prior $\Phi(f_t)$ models all the dependencies between the VFOA of meeting participants, while the other terms only model the effect of the conditional variable on the current focus.

The multi-person VFOA prior $\Phi(f_t)$: This prior models people's inclination to share VFOA targets. Fig. 3a) depicts the distribution of frames w.r.t. the number of people that share the same focus. As can be seen, people are sharing more often the same focus than if they would behave independently. Thus, we have set $\Phi(f_t)$ as:

$$\Phi(f_t) = \Phi(SF(f_t) = n) \propto \frac{d_n}{c_n} \tag{3}$$

where $SF(f_t)$ denotes the number of people that share the same focus in the joint state f_t, and d_n and c_n are defined in Fig. 3. Qualitatively, this term will favor states with shared focus according to the distribution observed on training data.

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:

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

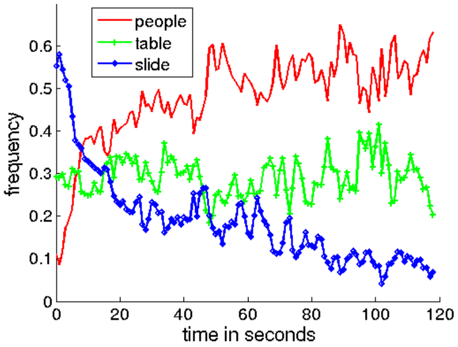

Figure 4: Empirical probability of focusing to another meeting participant, to the table, or to the slide screen, in function of the time that elapsed since the last slide change.

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 distributed on the other states, in order to avoid putting prior knowledge in this term except the smoothness.

Slide activity prior modeling: The slide variable a_t denotes the time that elapsed since the last slide change occurred. Slide change detection is done automatically through the processing of the view in Fig. 1, left). As illustrated in Fig 4 when 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:

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

where we assumed that the the individual person VFOA states were independently influenced by the slide activity variable. The probability $p(f_t^k|a_t)$ of observing a particular focus for person k given a_t is defined as:

$$\begin{array}{rcl} p(f_t^k = \text{ slide screen } |a_t) & = & p_{ss} = \alpha_1 e^{-\alpha_2 a_t} + \alpha_3 \\ p(f_t^k = \text{ other target } |a_t) & = & \dfrac{1-p_{ss}}{|\mathcal{F}_k|-1} \end{array} \tag{6}$$

where $|.|$ denotes the cardinality operator, and $\{\alpha_i\}_{i=1,2,3}$ are parameters learned from fitting the probability of focusing on the slide screen given a_t (see Fig. 4). Although Fig. 4 shows an interesting trend in the probability of focusing on people w.r.t. a_t, Eq. 6 assumed an equal probability amongst the targets different than the slide screen.

Speaking activity modeling: It is well known in social sciences that people in meetings are more likely to look at speakers than at non-speaker. This is illustrated in Fig. 3b). We followed this idea to model the speaking dependent term $p(f_t^k|s_t)$. Assuming that given the speaking status, people VFOA are independent, we have:

$$p(f_t|s_t) \propto \prod_{k=1}^{4} p(f_t^k|s_t) = \prod_{k=1}^{4} p(f_t^k|S_t^k) \tag{7}$$

where S_t^k denotes the set of speakers at time t which are not person k, and we further assumed that the VFOA of person k is independent of whether k speaks or not. To model $p(f_t^k|S_t^k)$, there are four cases, depending on the size of the set S_t^k. We assumed that the probability of focusing on an object is constant, independent of $|S_t^k|$, and denotes by p_o. It was equally divided amongst the object, i.e. $p(f_t^k = l, l \in \mathcal{O}|S_t^k) = \frac{p_o}{|\mathcal{O}|} = \frac{p_o}{3}$. Then, $1-p_o$ represents the probability of focusing on a person, which was divided amongst speakers and non-speakers according to: $p(f_t^k = l, l \in S_t^k|S_t^k) \propto \gamma$ and

Person position	seat 1	seat 2	seat 3	seat 4	mean
Independent	52.5	50.5	27.3	39.5	42.4
Group	50.4	51.1	32.4	43.3	44.3
Group+slide	48.9	51.3	35.3	45.3	45.2
Group+speech	51.3	52.3	33	45	45.4
Group+slide+speech	51.3	52.3	35.6	47.6	46.7

Table 2: FRR recognition rates per seating position.

$p(f_t^k = l, l \notin S_t^k | S_t^k) \propto \bar{\gamma}$, with $\gamma > \bar{\gamma}$. The value of p_o and of $\frac{\gamma}{\bar{\gamma}}$ are learned from training data. According to Table 1, p_o will be near 55% on average, and from Fig. 3b, $\frac{\gamma}{\bar{\gamma}}$ lie between 5 and 8.

