Leveraging the Robot Dialog State for Visual Focus of Attention Recognition

Samira Sheikhi
Idiap Research Institute
samira.sheikhi@idiap.ch

Vasil Khalidov
Idiap Research Institute
vasil.khalidov@idiap.ch

David Klotz
Bielefeld University
Germany

Britta Wrede
Bielefeld University
Germany
bwrede@techfak.uni-bielefeld.de

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

ABSTRACT

The Visual Focus of Attention (what or whom a person is looking at) or VFOA is a fundamental cue in non-verbal communication and plays an important role when designing effective human-machine interaction systems. However, recognizing the VFOA of an interacting person is difficult for a robot, since due to low resolution imaging, eye gaze estimation is not possible. Rather, head pose cue is used as a substitute for gaze, but leads to ambiguities in its interpretation as VFOA indicator. In this paper, we investigate the use of the robot conversational state, which the robot is aware of, as contextual information to improve VFOA recognition from head pose. We propose a dynamic Bayesian model that accounts for the robot state (speaking status, person he addresses, reference to objects) along with a dynamic head-to-gaze mapping function. Experiments on a publicly available human-robot interaction dataset, where a humanoid robot plays the role of an art guide and quiz master, shows that using such conversational context is effective in improving VFOA.

Categories and Subject Descriptors

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

Keywords

HRI, HCI, Gaze recognition, VFOA, Dialog Context

1. INTRODUCTION

Gaze is an important non-verbal cue with many functions in human interaction and discourse regulation [6], and can therefore play an important role 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) and a good cue to monitor people engagement [2]. For instance, in [4], the head pose is used to model predefined state-of-interactions of a customer in a bartender application. In our scenario (see Fig. 1), a humanoid robot (Nao) is used as an art guide, proposing and explaining different paintings surrounding him, and ultimately gives a quiz to participants. In this situation, VFOA can be used for addressee recognition as well as to monitor whether people follow the conversation (are people looking at the painting I am currently explaining?) or evaluate their level of interest.

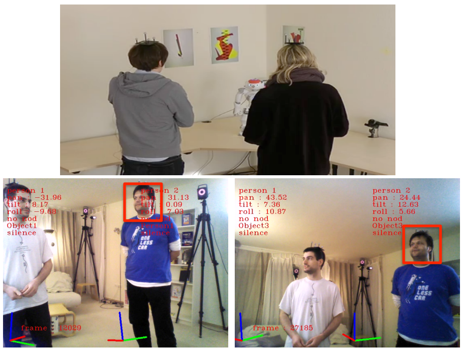

Figure 1: Vernissage data (top): The robot explains 3 groups of paintings to the participants, and then gives them a quiz. Our task is to monitor people attention, i.e. recognize whether they look at Nao, the other person, paintings, or elsewhere. Pose ambiguity (bottom): two frames from the video as seen by Nao. The same head pose is used for looking at the other person (left) and at the 3rd painting (right).

Motivation and related work. Recognizing the VFOA of people is however a difficult task. As standard sensor-based gaze tracking technologies can not often be applied, researchers have considered head pose as main gaze information [10, 1, 4, 5]. Head poses, however, are ambiguous: in realistic and dynamic scenarios, the same pose can be used to look at different targets, depending on the situation (see Fig. 1). To remove this ambiguity, researchers have explored two directions: (i) the use of other social cues, leveraging on the fact that the recognition of non-verbal cues should not be done in isolation, but jointly, as some behaviors provide context to others. In human-human interaction (mainly meetings), examples include speaker information [10] or higher conversational states [5], that can be complemented with group activity [1]; (ii) the improvement of the prediction of the gaze direction (including eye information) from the

ICMI'13, December 9–13, 2013, Sydney, Australia

ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522881.

head pose, allowing a better association of a head pose with looking at a given target [9].

While in human-human analysis applications most social cues to be used as context have to be inferred from the data and might suffer from being noisy, in the robotic or Embodied Conversational Agent (ECA) cases, the agent is fully aware of its own conversational acts. Such information can thus be more conveniently exploited to better interpret the non-verbal cues performed by interacting people. For instance, in [8], different types of features (lexical, timing, gesture displayed) performed by an ECA are exploited within a supervised learning framework to predict head nods and head shakes in combination with a vision-based head gesture recognizer. However, to our knowledge, while estimating the VFOA is considered by several systems [2, 4], the use of the robot dialog context to improve the recognition of a user attention (VFOA) has not been explored in the past.

