Recognizing the Visual Focus of Attention for Human Robot Interaction

Samira Sheikhi[1,2] and Jean-Marc Odobez[1,2]

[1] Idiap Research Institute, Switzerland
[2] École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. We address the recognition of people's visual focus of attention (VFOA), the discrete version of gaze that indicates who is looking at whom or what. As a good indicator of addressee-hood (who speaks to whom, and in particular is a person speaking to the robot) and of people's interest, VFOA is an important cue for supporting dialog modelling in Human-Robot interactions involving multiple persons. In absence of high definition images, we rely on people's head pose to recognize the VFOA. Rather than assuming a fixed mapping between head pose directions and gaze target directions, we investigate models that perform a dynamic (temporal) mapping implicitly accounting for varying body/shoulder orientations of a person over time, as well as unsupervised adaptation. Evaluated on a public dataset and on data recorded with the humanoid robot Nao, the method exhibit better adaptivity and versatility producing equal or better performance than a state-of-the-art approach, while the proposed unsupervised adaptation does not improve results.

Keywords: Human robot interaction, visual focus of attention, gaze, head pose.

1 Introduction

Endowing a humanoid robot with the capacity to interact with multiple persons at the same time requires the design of perceptual algorithms allowing the robot to analyze human behaviors and understand their intent. In particular, it is essential for the robot to be able to recognize communicative behaviors expressed by surrounding people.

In this paper, we addressed the recognition of gaze, and more precisely, the recognition of the VFOA (who is looking at whom or what). VFOA is an important cue for supporting interactions and dialog modeling: it is a good indicator of addresseehood (who speaks to whom, and in particular is a person speaking to the robot), but also a good cue to understand interaction between people or their level of interest. For instance, in a Museum scenario, if people are looking at the painting currently explained by Nao (our project robot), they are probably following the discourse. In order to create effective and natural conversational

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 99–112, 2012.

human-robot interfaces, it is desirable to have robots which can sense a user's gaze and infer appropriate conversational cues [14].

For estimating people's gaze, two main streams of work exist. Active sensing based methodologies based on infrared light are used very often. They are accurate but quite invasive and restrictive [5]. Computer vision techniques on the other hand use perceived information from gaze, head and body posture for recognizing VFOA [12]. This can be done using high definition images of the eyes. Still, it remains relatively constraining and usually restrict the mobility of the subject, considering the need for cameras with narrow field-of-views.

As an alternative, researchers have considered head pose as a clue for gaze [20], [18], [4], [15]. This idea is supported by the fact that turns of the head are a very informative cue in recognizing where the subjects are looking at [12]. Nevertheless, despite being very informative for recognizing VFOA, head pose it is an ambiguous cue: in realistic scenarios, the same head pose can be related to looking at different targets, depending on the situation; conversely, looking at a given target can be done using different head poses, as illustrated in Fig. 2. In this context, the following strategy is often exploited to recognize the VFOA:

- for a given person, track his head and estimate his head pose;
- map the head pose information to VFOA targets (looking at me, i.e. at the robot-, looking at another person, looking down, looking at a painting, elsewhere), and use this information within a recognizer to decode the latent sequence of VFOA targets. Note that the use of other cues like speaking utterance could be exploited as contextual information for recognition [4], [15], but is not addressed in this paper.

In practice, data driven approaches try to directly infer VFOA from head pose without estimating gaze as an intermediate step. Learned parameters, however, are then specific to the geometric configuration between the sensor (robot), the person, and VFOA targets. While this might be suitable in fixed settings [9], it is not adapted for a mobile robot dealing with moving people.

As an alternative we can exploit results from cognitive science studies about human gazing behavior and the dynamics of the head-eye motions involved in saccadic gaze shifts [10,8,11] to automatically determine which head poses should be associated with looking at a given target. This is done using a gaze model relating the head pose, a head-to-gaze ratio, and a head reference direction [3].

