This paper was published at the Int. Conf. on Multimodal Interaction, Santa Monica, October 2012

Investigating the Midline Effect for Visual Focus of Attention Recognition

Samira Sheikhi
Idiap Research Institute and EPFL, Switzerland
samira.sheikhi@idiap.ch

Jean-Marc Odobez
Idiap Research Institute and EPFL, Switzerland
odobez@idiap.ch

ABSTRACT

This paper addresses the recognition of people's visual focus of attention (VFOA), the discrete version of gaze indicating who is looking at whom or what. In absence of high definition images, we rely on people's head pose to recognize the VFOA. To the contrary of most previous works that assumed a fixed mapping between head pose directions and gaze target directions, we investigate novel gaze models documented in psychovision that produce a dynamic (temporal) mapping between them. This mapping accounts for two important factors affecting the head and gaze relationship: the shoulder orientation defining the gaze midline of a person varies over time; and gaze shifts from frontal to the side involve different head rotations than the reverse. Evaluated on a public dataset and on data recorded with the humanoid robot Nao, the method exhibit better adaptivity often producing better performance than state-of-the-art approach.

Categories and Subject Descriptors

I.2.9 [**Artifical Intelligence**]:Robotics-*Operator Interfaces*

Keywords

Gaze, visual focus of attention, HRI, HCI

1. INTRODUCTION

As a good indicator of people's interest and due to its major role in non-verbal communication, VFOA is an key cue for human interaction analysis and in Human-Computer/Robot interactions (HCI, HRI) design [3].

Sensor based methodologies can be used to estimate people gaze. They are accurate but quite invasive and restrictive. Computer vision techniques relying on perceived information from gaze or head has also made good progresses, but still usually restrict subject mobility considering the need for cameras with narrow field-of-views for looking at the eyes.

As an alternative, researchers have considered the use of head pose as a clue for gaze [7, 9, 1, 10, 4]. Head poses, however, are ambiguous: in realistic scenarios, the same head pose can be used to look at different targets, depending on the situation. To address this issue, researchers have proposed to exploit other cues: speaker information or conversational regimes [7] that can be exended with contextual knowledge from the group activity [1].

While context is important, a central issue is how to set the expected head pose of an observer that looks at a given target? in other words, how to define a mapping from the gaze target direction to the corresponding head pose. This is essential to set the parameters of recognizers like Hidden Markov Models (HMM) and decode the sequence of VFOA targets given the head pose sequence. Methods often rely on manual setting, potentially followed by adaptation [7].

One of the few work addressing this problem is [1]. Exploiting results on human gazing behavior and head-eye dynamics involved in saccadic gaze shifts [6, 5], they introduced a linear gaze model relating the head pose, gaze direction, and head reference (coined gaze midline in [5]) as illustrated in Fig. 1(a). While the method worked when applied to meetings, it suffers from two drawbacks: the reference direction, which corresponds to the direction perpendicular to the shoulder, was assumed to be fixed and set according to the setup. This approach might not be feasible in more dynamic settings, like in HRI with multiple people where the robot is not always the main focus, and more generally in scenarios involving people free to move and re-orient themselves, as illustrated in Fig. 2. The second drawback, pointed out in several psychovisual works, is that the mapping not only depends on the the gaze direction and midline reference, but also on the head or gaze direction before the shift, resulting in different head poses for looking at different targets even for the same head reference direction.

In this paper, we investigate both problems. In absence of shoulder orientation measures, we introduce an implicit estimation of the midline direction, and propose two approaches inspired by [5] to improve gaze-to-head mapping and gaze shift models. Experiments on meeting benchmark data and on data recorded by a robot (Nao) shows the benefit of several modeling components. To our knowledge, this has not been done before. The work in [9] is the closest to ours. In a dynamic scenario authors proposes to use a discrete set of head-to-gaze ratios and estimate the most likely based on gaze prediction and differential head pose indicators. However, the reference direction setting was not addressed, and their approach is quite different from our proposition that compensates the models' limitation relying on models inspired from human behaviors [5].