Contributions. To improve VFOA recognition, we propose to leverage on two types of robot dialog acts that can affect VFOA expectations: communicative acts (people look more at speakers; and this is particularly true from addressed persons), and lexical (implicit or explicit references to scene object(s)). We propose an Input-Output Hidden Markov Model (IO-HMM) to integrate such information in a recognition process that also leverages on a better head-pose to gaze dynamic mapping process [9]. Experiments are conducted on a recent HRI database with available ground-truth [3] featuring natural human robot interactions. Results using either head pose ground truth or pose estimated by an automatic tracking algorithm show the viability of the method.

2. VFOA RECOGNITION MODEL

We formulate the recognition problem as the estimation of the VFOA state $F_t \in \mathbb{F}$ of a given participant at each time t, with $\mathbb{F}$ defined as $\{Nao, partner, pai_1, pai_2, pai_3, other\}$, where pai_j refers to painting number j and $other$ stands for VFOA that is not attributed to any other label (see Fig. 1). Below we define the conversation context as derived from the robot and then present our contextual recognition model.

2.1 Robot Conversation Context

Given our task, the question is which of the robot actions affect people VFOA, and how? In interactions, these mainly relate to the communication functions of gaze and their relationships with speaking turns [6]. However, it is also known that objects that play a central role in the conversation may attract the attention, whereby overruling the communication patterns [11] observed in natural conversation. In our art guide scenario this corresponds to physical locations in the room and particularly paintings. We thus defined the robot interaction context as described below.

Speaking context. Since people look more at speakers than at non-speakers, we defined a speaking context state variable $s_t \in \{0, 1\}$ as whether Nao speaks or not at time t.

Addressee context. It is known that speakers monitor their addressees' attention by gazing at them, and expect gaze in return [6]. We thus defined the addressee context $a \in AC = \{pers_1, pers_2, group\}$ of a speech segment the situation when the robot addresses the first person, the second person, or both. In our data, this context is automatically derived from the dialog system, which is aware of who is addressed (either a person, or a group) along with the way to address them, which in our set-up was accomplished for a given individual by naming him and turning the head towards him, or by directing the head in between participants when both persons were addressed. VFOA statistics depending on the addressee status are shown in Fig. 2, during the robot speech ($x = 0$) or x seconds after the end of the speech. In spite of the noise, we can notice that addressed people tend to stay more in visual contact with the robot, while non-addressed people disengage quicker to look at the other person or elsewhere. There is overall no strong temporal variation of VFOA probabilities (after the utterance), so to avoid overfitting, it is reasonable to assume a constant model for x>0. We defined the addressee context state a_t at t as the addressee context derived from the current (if $s_t = 1$) or preceeding (if $s_t = 0$) robot utterance.

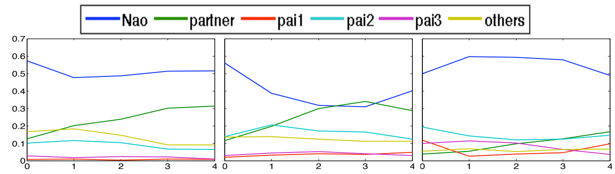

Figure 2: VFOA statistics of an individually addressed person (left), a non-addressed person (middle), and addressed person, when both persons are addressed (right). The x axis denotes the time since the end of the robot's utterance. The statistics for $x = 0$ are collected during the robot's utterance.

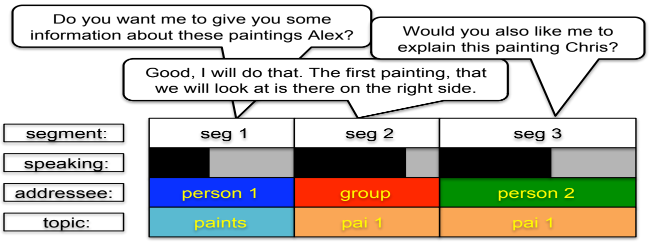
Do you want me to give you some information about these paintings Alex?
Would you also like me to explain this painting Chris?
Good, I will do that. The first painting, that we will look at is there on the right side.
segment:
speaking:
addressee:
topic:
seg 1
seg 2
seg 3
person 1
group
person 2
paints
pai 1
pai 1

Figure 3: Illustration of the context assignments.

