IDIAP RESEARCH REPORT

A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of the Focus of Attention from Head Pose

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

IDIAP–RR 07-20

January 2007

to appear in
International Conference on Multimodal & Expo

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne

IDIAP Research Report 07-20

A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of the Focus of Attention from Head Pose

Jean-Marc Odobez Sileye O. Ba

january 2007

to appear in
International Conference on Multimodal & Expo

Abstract. In this paper, the recognition of the visual focus of attention (VFOA) of meeting participants (as defined by their eye gaze direction) from their head pose is addressed. To this end, the head pose observations are modeled using an Hidden Markov Model (HMM) whose hidden states corresponds to the VFOA. The novelties are threefold. First, contrary to previous studies on the topic, in our set-up, the potential VFOA of a person is not restricted to other participants only, but includes environmental targets (a table and a projection screen), which increases the complexity of the task, with more VFOA targets spread in the pan and tilt (as well) gaze space. Second, the HMM parameters are set by exploiting results from the cognitive science on saccadic eye motion, which allows to predict what the head pose should be given an actual gaze target. Third, an unsupervised parameter adaptation step is proposed which accounts for the specific gazing behaviour of each participant. Using a publicly available corpus of 8 meetings featuring 4 persons, we analyze the above methods by evaluating, through objective performance measures, the recognition of the VFOA from head pose information obtained either using a magnetic sensor device or a vision based tracking system.

1 Introduction

The automatic analysis and understanding of human behavior constitutes a rich and interesting research field. One particular behaviour component of interest is the *gaze*, which indicates where and what a person is looking at, or, in other words, what the *visual focus of attention (VFOA)* of the person is. In many contexts, identifying the VFOA of a person conveys important information about that person. For instance, tracking the VFOA of people in a public space could be useful to measure the degree of attraction of a given focus target such as advertisements or shop displays [9]. In meeting contexts, the VFOA of people is an important non verbal communication cue with functions such as establishing relationship (through mutual gaze), regulating the course of interaction, expressing intimacy, and exercising social control [6]. Thus, recognizing the VFOA patterns of people or of group of people can reveal important knowledge about the participants' activity and role, and the conversation structure.

In this meeting context, the goal of this paper is to analyze the correspondence between the head pose of people and their gaze in more natural and realistic scenarios than those previously considered [10, 7], and propose methods to recognize the VFOA of people from their head pose. In contrast to previous approaches [10, 7], our scenario involves people looking at slides or writing on a sheet of paper on the table. As a consequence, people have more potential VFOA targets in our set-up (6 instead of 3 in the cited works), leading to more ambiguities between VFOA. And, due to the physical placement of the VFOA targets, the identification of the VFOA can only be done using the complete head pose representation (pan and tilt), instead of just the head pan as in [10, 7].
To recognize the VFOA of people, we investigate the use of an HMM model to segment pose observation sequences into VFOA temporal segments. The head pose observations are represented using VFOA conditional Gaussian distributions, whose means indicate the head pose associated with each VFOA target. In our previous work [1], these means were set using training data. However, as collecting training data can be tedious, in the current paper, we investigate the use of results of studies on saccadic eye motion modeling [3] and propose an approach (referred to as cognitive in the paper) that models the head pose of a person given his upper body pose and his effective gaze target. In this way, no training data are required to learn the VFOA parameters, but some knowledge of the 3D room geometry is necessary. In addition, to account for the fact that people have their own head pose preferences for looking at the same given target, we adopted an unsupervised Maximum A Posteriori (MAP) scheme to adapt all the HMM parameters to individual people. This departs from [10], where the HMM parameters are directly learned on the test data after k-means initialization, an approach that could only work due to their simpler setting, as commented above.

To evaluate our VFOA modeling, we conducted comparative and thorough experiments on a large publicly available database, comprising 8 meetings for which both the head pose ground-truth and VFOA label ground truth are known. Due to this feature, we can differentiate between the two main error sources in VFOA recognition: (1) the use of head pose as proxy for gaze, and (2) errors in head pose estimation.