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

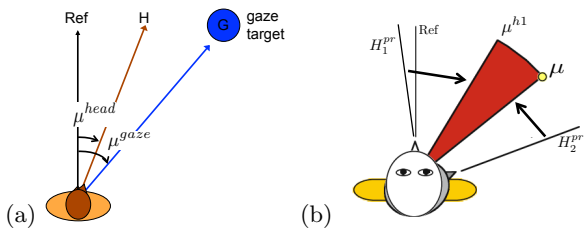

Figure 1: Gaze models. (a) Geometrical model. The person is assumed to be looking at the reference direction, or midline (grossly, the body orientation). Then, looking at a gaze target is accomplished by rotating both the eyes and head, with the head part being a fixed fraction of the full gaze rotation. (b) Midline effect [5]. The interval of head positions $[\mu^{h1}, \mu]$ corresponding to the target at position μ. When the gaze is moved to μ from initial position H_1^{pr}, the head is moved to μ^{h1} according the geometrical model. When the gaze shift is centripetal from H_1^{pr} to μ, the head is moved to μ. For initial head positions between μ^{h1} and μ, an eye-only saccade to μ is made (the head position remains the same).

2. VFOA RECOGNITION MODEL

2.1 VFOA recognition using a HMM model

In this model, the distribution of head poses associated to a given VFOA target is represented by a distribution (a Gaussian) with different parameters, whereas transitions between VFOA targets is represented by a transition matrix A. More specifically, let H_t and F_t indicate head pose and focus values at time t, and $\mu^h \in (\mathbb{R}^2)^K$ and $\Sigma_H \in (\mathbb{R}^4)^K$ denote the means and covariances of each of the K Gaussians associated with the targets. The HMM equations can be written as follows:

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

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

Parameter setting. This is a major issue. Following previous work, the covariance of each target can be set according to its size and head pose estimation variability. In absence of other prior, the transition matrix A can also reasonably be set to satisfy our expectation of preserving the VFOA continuity in the sequence.

However, although they play the most important role in the model, setting the means of the Gaussians μ^h is not possible in an easy way. Using training data is not really an option, since VFOA annotation is difficult, time consuming, and data needs to be gathered and annotated for each configuration of the observer, targets and settings. This is especially problematic if people are free to move.

A solution to the above difficulty is to use **gaze models** derived from gazing behavior [6, 5]. Accordingly, gazing at a target is accomplished by rotating both the eyes ('eye-in-head' rotation) and the head as illustrated in Fig. 1. More precisely, as a first approximation, the means of the Gaussians can be set as a fixed linear combination of the gaze and *head reference* directions. For a gaze target n, we have:

$$\mu^{hb}(n) - R = \alpha\ (\mu(n) - R) \Rightarrow \mu^{hb}(n) = \alpha\ \mu(n) + (1-\alpha)R \quad (3)$$

where $R \in \mathbb{R}^2$ denotes the reference direction and $\mu \in (\mathbb{R}^2)^K$ 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. Eq. 3 can be used to set the Gaussian mean corresponding to target n in our HMM model. Our baseline will consist of the above model with the reference R set to a constant value as done in most previous works.

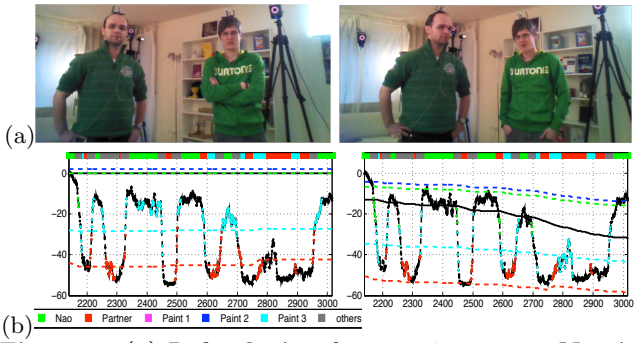

Figure 2: (a) Left: during frames 1700-2200, Nao is the main speaker, participants tend to look straight at him. Right: afterwards (quiz part) participants alternatively look at the robot and the second person (amongst others). Their reference direction is thus different, and so are the pose for looking at Nao. (b) head pose (pan angle) of the person on the right in image (a). The ground truth VFOA is displayed in the top bar, with color codes below. The head pose data is displayed in black when the recognition is correct, and in the color of the wrongly recognized VFOA otherwise. Dashed lines indicate the pan pose mean for looking at each target for the baseline (left), or the proposed model (right). In this later case, the black line shows the head reference. With the dynamic reference, head poses for looking at each of the target are better predicted, like for looking at Nao (despite its high variability: pan near 0 at frame 2150, near -17 at frame 2550).