This reference direction, which corresponds to the direction perpendicular to the shoulder, was assumed to be fixed in [3] and set according to the setup. This assumption might not hold true in potentially more dynamic settings, e.g. those involving the robot. In these situations, we believe that an explicit or implicit estimation of the reference direction can result in more accurate VFOA recognition. In this context the contributions of the paper are the investigation of two models to dynamically estimate the reference pose, within a VFOA recognition task, and their evaluation on 3 datasets (meeting and robotics domain).

Section 2 goes through the related works. Section 3 reminds the basic Hidden Markov Model (HMM) used to recognize VFOA, the parameter setting issue, and introduce the standard gaze model. In Section 4, we introduce our new models,

providing the intuition behind them and their formal description. Results on meeting benchmark data and on Humavips Nao data are presented in Section 5, while Section 6 concludes the paper.

2 Related Work

In HRI and HCI context, many conversational systems need VFOA information for analyzing and performing necessary interactions. However, with respect to VFOA recognition, most works use either sensor-based or high definition image approaches which are not usually applicable for interaction with robots. The remaining works mostly take a very simplified version of the problem or do not explicitly mention VFOA recognition at all. For instance, in [7], [6] it is not mentioned how the VFOA is extracted and it is only used by the other modules. In [6] the problem is relaxed by only inquiring if the person is looking at the system or not. In [13] detecting a frontal face at a suitable spatial location is enough to adjust the classification to a higher level of engagement. In other works such as it is not clear how they solve the task. In [16] gaze is expressed in terms of head movements but it is not mentioned how to extract it and in [17] it is admitted that gaze is a very fundamental cue in human-human and HRI, but still nothing is mentioned about its extraction.

In another context (meeting), several works explored Dynamic Bayesian Networks (DBN) relying on head pose only [18] or multimodal data [15], [4] for VFOA recognition. All of them rely on Gaussians to model the distribution of head pose for looking at a given target, but only [3] uses a gaze model that does not require annotated data for setting the Gaussian means, or the manual setting of prior values. This allows for an easy exploitation for different observer-VFOA target configuration. Still, the head reference used in [3] is considered to be fixed and and set to the middle of the VFOA targets, it does not evolve dynamically preventing its adaptation to the focus context.

The approach in [19] follows the same approach for setting the means of the Gaussians. In a dynamic scenario they propose to use a discrete set of different head-to-gaze ratios according to the gaze dynamics. From the set of different ratios they take the one with the highest weight for each situation. This is quite different from our proposition since we try to compensate the model limitation by estimating real reference head directions.

3 VFOA Recognition Using HMM

3.1 The HMM Model

A basic solution for inferring the VFOA from head poses is to model the distribution of head poses with a K component Gaussian mixture model where K is the number of existing targets [18]. This method assigns the head poses lying on a specific Gaussian with the corresponding visual target. The model can be easily extended to an HMM as shown in Fig. 1(a), allowing to incorporate temporal information and obtaining more continuous and consistent VFOA results.

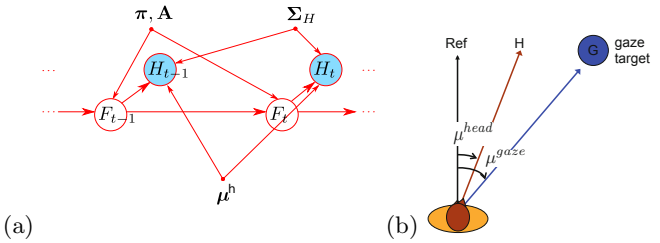

Fig. 1. (a) HMM graphical model for VFOA recognition with raw parameters. (b) Head-Gaze relationship. The person is assumed to be looking at the reference direction at rest (this direction grossly corresponds to the body orientation). Then, looking at a gaze target is accomplished by both the eyes and head. As a first approximation, the head rotation is a linear fraction of the full gazing rotation.