This paper is organized as follows. Section 2 describes the task. Section 3 presents our video head pose tracker and its performance. Section 4 describes the VFOA model, along with the cognitive parameter setting and the unsupervised MAP framework used to adapt our VFOA model to unseen data. Section 5 presents our experimental setup and reports our results. Section 6 concludes the paper.

2 Task and Database

Our goal is to evaluate how well we can infer the VFOA state of a person using head pose in real meeting situations. An important issue is: what should be the definition of a person's VFOA state? At first thought, one can consider that each different gaze direction could correspond to a potential VFOA. However, studies on the VFOA in natural conditions [5] have shown that humans tend to

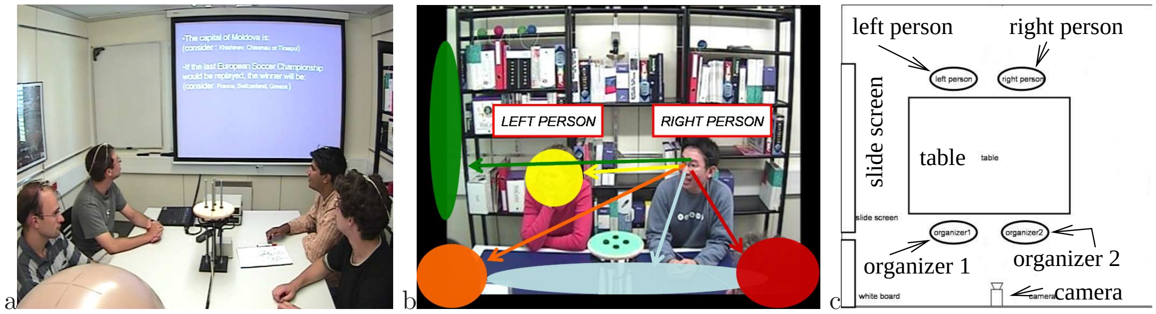

Figure 1: a) meeting room b) sample image and potential VFOA targets for the right person c) geometric configuration of the room.

look at targets which are either relevant to the task they are solving or of immediate interest to them. Additionally, one interprets another person's gaze not as continuous 3D spatial locations, but as gaze towards objects that one has identified as potential targets. This process, called the shared-attentional mechanism [6], suggests that in general VFOA states correspond to a finite set of targets of interests. Thus, in our meeting context set-up, the set of VFOA targets of interest, denoted $\mathcal{F}$, has been defined as: the other meeting participants (PR and PL stands for person right and left, $O1$ and $O2$ for organizer 1 and 2), the slide-screen SS, the table TB, and a label unfocused U when none of the previous could apply. As a result, for the 'person left' in Fig. 1c, we have: $\mathcal{F} = \{PR, O2, O1, SS, TB, U\}$.

As data, we rely on the IDIAP Head Pose Database (IHPD[1]), which was collected with head pose ground truth (GT) produced by a magnetic field head orientation sensor. Each participant's discrete VFOA was hand annotated as well, on the basis of its gaze direction. This allows us to evaluate the impact of using the estimated vs the true head pose as input to the VFOA recognition algorithms. The meeting durations ranged from 7 to 14 minutes. In shorter recordings (less than 2-3 minutes), we found that participants tend to overact with the effect of using to a greater extent their head to focus on other people/objects. In our meetings, the attention of participants sometimes drops and people may listen without focusing on the speaker, which produces more realistic meeting scenario.

3 Head Pose Tracking

Probabilistic Head Pose Tracker. To estimate the head pose, we used the computer vision tracker described in [2], which 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 [2], we applied such a framework to the joint tracking of the head and of its head pose. The state space contains both continuous variables (head location, scales, in-plane rotation) and a discrete variables l denoting an element of the discretized set of possible out-of-plane head poses. As 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 of the pose l, a texture and color 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 value. To increase the sampling performance of the PF, a Rao-

[1]Details and data available at *http://www.idiap.ch/HeadPoseDatabase/*

con-dition	right persons		left persons		pan near frontal ($\|\alpha\| < 45°$)		pan near profile ($\|\alpha\| > 45°$)	
stat	mean	med	mean	med	mean	med	mean	med
pan	11.4	8.9	14.9	11.3	11.6	9.5	16.9	14.7
tilt	19.8	19.4	18.6	17.1	19.7	18.9	17.5	17.5
roll	14	13.2	10.3	8.7	10.1	8.8	18.3	18.1

Table 1: Pan/tilt/roll error statistics (in °) for person left/right, and different configurations of the true head pose.