4.3 Observation Models

The observations consist of the head poses automatically extracted using the computer vision tracker described in [1]. The tracker relies on a Bayesian approach to jointly estimate the head location and pose using head pose appearance models learned from an external database (www-prima.inrialpes.fr/Pointing04). For a person k, the pose h_t^k consists of the estimated pan and tilt pose values.

Head pose observation model: Assuming that given the VFOA state, people head poses are independent of each other, the observation model can factorize as $p(h_t|f_t) = \prod_{k=1}^4 p(h_t^k|f_t^k)$. Then, for regular VFOA label, the individual conditional probabilities distribution were modeled as Gaussian distribution, i.e. $p(h_t^k|f_t^k = j) = \mathcal{N}(h_t^k; \mu_k^j, \Sigma_k^j)$ where μ_k^j and Σ_k^j resp. denote the mean and covariance of the distribution. For the unfocused VFOA label, $p(h_t^k|f_t^k = \text{ unfocused }) = u$ is modelled as a uniform distribution.

Unsupervised observation model adaptation: The observation model parameters (μ_k^j, Σ_k^j) can be learned from training data. However, as people have personal ways at looking at specific targets, and since the tracker can introduce systematic bias to the estimated head pose of different person, the learned parameters might not be suited for new people. To address this issue, we applied the unsupervised Maximum a posteriori (MAP) adaptation method to produce in an unsupervised fashion (i.e. only the head observations are used, without any VFOA label information) head pose observation models that compensates for people's personal characteristics [3]. Adaptation was conducted separately for each individual, and the same adapted models were used in all experiments.

5 Experimental Results

Experimental setup: 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. Five models are compared, depending on the terms that are used in Eq. 2. The first model (called *independent*) relies only on the temporal smoothing term, thus implicitly assuming an independent recognition of each person VFOA from the head pose. The second model exploits in addition the group prior $\Phi(f_t)$. The $3rd$ and $4th$ model adds to the $2nd$ model the slide (resp. speaking) contextual model. The $5th$ model is the complete model.

Results: Table 2 summarizes the results for the different models. First of all, we can notice that the recognition rates are not very high, which provides an idea of the task difficulty. As could be expected, VFOA recognition in seat 1 & 2 is better than in seats 3 & 4, as the latter have higher VFOA ambiguities in the head pose space. For instance, in image at center in Fig.1, the head pose of the person at seat 3 can correspond to looking at the slide screen or at seat 1. Comparing the results of the different models, we can notice that enforcing joint focus in the recognition process is already

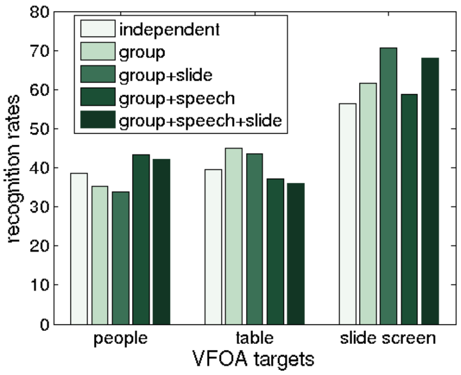

Figure 5: VFOA Recognition rates for the people, table, and slide screen VFOA classes.

bringing around 2% improvement over the independent modeling. When adding one contextual cue, the results further increase by one percent approximately. When everything is combined, we obtain a final increase of 4.3% over the independent modeling, reaching a 46.7% recognition rate. Interestingly enough, we observe that seats 3 and 4 beneficiate the most from adding the group and contextual cues with around 8% of performance increase. This was to be expected since these seats are the ones with the most ambiguous head pose space. Fig. 5 shows the average VFOA recognition for the people VFOA targets, the table target, and the slide screen target. It illustrates the benefit of using the group dynamic and contextual priors. Adding the group priors improves the recognition of VFOA targets (table, slide screen) related to group behaviors. The slide screen context improves the slide screen recognition and the speaking context leads to improvements in recognizing the different people as VFOA targets.

6 Conclusion

This paper presented an IOHMM model for the joint multi-party VFOA recognition from head pose and multi-modal contextual cues. The conducted experiments showed that the proposed approach performed significantly better than when an independent recognition of people's VFOA using only head pose is done. In particular, the we proposed model beneficiated mostly to the VFOA recognition of people seating at places where the head pose alone is ambiguous to determine the VFOA.

Future research directions include adding table activities such as writing or using laptop in the contextual cues, as well as investigating the joint unsupervised adaptation of the group VFOA head pose observation model instead of the independent one currently done.

7 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 Workshop on Multimodal Multiparty Meeting Processing (MMMP)*, pages 9–16, 2005.

[2] O. Kulyk, J. Wang, and J. Terken. Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In *Proc. of the MLMI workshop*, volume 3869 of *LNCS*, pages 150–161, 2006.

[3] 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 Int. Conf. on Multi-media & Expo*, 2007.

[4] 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. of Int. Conf. on Multimodal Interfaces*, pages 191–198, 2005.

[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 Int. Conf. on Multi-media & Expo*, 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.