Let H_t and F_t indicate head pose (represented by a pan and tilt angles) and focus values at time t, and A denote the transition matrix in the HMM. Moreover let $\mu^{head} \in \mathbb{R}^{K\times 2}$ and $\Sigma_H \in \mathbb{R}^{K\times 4}$ denote the means and covariances of the K Gaussians. The HMM equations can then be written as follows:

$$P(H_t|F_t = n) = \mathcal{N}(H_t|\mu^{head}(n), \Sigma_H) \tag{1}$$

$$P(F_t = m|F_{t-1} = n) = A_{nm} \tag{2}$$

3.2 The Parameter Setting Issue

A major question is how to set the HMM parameters: the means μ^{head}, covariances Σ_H and transition matrix A. Following previous work, covariances can be set according to the size and proximity and of targets. The transition matrix A can also be set to satisfy our expectation of preserving the continuity in the sequence, an no other preferences. However, setting the means of the Gaussians μ^{head} is not possible in an easy way as it is highly related to the configuration of the observer and the targets and plays the most important role in the model.

The Training Approach. relies on annotated data to estimate the model parameters. However, annotating the VFOA of people in videos is difficult and time consuming, as training data needs to be gathered and annotated for each possible configuration of participant, targets and settings. This is especially problematic if people are free to move.

The Geometric Gaze Modeling Approach and Head Reference Direction. To overcome the above difficulty we can use cognitive findings on gazing behavior [8,11] which state that gazing at a target is accomplished by rotating

both the eyes ('eye-in-head' rotation) and the head (and sometimes even the body in the same direction) as illustrated in Fig. 1(b). The relative contribution of the head and eyes towards a given gaze shift is found to follow simple rules [8], [11]. More precisely, the means of the Gaussians corresponding to each specific target can be set as a fixed linear combination of the target direction and the *head reference* direction. For a gaze target indexed by n, we have:

$$\mu^{head}(n) - R = \alpha\ (\mu(n) - R) \quad if \quad |\mu(n) - R| > \lambda_\alpha \tag{3}$$

or equivalently (if we set λ_α to 0):

$$\mu^{head}(n) = \alpha\ \mu(n) + (1 - \alpha)R \tag{4}$$

where $\mu^{head}(n) - R$ is the rotation made by the head to look at the direction of the target, $R \in \mathbb{R}^2$ denotes the head reference direction and $\mu \in \mathbb{R}^{K\times 2}$ denotes the target directions. The coefficient α is usually set between 0.5 and 0.7 for pan and between 0.3 and 0.5 for the tilt angle. For a given application, a suitable value can be obtained by studying the existing behavior on training data of different individuals.

Equation 4 can be used to set the mean of the Gaussian corresponding to target n in our HMM model. In previous work, the reference vector R was set to a constant value (eg the median of the target directions in [3]). Assuming this as a baseline, the probability of observations given the VFOA states is then given by the following equation:

$$P(H_t|F_t = n, \mu_t) \sim \mathcal{N}(H_t|\ \alpha\mu_t(n) + (1 - \alpha)R,\ \Sigma_H) \tag{5}$$

4 Exploiting Temporal Head Reference Estimates

Setting the means of the Gaussians using the cognitive model requires the knowledge about the value of the reference R as well as the directions of the targets. Equation 4 shows the importance of the reference for recognizing correct targets. Note that using a wrong value for R produces shifted mean values for all of the targets $\mu^{head}(n)$ simultaneously, which can have dramatic effects.

This importance of knowing the head reference is also illustrated in Fig. 2. It shows that, unless the head reference directions (people shoulder's orientation) are more or less constrained by the setting (e.g. like when people are seated in a meeting) or the situation is known (e.g. in the quiz scenario, people are dominantly facing the robot), when can not use a constant reference direction in our model. In more versatile situations and interactions, we have many variations and shifts in the reference as people are free to move. These reasons motivate us to find a suitable way for setting the reference dynamically. Therefore, we proposed two different solutions for setting the reference and their corresponding probabilistic models as explained in the following sections.