Blackwellization approach was used, which results in a reduction of the number of samples for similar tracking performance.
The Head Pose Tracking Evaluation was done on the IHPD database using a two-fold protocol (some of the dynamic parameters were learned on the data). No initial and manual individual registration was needed as in [7]. As Performance measures, we used the average and median errors in pan, tilt and roll angles (e.g. the average over time and meeting of the absolute difference between the pan of the GT and the tracker estimation). The statistics of the errors are shown in Table 1. Overall, given the small head size, and that the appearance training set is composed of faces recorded in an external set up (different people, different viewing and illumination conditions), the results are good. However these results hide a large discrepancy between individuals (e.g. the average pan error per individual ranges from 7° to 30°), which depends mainly on whether the tracked person's appearance is well represented by people in the appearance training set. Table 1 also shows that the pan and roll tracking errors are smaller than the tilt errors (tilt estimation is more sensitive to the quality of the face localization, as pointed out by other researchers) and details the errors depending on whether the true pose is near frontal or not. In near frontal position, the head pose tracking estimates are more accurate, in particular for the pan. This can be understood since for near profile poses, a variation in pan introduces much less appearance change than the same variation in a near frontal view.

4 Visual Focus of Attention Modeling

4.1 VFOA modeling with an HMM

Let $s_t \in \mathcal{F}$ denote the VFOA state, and z_t the head pointing direction of a person at time t as defined by the head pan and tilt angles, i.e. $z_t = (\alpha_t, \beta_t)$, since the head roll has no effect on the head direction by definition. Denoting also $s_{0:T}$ and $z_{1:T}$ the VFOA and observation sequence, respectively, the joint posterior probability density function of states and observations can be written:

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

In this equation, the emission probabilities $p(z_t|s_t = f_i)$ are modeled as Gaussian distribution $\mathcal{N}(z_t; \mu_i, \Sigma_i)$ with mean μ_i and full covariance matrix Σ_i for the $K-1$ VFOA which are not *unfocused*, and by a uniform distribution $p(z_t|s_t = U) = \frac{1}{180\times180}$ in the unfocused case. The state transitions $p(s_t = f_j|s_{t-1} = f_i) = a_{i,j}$ are represented by a transition matrix $A = (a_{i,j})_{i,j=1:K}$. Thus, the set of HMM parameters is $\lambda = \{\mu = (\mu_i)_{i=1:K-1}, \Sigma = (\Sigma_i)_{i=1:K-1}, A\}$. With this model, given the observation sequence, the VFOA recognition is done by estimating the optimal sequence of VFOA which maximizes $p(s_{0:T}|z_{1:T})$. This optimization is efficiently conducted using the Viterbi algorithm [8].

In many meeting settings, where people are most of the time seated at given physical positions, the model parameters can be set using a traditional machine learning approach. Given training data with VFOA annotations and head pose measurements, we can readily estimate all the parameters of the

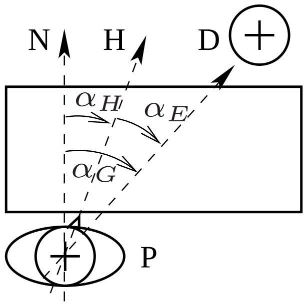

Figure 2: Model of gazing and head orientation.

HMM models. In our case, however, to avoid over-specialization to a specific type of meetings, the transition matrix was not learned but designed to exhibit a uniform stationary distribution and to favor the probability transitions of keeping the same focus according to: $a_{i,i} = \epsilon < 1$, and $a_{i,j} = \frac{1-\epsilon}{K-1}$ for $i \neq j$.