2.2 Model G1: Dynamical Head References and the Midline Effect

Reference setting. Setting the Gaussians means using the above model requires the knowledge of R and of the target directions. Eq. 3 shows the importance of the reference: using a wrong value for R shifts mean values for all targets $\mu^h(n)$ simultaneously, which can have dramatic effects.

This importance of knowing the head reference is also illustrated in Fig. 2. Unless the reference direction (shoulder orientation) is constrained by the setting (eg when people are seated) using a constant reference can be problematic. General interactions will result in more variations and shifts in the reference as people are free to move, motivating the need for setting the reference dynamically.

Midline effect. As documented in [5], and illustrated in Fig. 1, the gaze model defined by Eq. 3 is only valid if the gaze shift goes from the reference to a given target. Indeed, in [5] shows that how much of a gaze shift is accomplished by the head or by the eye depends on the position of the head (which is not aligned with the reference in general) at the start of the gaze, and whether the shift goes through the reference or not, hence the term 'midline effect' used in [5][1]. From the analysis of the psychovision literature, the

[1]In [5], the reference is called midline.

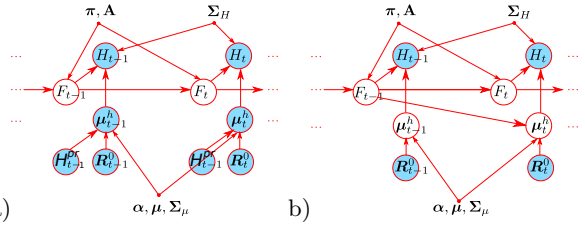

Figure 3: Graphical models. (a) Model G1. The head reference direction and the mean head pose of the Gaussians are now variables over time, and the recent head pose can be exploited. (b) model G2. The mean head pose for looking at a target depends on the gaze target at the previous time step.

authors derived a gaze model which is illustrated in Fig. 1b, and that we have investigated in our research.

First model G1. Our goal is to derive a gaze model that accounts both for a dynamically estimated reference, and for the midline effect. To address the first point, we relied on the following intuition. 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 time is a good indicator of his reference direction, and can be used as reference estimate. Therefore we set the reference value at frame R_t^0 as the head pose average over a previous time window:

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

To investigate the midline effect, we need to know what was the value of the head pose before the gaze shift occurs. To this end, we introduced the variable H^{pr} defining the previous head pose and used as estimate of this variable the average of the head poses computed over a window of size w^p separated by a gap δ^p from the current instant:

$$H_t^{pr} = \frac{1}{w^p} \sum_{i=t-w^p-\delta^p}^{t-\delta^p} H_i \quad (4)$$

Finally, the G1 gaze model was implemented by setting the head pose mean $\mu_t^{hg1}(n)$ of the n^{th} target at time t using the rules (for $\mu(n) > 0$ and omitting (n) for simplicity):

$$\mu_t^{h1} = \alpha\mu + (1-\alpha)R_t^0 \quad (5)$$

$$\text{if } H_t^{pr} < \mu \text{ then } \mu_t^{hg1} = \mu_t^{h1} \quad (6)$$

$$\text{otherwise } \mu_t^{hg1} = Min\left(\mu, \mu_t^{h1} + \alpha_H(H_t^{pr} - \mu_t^{h1})\right) \quad (7)$$

The above equations can be adapted for $\mu(n) \leq 0$. The factor α_H indicates how much we take into account the previous head pose in the estimate. When $\alpha_H = 0$, we always have $\mu_t^{hg1} = \mu_t^{h1}$, which means that the head pose means are set using the standard geometric model, but using a dynamically set reference. When $\alpha_H = 1$, the implemented model is exactly the one proposed by [5].