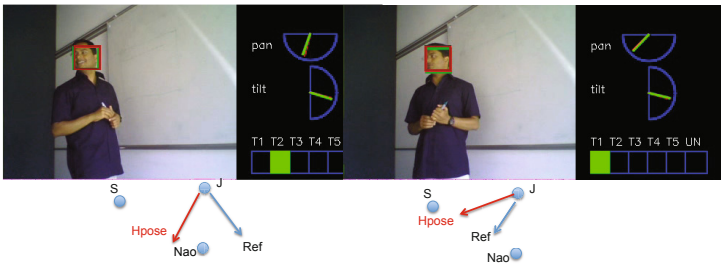

Fig. 2. Different reference directions (shoulder orientations) lead to different poses for looking at the same target. In both images, person J looks at person S. These images illustrate that the geometric model is holding true: the head orientation is approximately half-way between the reference direction and the gaze direction. On the left image, using looking at Nao as reference direction could lead to a wrong interpretation of the head pose on the right as looking at Nao.

4.1 First Model G1

Intuition: For the first model we tend to use a general notion for the reference which is in average acceptable. The principle is that a person tends to orient himself towards the set of gaze targets he/she spends time looking at. Such a body position makes it more comfortable and less energy consuming to rotate his head towards different gaze targets. As a corollary, this means that his average head pose over a time window is a good indicator of his reference direction, and can be used as an estimate of this direction. Although such an estimate might not be very sensitive to local changes and temporal variations and does not account for the previous head pose (that is involved in gaze shifts according to the cognitive studies that led to the geometrical model), it can provide a robust angle estimate that reflects the overal balanced direction of the head or body. Therefore we can set the reference value at each frame R_t^0 to the average of the person's head pose over a previous time window:

$$R_t^0 = \sum_{i=t-w}^{t} H_i/w$$

Setting the reference in this way is also linked to the midline effect [12] which plays an important role in the head direction needed looking at a target.

Model: Since the reference directions set this way are known from the head poses, they serve as the observations in the model as illustrated in Fig. 3. Here again μ_t^{head} denotes the expected means for the head poses. The dynamics between the hidden states are the same as previously and the rest of the relationships are formulated as follows:

$$P(\mu_t^{head}(n)|R_t^0) \sim \mathcal{N}(\mu_t^{head}(n)|\ \alpha\mu(n) + (1-\alpha)R_t^0,\ \Sigma_\mu) \quad (6)$$

$$P(H_t|F_t = n, \mu_t^{head}) \sim \mathcal{N}(H_t|\ \mu_t^{head}(n),\ \Sigma_H) \quad (7)$$

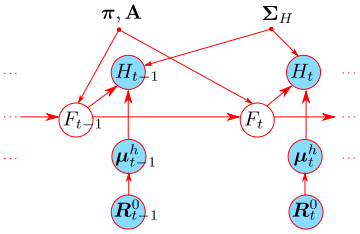

Fig. 3. First model. The head reference direction and the mean head pose of the Gaussians are now variables over time. However, they are observed variables: the head direction is defined as the average of the head poses over a temporal window, and the mean head poses are then deduced from the geometric gaze model.

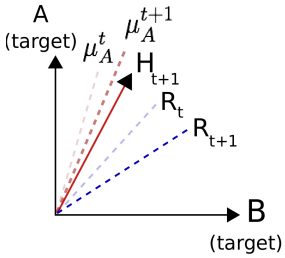

Fig. 4. Unsupervised reference adaptation. Assume that at time $t+1$ the head pose H_{t+1} is associated with target A of current head pose mean μ^t_A in the picture. Trusting the current recognition, adaptation will move the mean μ^{t+1}_A at time $t+1$ closer to the observation and as a result of the gaze geometrical model (assuming there is no change in target positions during this time interval), the reference direction will move accordingly.

The recognition of the VFOA values is straightforward by running the classical inference algorithm on this HMM.

4.2 Second Model G2

Intuition: Setting the reference using long term head pose value statistics might not be sufficient, as more local (short term) gaze activity can come into play. We can thus try another strategy and adapt it in an unsupervised way in order for the model to better fit the observations. This is illustrated and explained in Fig. 4: we would like to change the reference R in order to maximize the probability of observing these new head pose values given their recognized targets.