4.2 Cognitive Parameter Setting.

Annotating the VFOA of people in video recording is tedious and time consuming, as training data needs to be gathered and annotated for each meeting position and its possible VFOA targets. In the case of moving people, this can become even more complex. As an alternative to the training approach, we propose to take inspiration from the neurophysiology and cognitive science research on the modeling of the dynamics of the head/eye motions involved in saccadic gaze shifts [3, 5]. The proposed cognitive model is presented in Fig. 2, and reflects the fact that gazing at a target is usually accomplished by rotating both the eyes ('eye-in-head' rotation) and the head in the same direction. Given a person P whose reference (or rest) head pose corresponds to looking straight ahead in the N direction, and given that she is gazing towards D, the head points in direction H according to:

$$\alpha_H = \kappa_\alpha \, \alpha_G \qquad \text{if } |\alpha_G| > \xi_\alpha, \qquad \text{and 0 otherwise,} \tag{2}$$

where α_G and α_H denotes the pan angle to look at the gaze target and the actual pan angle of the head pose respectively. The parameters of this model, κ_α and ξ_α, are constants independent of the VFOA gaze target, but usually depend on individuals [3]. While there is a consensus among researchers about the linearity aspect of the relation, some researchers reported observing head movements for all VFOA gaze shift amplitude (i.e. ξ_α=0), while others do not. In this paper, we will assume $\xi_\alpha = 0$. Finally, in [3], it is shown that the tilt angle β follows a similar linearity rule.

Assuming we know the approximate room locations of the people's head, VFOA target, and camera[2], this cognitive model can be used to predict the values of the mean angles μ of the HMM VFOA model. The reference direction N (Fig. 2) will be assumed to grossly correspond to the gravity center of the gaze targets. For both person left and right, it corresponds to looking at O1 (cf Fig. 1c).

4.3 Maximum A Posteriori (MAP) Adaptation

The HMM model, with parameters learned through either the training or the cognitive approach, is generic and can be applied to any new person seated at the location corresponding to the learned model. In practice, however, we observed that people have personal ways of looking at targets. For example, some people use their eye more and turn their head less towards the focused target than others. In addition, our head pose tracking system is sensitive to people head appearance, and

[2]The relation in Eq. 2 is valid in the person's head reference. The camera position is needed in order to transform the obtained pose values into head poses w.r.t. to the camera.

can introduce a systematic bias in the estimated pose for a given person, especially for head tilt. Thus, the generic parameters might not be the best for a given person, and we propose to exploit the MAP adaptation principle to produce, in an unsupervised fashion, models adapted to people's characteristics.
The MAP principle is the following. Let z denote the sequence sample to which we want to adapt the parameters $\lambda \in \Lambda$ of the HMM model. The MAP estimate $\hat{\lambda}$ is then defined as:

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} p(\lambda|z) = \arg\max_{\lambda \in \Lambda} p(z|\lambda)p(\lambda) \tag{3}$$

where $p(z|\lambda)$ is the data likelihood and $p(\lambda)$ is the prior on the parameters. The goal is thus to find the parameters that best fit the data, while avoiding too large deviation from sensible values thanks to the parameter prior. The choice of the prior distribution is crucial for the MAP estimation. In [4] it is shown that for an HMM model, by selecting the prior pdf on λ as the product of appropriate conjugate distributions of the data likelihood[3], then the MAP estimation can be solved using the Expectation-Maximization (EM) algorithm. We followed this approach, modeling the prior on the Gaussian parameters as a Normal-Whishart distribution, and the prior on each row $p(.|s = f_i) = a_{i,\cdot}$ of the transition matrix with a Dirichlet distribution. Details of the EM equations are given in [4].