2.3 Model G2: implementing gaze shifts

When implementing the midline effect, the previous model has one drawback: at each time step, a gaze shift is assumed. In other words, even if the person is focusing on target i, the previous head pose H_t^{pr}, estimated through recursion over a short window, evolves and introduces an evolution of what the head pose mean of the target i itself should be.

Figure 4: Settings. Left. Meeting, with VFOA targets for the person on the right. Middle. Nao D1. Right. Nao D2, from vernissage recordings, with the VFOA targets for one of the 2 participants.

As alternative to the G1 model, we define the gaze situation prior to the gaze shift by the gaze at the previous instant. In this case, the head pose mean to look at target n at time t, given the previous focus F_{t-1}, is given by:

$$\mu_t^{hg2}(n) = \alpha_1\mu(n) + \alpha_2\mu(F_{t-1}) + (1-\alpha_1-\alpha_2)R_t^0 \quad (8)$$

Thus, in absence of gaze shift ($F_{t-1} = n$), the head pose mean is simply given by the geometrical model with $\alpha = \alpha_1 + \alpha_2$, while in case of a gaze shift $F_{t-1} \neq n$, the head pose is not only affected by the reference and new gaze direction $\mu(n)$, but also by the previous gaze (the head should be closer to the previous gaze than what would be predicted by the geometrical model).

Fig. 3(b) shows the G2 graphical model. The link between the hidden states F_{t-1} and μ_t^h renders the inference more complex than in a standard HMM. In practice, we conducted the inference sequentially, using the estimated focus at time $t-1$ to estimate the optimal focus at time t.

3. EXPERIMENTAL RESULTS.

3.1 Data Sets and Experimental Protocol

We used three datasets for experiments, with setup illustrated in Fig. 4. The **Meeting dataset** was taken from [2]. It has 8 meeting sessions (1 hour of data), and provides the ground truth head poses for the two persons in front of the camera. Each person has 5 possible gaze targets: the 3 other persons (the one on his side, two on the other table side), the slide screen and the table.

The two other datasets were recorded with the humanoid robot Nao and in both cases, head poses were extracted with an automatic algorithm [8]. In the first dataset, **D1**, two participants were seating in front of Nao and discussing about Nao's features. At one point, one of the person leaves and a third person comes in. The session lasts 22 minutes (min). Each person has 3 visual targets: the other participant, Nao and a booklet which they refer to during the recording. The 3rd dataset **D2**, involves two participants standing in front of Nao and free to walk around and look at different objects. In a first part, Nao explains three paintings to them; in a second part, Nao makes them a quizz where the participants can discuss before giving their answer. VFOA annotations are available for 3 recordings, for a total of around 22:30min per side, ie 45min in total. As shown in Fig.4, the VFOA labels were Robot (NAO), Partner, Painting1, Painting2, Painting3, and in addition, Others, that we used when people were looking elsewhere.

Performance Measure: We uses the Frame based Recognition Rate (FRR), that is, the percentage of frames where the VFOA has been correctly recognized.

Algorithms: For each datasets, 3 models were tested. The baseline is the HMM model with a fixed reference value. G1 uses the head pose average R_t^0 as reference, and sets the

Table 1: Performance on the Meeting data

Person	Training	Baseline	Model G1	Model G2
Person on left	same seat	64.7	65.7	**68.4**
Person on left	other seat	64.5	66.7	**69.3**
Person on right	same seat	57.0	**58.7**	58.5
Person on right	other seat	43.9	**59.0**	57.8

head pose using the new formula to account for the midline effect. G2, in addition to using R_t^0, implements gaze shifts by using the previous gaze.

3.2 Parameter Setting

Meeting Data. We set the gaze target directions from the physical setting and the Gaussians variances as in [2]. The reference direction for the baseline was set as the middle between VFOA gaze target as in [2]. The remaining parameters were adjusted by cross-validation for each model, by considering two different set-ups: cross validation using training data either from the same seat, or from the other seat. The second case is useful to see whether our model is sensitive to a specific setting or it is more general.