Model: to accommodate the long term statistics with the short term adaptation, we add a new variable R_t denoting the real reference direction at each frame t and use the estimated head reference direction R^0_t from the average head pose as a prior to this variable. This is illustrated in Fig. 5. By adding R_t as a hidden variable to the model we would then infer the reference values around R^0_t such that the predicted means μ^{head}_t can best fit the observations.

Notice that the same equations 7 and 6 from the previous model are still valid here by substituting R_t for R^0_t. We also expect the reference to be continuous over successive frames and thus the value of R_t should be dependent on its value R_{t-1} at the previous frame. Therefore, we set it as a linear combination of R_{t-1} and R^0_t. The main intention for including the prior value R^0_t in our model is to avoid R from deviating too much from a reasonable range. The following equation formulates this relationship:

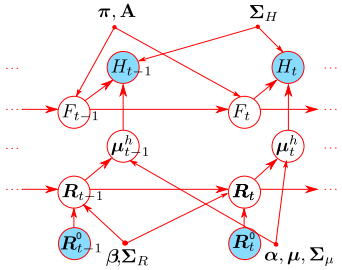

Fig. 5. Second model, unsupervised adaptation for the reference

$$P(R_t|R_{t-1}, R^0) \sim \mathcal{N}(R_t|\ \beta R_{t-1} + (1-\beta)R^0,\ \Sigma_R) \tag{8}$$

Inference: Given the probabilistic model, we wish to determine the sequence of visual focus of attentions F_t (VFOA) of a person from the observed head poses H_t. All the parameters in the model are assumed to be given and remain fixed throughout the inference. We can note that if the head poses μ^{head} are known, our cognitive VFOA model splits into two parts: the VFOA values follow a standard HMM model, whereas the reference variables follow a Kalman filter model. We thus use the following approximate procedure for inference. At time t, we first apply the prediction step for the R_t value and then the targets μ^{head} mean, apply the HMM filtering step for the VFOA state, and then apply the update steps for μ^{head} and R_t given the estimated VFOA for which efficient inference procedures exist.

5 Experimental Results

5.1 Data Sets and Experimental Protocol

For our experiments we use three sets of data. In the first dataset we have the recordings of eight meeting sessions with a total duration of 145 minutes. All of the meetings are recorded under the same condition and with similar configuration as shown in Fig. 6, with four people (Person left Pl and Person right Pr seen on the image, and two organizers O1 and O2 seating in front of them) discussing statements displayed on slides. We perform our study on the two persons on the seats in front of the camera. For this dataset we have ground truth head poses, captured from flock of bird sensors which will be used for analysis. Each of the participants has five possible gaze targets: three other persons, the slide screen and the table.

In the second dataset (D1) we have a video recorded by our robot, Nao. The total duration of this video is 22 minutes. In this case there are two participants seating in front of Nao as shown in Fig. 6. For this dataset, we do not have

Fig. 6. Datasets. (Left) Meeting setting, with VFOA targets for the person on the right (PR). (Middle) Nao dataset 1 (D1) with two participants in front of Nao. (Right) Nao dataset 2 (D2), from Vernissage recordings, with VFOA targets for one of the two participants.

ground truth head poses and analysis are done using the tracked head poses [2] which does joint tracking and head pose estimation [1]. Each of the participants have three visual targets: the other participant, Nao and a booklet which they refer to during the recording.

In the third dataset (D2) comes the Vernissage (the word refers to the preview of an art exhibition) data recording (Fig. 6). There we have one session during which people participate in a quiz given by the robot. The recordings take around 6:30 minutes. Here again we used tracked head poses for analysis. As shown in Fig. 6, there are five main VFOA targets in this recording: Robot (NAO) Person1 (partner), Painting1, Painting2, Painting3. In addition, we use the label Others when people look elsewhere (often down in front of them, with not much head pose change).