5 Experiments

5.1 Evaluation Set Up

Evaluation was conducted using the IHPD database (Section 2). As performance measure, we used the Frame based Recognition Rate (FRR) which corresponds to the percentage of frames during which the VFOA has been correctly. recognized. Since we are also interested in the temporal patterns followed by the VFOA events (i.e. temporal segments with the same VFOA label), which contain information related to interaction, we considered the following measures:

$$\rho_E = \frac{N_{matched}}{N_G}, \pi_E = \frac{N_{matched}}{N_R} \text{ and } F_E = \frac{2\rho_E \pi_E}{\rho_E + \pi_E}, \tag{4}$$

where N_G and N_R denote the number of events in the ground truth G and recognized R sequences of VFOA events, respectively, and where $N_{matched}$ represents the number of events in R that match the same event in G after a string alignment procedure that takes into account the temporal extent of the events. The recall ρ_E measures the percentage of ground truth events that are correctly recognized while the precision π_E measures the percentage of estimated events that are correct. The F-measure F_E accounts for both ρ_E and π_E.
Protocol: Comparisons between the learning and cognitive approaches will be made. In the learning case, a leave-one-out approach is used: each meeting recording is in turn left aside for testing, while the 7 other recordings are used for parameter learning. In the cognitive case, we used $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.5)$ to set the Gaussian means in the HMM models (variances were set by taking into account object size and pose). The transition matrix was the same in both cases (cf 4.1). In addition, in MAP adaptation the means and covariances in the Normal-Whishart priors were set to the values used without adaptation, and other parameters were set through cross-validation.

5.2 Results

Fig.3 illustrates VFOA recognition results, and Table 2 and 3 shows the results obtained for the left and right persons (see Fig. 1), under the different experimental conditions.
Overall results: From the result tables, one can first notice that in all conditions, the results for

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

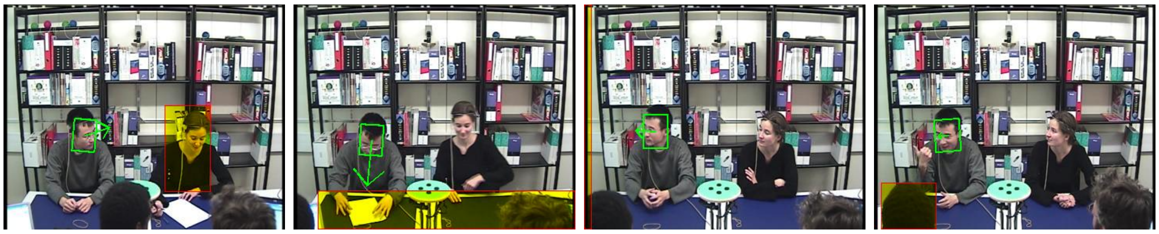

Figure 3: Example of results. The tracking result and head pointing direction appear in green. The recognized focus target appear in yellow (from image left to image right: PR, TB, SS and O1).

person left are much higher than for person right (more than 10%). This can easily be explained by the strong ambiguity for person right between looking at person left and at the slide screen (see Fig. 1), as confirmed by the confusion matrices [1]. Indeed, the average angular distance between close targets is around 20° for person right, a distance which can easily be covered using only eye movements rather than head pose rotation.

When using head pose tracking estimates rather than sensor data, we can observe some substantial performance degradation. A closer look at the numbers shows that overall the decrease is smaller for person right than for person left. This can be due to the better tracking performance -in particular regarding the pan angle- achieved on people seated at the person right position (cf Table 1). This correlation between tracking error and VFOA recognition performance is confirmed by the analysis of individual people's result.

Results with the cognitive approach: Fig. 4 shows the geometric VFOA Gaussian parameters (mean and covariance) obtained from the cognitive model. As we can see, the VFOA pose values predicted by the model are consistent with the average pose values computed for individuals using the GT pose data. Indeed, the computation of the average prediction error (average distance between each colored $\triangle$ symbols and the same color $+$ symbols in Fig. 4) is similar (around 6°) to the same error measure computed using the training approach. The recognition performance shows that, when using GT head pose data, the results with the cognitive approach are slightly worse than with the training approach. However, when using the pose estimates, the results are much better with the cognitive approach. Given that the modeling does not request any training data, this is an interesting and encouraging result.