Nao Data. For **D1 and D2** the gaze directions were defined from the geometrical setting. The reference direction for the baseline was set as looking at Nao, which is a reasonable choice in an HRI scenario. Standard deviations of Gaussian were set to 10 and 8 for pan and tilt. Other parameters were adjusted by leave-one-out cross validation on the three participants in D1, and on the 6 sequences (2 participants in 3 recordings) for D2.

3.3 Results

Meeting data. Table 1 shows the results of the three models. The first model outperforms the baseline, particularly in more mismatched conditions, when parameters are learned from the other seat, exhibiting therefore a better adaptation capacity. The main (mismatched) parameters leading to the degradation is the parameter α of the gaze model (see Eq. 3) 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. The different values of α_{pan} obtained from two different seats could be due to the fixed choice of the reference which leads to different values for these two settings. This effect does not exist for the first model G1 and the chosen parameters through cross-validation are completely consistent with an optimal value for both seats around 0.7 for α_{pan}. On the other hand, we can see that G2 performs better than the G1 in most cases. We observed that the improvement happened mainly during consecutive gaze shifts involving stable head pose changes, and was also observed in the Nao case.

Nao Data. For **D1** the results are summarized in Table 2. Despite the quite different setting, the conclusions are similar to the meeting data. However, model G1 outperforms the baseline with a larger difference. This is particularly true for the first person, who, being more dynamic during the interaction, shifted her body orientation towards both the robot and the other participants, wheras the two other people remained more firmly seated in the sofa and oriented towards Nao, which better matches the looking at Nao assumption of the baseline. Also, model G2 performs better than model G1 for all of the sequences. Table 3 contains the results for **D2**. The same conclusion than with D1 can be drawn: G1 outperforms the baseline with a noticable difference and G2 is performing slightly better than G1.

Table 2: Performance First Nao Data (D1)

Person	Baseline	Model G1	Model G2
Person1	49.78	64.36	**65.32**
Person2	91.52	93.73	**95.45**
Person3	67.83	66.12	**68.01**

Table 3: Performance on Second Nao Data (D2)

Person	Baseline	Model G1	Model G2
Person on right	54.4	59.0	**59.2**
Person on left	55.9	59.1	**60.9**

When looking at the parameters selected by cross-validation on the 3 datasets for model G1, most of the time the value of α_H was 0: the main improvement with G1 was thus due to the use of an adaptive reference, rather than to the midline effect illustrated in Fig. 1b.

4. CONCLUSION.

We investigated two drawbacks of the previous work for mapping vfoa targets to head pose values. To overcome the problem with the fix reference (midline) we provided an implicit estimate of it. To account for the dependency of the gaze-to-head mapping to the previous head or gaze before the shift we proposed two approaches inspired from human behaviors [5]. Experiments on meeting benchmark data and on two datasets recorded by a humanoid robot shows the benefit of several modeling components.

As future work, using image-based gaze directions [4] would be beneficial, and can be combined with our approach. However, this is true only if they can be extracted sufficiently reliably from the available images.

Acknowledgments. This work was supported by the European Community's Seventh Framework Programme through the HUMAVIPS project (Theme Cognitive Systems and Robotics, Grant agreement no. 247525). We also thank Vasil Khalidov for his help and fruitful discussions.

5. REFERENCES

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

[2] S. O. Ba and J.-M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Trans. Sys. Man Cyber. Part B*, 39:16–33, February 2009.

[3] D. Bohus and E. Horvitz. Facilitating multiparty dialog with gaze, gesture and speech. In *ICMI*, 2010.

[4] S. Gorga and K. Otsuka. Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *ICMI*, 2010.

[5] D. A. Hanes and G. McCollum. Variables contributing to the coordination of rapid eye/head gaze shifts. *Biol. Cybern.*, 94:300–324, March 2006.

[6] S. R. Langton, R. J. Watt, and I. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends Cogn Sci*, 4(2):50–59, feb 2000.

[7] 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 *ICMI*, 2005.

[8] E. Ricci and J.-M. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *ICIP*, 2009.

[9] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *ICMI*, Oct. 2008.

[10] Z. Yücel and A. Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *Int. Conf. on Affective Computing and Intelligent Interfaces*, 2009.