Performance Measure: As performance measure we use "Frame based Recognition Rate (FRR)" which corresponds to the percentage of frames during which the VFOA has been correctly recognized.

Algorithms: We have performed our experiments with three different models as summarized in Table 1. The baseline is the basic HMM model using an initially set and fixed reference value for all of the frames. The other models were presented in the previous Section: G1 uses the head pose average over a temporal window as the reference; and G2 adapts the reference value in an unsupervised fashion, using the head pose average value as prior at each frame.

Table 1. Tested algorithms

Model	Reference Prior	Unsupervised Adaptation
Baseline	set as the initial reference	No
Model G1	head pose average over a window	No
Model G2	head pose average over a window	Yes

5.2 Parameter Setting

Meeting Data. We set the variances of the Gaussians according to the size of the targets. For the meeting data we use the same values as in [3] which are $\sigma_\alpha(O_1, O_2, P_R, P_L) = 12$, $\sigma_\alpha(SS) = 25$, and $\sigma_\alpha(TB) = 20$ for the pan, and $\sigma_\beta(O_1, O_2, P_R, P_L) = 12$, $\sigma_\beta(SS, TB) = 15$ for the tilt. Here, σ_α and σ_β show the variances for pan and tilt values. Moreover, O_1, O_2 indicate the observers, P_R, P_L the persons on the right and left, and SS, TB indicate the slide screen and the table respectively.

For the gaze directions, they were assumed to be fixed for each recording (thus neglecting people's motion), and currently defined from the geometrical setting. The initial value for the reference direction is particularly important for the baseline for which it remains the same over time, but not important for the other models as the reference value is quickly set as the average over head pose values. For the baseline, we experimented with setting the reference as the middle of the gaze target directions, which was shown to work the best in previous works [3].

Table 2. Parameter set using cross-validation

Parameters	Baseline	1st Model G1	2nd Model G2
α_{pan}	$\surd$	$\surd$	$\surd$
α_{tilt}	$\surd$	$\surd$	$\surd$
self-loop	$\surd$	$\surd$	$\surd$
window-size	$\times$	$\surd$	$\surd$
σ_R	$\times$	$\times$	$\surd$
ratio σ_μ/σ_H	$\times$	$\times$	$\surd$
β	$\times$	$\times$	$\surd$

The remaining parameters of the models which are summarized in Table 2 were adjusted by cross-validation separately for each of the models. For the meeting data there are two different set-ups for people seating on the first and second seats and therefore we did the cross validation once by taking the training set from the same seat and once by taking it from the other seat. The second case is useful to evaluate whether our model is sensitive to a specific setting or it is more general. For training with data from the same seat, leave-one-out cross validation was used, taking seven meetings for training and testing on the eighth meeting. For training with data from the other seat, all eight meetings from the other seat were used for training and obtained parameters were used to test on the other seat.

Table 3 summarizes the parameters which were obtained in cross validation for the baseline and Table 4 summarizes chosen parameters for the first model G1. There is a strong agreement between the parameters which were obtained for left and right people. Also there is a strong overall consistency between the

parameters trained using the first seat data and those of the second seat. For the baseline model, α_{pan} obtained from two different seats is different. This could be through to the fact that the reference which is used there (the middle of the targets) is a poor reference and force and introduces different results for these two different settings. This effect does not exist for the fist model G1 and the chosen parameters are completely consistent.

Table 3. Chosen parameters for the meeting data and baseline model

Person	Training	Parameters: α_{pan}, α_{tilt}, self-loop
Person on left	same seat	0.5 - 0.5 - 0.75
Person on left	other seat	0.8 - 0.5 - 0.75
Person on right	same seat	0.8 - 0.5 - 0.75
Person on right	other seat	0.5 - 0.5 - 0.75

Table 4. Chosen parameters for the meeting data and first model G1

Person	Training	Parameters: α_{pan}, α_{tilt}, self-loop, wind-size
Person on left	same seat	0.7 - 0.5 - 0.75 - 500
Person on left	other seat	0.7 - 0.4/0.45 - 0.75 - 500
Person on right	same seat	0.7 - 0.4/0.45 - 0.75 - 500
Person on right	other seat	0.7 - 0.5 - 0.75 - 500