Results with model adaptation: Overall, we observe for person left that adaptation brings a modest improvement when using GT pose data and a much larger one when using tracking estimates. The explanation is that, since for the left persons the different VFOA targets are rather un-ambiguous in the head pose space, with the clean GT data most of the VFOA information contained in the head pose is already captured by the baseline training approach. With the pose estimates, however, adaptation can cope with the variability in the individual people tracking errors (possibly systematic, e.g. underestimation), and provide a better estimate of the VFOA parameter which results in better performance. For person right, we notice improvement with both the GT and tracking head pose data. In this case, the presence of VFOA ambiguities allows to observe improvement event with GT input data. It is interesting to notice that the adaptation brings the same type of improvement when the cognitive approach is used instead of the training approach, despite the fact that the raw results (without adaptation) are different, as commented above. All together, in comparison with the baseline training approach, the use of both the cognitive setting and the adaptation leads to better results, with a significant improvement when the input data are head pose estimates.

	gt	gt-co	gt-ad	gt-co-ad	tr	tr-co	tr-ad	tr-co-ad
FRR	**72.3**	69.3	**72.3**	70.8	47.4	55.2	53.1	**59.5**
ρ_E	**72.1**	61.4	68.8	64.4	38.4	**42**	40.5	41.9
π_E	55.1	**70.2**	64.4	67.3	59.3	63.7	62.5	**69.9**
F_E	61.2	65.2	**66.6**	65.3	45.2	48.2	47.9	**50.1**

Table 2: VFOA recognition results for person left using GT (gt) or tracking head pose estimates (tr) as input data, with the learning or the cognitive (-co) approach, and with (-ad) or without adaptation.

	gt	gt-co	gt-ad	gt-co-ad	tr	tr-co	tr-ad	tr-co-ad
FRR	57.3	51.8	**62**	58.5	38	41.1	41.8	**42.7**
ρ_E	58.4	43.7	**63**	52.2	37.3	41.9	43.6	**43.8**
π_E	63.5	69	64.5	**71.5**	55.1	**61.1**	56.1	**61.1**
F_E	59.5	53	**62.7**	59.2	43.8	49.1	48.8	**50.1**

Table 3: VFOA recognition results for person right.

6 Conclusion

In this paper, we have presented an approach to recognize the VFOA from head pose data, which relies on a MAP adaptation technique and a cognitive parameter setting that does not require training data. Compared with a standard approach with training data, the combination of the two new features provides much better results when using head pose estimates.
Future research will be directed along two tracks. In the first one, we will further investigate and validate the cognitive model in the case of moving VFOA targets. In the second one, we will look for new models to model the joint VFOA and speaking status of all meeting participants from ultimodal data (ausio, vision slide screen changes) and derive conversation structure, as done in [7].

References

[1] S. Ba and J.M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington DC, May 2006.

[2] Sileye O. Ba and Jean Marc 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.

[3] Edward G. Freedman and David L. Sparks. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77:2328–2348, 1997.

[4] J.L. Gauvain and C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–213, 1992.

[5] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *TRENDS in Cognitive Sciences*, 9(4):188–194, 2005.

[6] S.R.H. Langton, R.J. Watt, and V. Bruce. Do the eyes have it ? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–58, 2000.

[7] 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 Multimedia and Exposiiton (ICME'06)*, Toronto, June 2006.

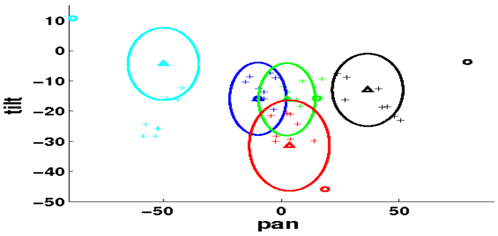

Figure 4: VFOA Gaussian distributions for person left (pose w.r.t. camera): gaze target direction (◯ symbols); corresponding head pose contribution according to the cognitive model (△); average head pose of individual people from GT pose data (+). Ellipses display the standard deviations used in the cognitive modeling. PR=black, SS=cyan, $O1$=blue, $O2$=green, TB=red.

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

[9] K. Smith, S.O. Ba, D. Gatica-Perez, and J.-M. Odobez. Multi-person wandering focus of attention tracking. In *International Conference on Multimodal Interfaces*, Banff, Canada, November 2006.

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