Nao First Dataset (D1). For the gaze directions, same as the meeting data, they are assumed to be fixed for each recording and defined from the geometrical setting. The initial value for the reference direction is considered to be at Nao's direction which is a reasonable choice in human robot interaction scenario. We set the standard deviations of the targets to 8° for the pan and 4° for the tilt angle. Notice that these values are smaller compared to the meeting data. This choice is made both our Nao datasets where we use tracker results for head poses since those head poses are usually smaller than the ground truth pose values.

For the rest of the parameters, as there are a few number of people participating in this dataset with very different gazing behaviors cross-validation will not produce reliable parameters. To choose the parameters we consider the meeting data as the training set and use parameters obtained from that data for running our algorithms on Nao's data. Note however, that the resulted α_{pan} value from meeting data is 0.7. In Nao data this ratio is big considering the tracked head poses which are a little underestimated. Therefore we do our experiments with a smaller value of 0.65.

Nao Second Dataset (D2). Standard deviations of the pan angles were set to 8°, 8°, 8°, 9°, 10° respectively for Robot, Partner, Painting1, Painting2 and

Painting3, according to their size and proximity. The tilt angle standard deviations were set to 4° for all targets. The remaining parameters are all set in the same way they are set for D1.

5.3 Results

Meeting Data. Table 5 shows the results of the baseline, G1 and G2 models. As can be seen, the first model outperforms the baseline. This is particularly true in more mismatched conditions, when parameters are learned from another seat rather than the same seat, thus exhibiting a better adaptation capacity. In particular, we can notice the performance degradation for person right (PR) when using the optimal parameters for person on left (PL). The main (mismatched) parameters leading to the degradation is the parameter α of the gaze model (see Eq. 4) that directly impact the prediction of the head poses: for PL, the optimal parameters is around 0.8, whereas for PR, it is around 0.5. Using the head pose average for the reference is indeed a more stable choice for this important parameter, with an optimal value for both seats around 0.7.

On the other hand, we can see that the 2nd model G2 performs very closely to the G1 model, a behavior that will be seen in other recordings as well. This means that in practice the unsupervised adapted head reference remains very close to the prior, and thus the models behave very similarly.

Table 5. Performance evaluation on Meeting data

Person	Training	Baseline	Model G1	Model G2
Person on left	same seat	64.7	**65.7**	65.5
Person on left	other seat	64.5	66.7	**66.8**
Person on right	same seat	57.0	**58.7**	58.6
Person on right	other seat	43.9	**59.0**	59.0

Nao First Data D1. The results of the baseline, G1 and G2 are summarized in Table 6. Despite the quite different setting (situation, number of gaze targets, use of estimated head pose vs ground truth head pose), the conclusions are similar to the meeting data. More precisely, model G1 outperforms the baseline, particularly for people on the left with a large difference, and model G2 performs almost the same as model G1.

Table 6. Performance evaluation on Nao data D1

Person	Baseline	Model G1	Model G2
Person on left	69.4	**78.7**	**78.8**
Person on right	**66.5**	66.2	66.2

As the table shows, we have a high gain using our models for the person on the left while the performance of the models are very close for the person on

the right. The main reason for this behavior is that considering the participants body configuration Nao's direction is quite a suitable choice for the reference for the person on the right but it cannot perform as a good reference estimation for the person on the left.

Nao second Data D2. Table 7 shows VFOA recognition rates obtained our first model compared to the baseline model. As it is shown in the table for person on the right we get better results using our model G1 whereas for the person on the left this is not true. We need to verify these results using more sequences from this dataset to find the overall performance behavior.

Table 7. Comparison of FRR of Baseline and the 1st model G1 on quiz recording 9

Person	Baseline	Model G1
Person on Left	**54.6**	51.2
Person on Right	58.8	**59.6**

6 Conclusion

In this Section, we have presented our research towards designing better gaze models for improved VFOA recognition. We have shown that the implicit estimation of the head reference has a positive impact on performance. It is important to underline that for the different recordings the choice of head reference for the baseline is a very good approximation of the actual value. In practice, such a value might be difficult to set. As we have seen, in the robot interaction application, the same strategy (looking at Nao) does not produce good results in all conditions. Similarly, using the middle of the gaze target directions might work, but assumes that the robot is aware of all target directions a person can be looking at. This might not hold true in all cases.

Our future work contains an assessment of the model on larger amounts of recordings with the robot from the Vernissage dataset; assessment of the model using the true head poses, to mitigate the effect of head pose estimation on performance evaluation. Moreover, we would compare the results of our VFOA models using tracked head poses versus the ground truth head poses to study how the head pose estimation error affects the VFOA recognition.

Acknowledgments. The authors gratefully acknowledge the financial support from the HUMAVIPS project, funded by the European Commission Seventh Framework Programme, Theme Cognitive Systems and Robotics, Grant agreement no. 247525.

References

1. Ba, S.O., Odobez, J.-M.: Evaluation of multiple cue head pose estimation algorithms in natural environements. In: IEEE Int. Conf. on Multimedia and Expo (2005)

2. Ba, S.O., Odobez, J.-M.: Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 276–286. Springer, Heidelberg (2008)
3. Ba, S.O., Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. Trans. Sys. Man Cyber. Part B 39, 16–33 (2009)
4. Ba, S.O., Odobez, J.-M.: Multiperson visual focus of attention from head pose and meeting contextual cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 101–116 (2011)
5. Babcock, J.S., Pelz, J.B.: Building a lightweight eyetracking headgear. In: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, ETRA 2004, pp. 109–114. ACM, New York (2004)
6. Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proc. of the SIGDIAL Conference, Stroudsburg, USA, pp. 225–234 (2009)
7. Bohus, D., Horvitz, E.: Open-world dialog: Challenges, directions, and prototype. In: Proceedings of IJCAI 2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems (2009)
8. Freedman, E.G., Sparks, D.L.: Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. Journal of Neurophysiology 77(5), 2328–2348 (1997)
9. Gaschler, A., Huth, K., Giuliani, M., Kessler, I., de Ruiter, J., Knoll, A.: Modelling state of interaction from head poses for social human-robot interaction
10. Hanes, D.A., McCollum, G.: Variables contributing to the coordination of rapid eye/head gaze shifts. Biol. Cybern. 94, 300–324 (2006)
11. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in Cognitive Sciences 9(4), 188–194 (2005)
12. Langton, S.R., Watt, R.J., Bruce, I.: Do the eyes have it? cues to the direction of social attention. Trends Cogn. Sci. 4(2), 50–59 (2000)
13. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: 9th IEEE Int. Workshop on Advanced Motion Control (2006)
14. Morency, L.-P., Darrell, T.: Conditional Sequence Model for Context-Based Recognition of Gaze Aversion. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 11–23. Springer, Heidelberg (2008)
15. Otsuka, K., Takemae, Y., Yamato, J.: A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of the 7th International Conference on Multimodal Interfaces, ICMI 2005, pp. 191–198. ACM, New York (2005)
16. Sidner, C.L., Lee, C.: Engagement rules for human-robot collaborative interactions. In: IEEE Int. Conf. on Systems, Man and Cybernetics, vol. 4 (2003)
17. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1), 140–164 (2005)
18. Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, ICMI 2002, p. 273. IEEE Computer Society, Washington, DC (2002)
19. Voit, M., Stiefelhagen, R.: Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proc. of the 10th Int. Conf. on Multimodal interfaces (ICMI), Chania, Crete, Greece (2008)
20. Yücel, Z., Salah, A.A.: Resolution of focus of attention using gaze direction estimation and saliency computation. In: Proceedings of the International Conference on Affective Computing and Intelligent Interfaces (2009)