Topic context. Given our scenario, the topic context set is defined as $OC = \{pai_1, pai_2, pai_3, paintings, none\}$ corresponding to whether the robot informs or refers to a specific painting, all paintings, or none of them. The topic context state $o_t \in OC$ at t is thus defined as topic context of the robot utterance that preceeds t.

Overall conversational context C_t. As a summary, at each instant t the different context states s_t, a_t and o_t are automatically assigned according to the spoken utterances and temporal segments, as illustrated in Fig. 3. The final context state C_t is then defined as the Cartesian product of all contexts, i.e. $C_t = (s_t, a_t, o_t)$, and will influence the VFOA recognition as explained in the next Section.

2.2 Conversation Aware VFOA Recognition

To address VFOA recognition, we propose the IOHMM graphical model of Fig. 4. Broadly speaking, the middle part (box) shows the main process, which models how the sequence of VFOA states generates a sequence of head poses $H_t \in \mathbb{R}^2$ (represented by pan and tilt angles). This process is affected in two ways: by the gaze-head mapping model in the bottom part, whose goal is to dynamically predict the expected head pose μ_t^h for each VFOA target; and by the conversation context C_t (top part). More precisely, the VFOA is inferred by maximizing the posterior probability

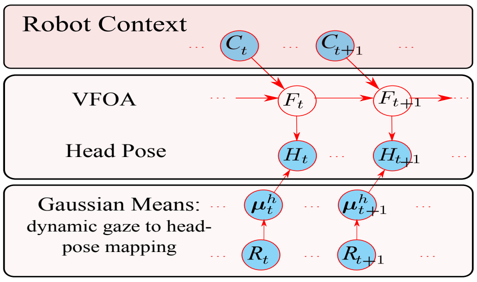

Figure 4: VFOA recognition from head pose.

of the sequence of VFOA states $F_{1:t}$ given all observed variables: head pose $H_t \in \mathbb{R}^2$ and context C_t. The posterior for the graphical model of Fig. 4 is expressed as:

$$p(F_{1:t}|H_{1:t}, C_{1:t}, \mu^h_{1:t}, R_{1:t}) \propto \prod_{t=1:t} p(H_t|F_t, \mu^h_t) p(F_t|F_{t-1}, C_t)$$

$$\text{with } p(H_t|F_t = f, \mu^h_t) = \mathcal{N}(H_t|\mu^h_t(f), \Sigma_H(f)) \qquad (1)$$

$$\text{and } p(F_t|F_{t-1}, C_t) \propto p(F_t|F_{t-1}) p(F_t|C_t) \qquad (2)$$

where the different terms are explained below.

Data likelihood. The term in Eq. 1 represents the likelihood of an observed head pose for a given focus, and is modeled as a Gaussian distribution with mean $\mu^h_t(f)$ and variance $\Sigma_H(f)$. The means μ^h play a crucial role for VFOA recognition, as they represent the expected head pose for looking at each target. Often, researchers assume a fixed setting, with people facing the camera at a given distance and set time-independent means manually or through learning [2, 4]. In this work, we follow the approach of [9] that leverages on body-head-gaze behavioral studies and better accounts for natural gaze shifts. Accordingly, the means were set dynamically as a combination of the direction in which the person should gaze to look at a target, and of the body orientation R_t (estimated as a proxy through the average of the past head poses). See [9] for more details.

Contextual prior. Eq. 2 denotes the prior on the focus, which we assumed can be decomposed in two parts. The first one is the temporal prior $p(F_t|F_{t-1})$, which allows temporal smoothing by setting large probabilities to stay in the same state and equal low probability to transit to other states. The second one $p(F_t = f|C_t = c) = B_{cf}$ denotes our Robot context prior which affects the recognition by altering the expectations about what people look at depending on the context, and is parameterized by the probability tables B.

Learning the context tables. There are several ways to set the tables, depending on the goals and assumptions. Here, we use a learning approach, with smoothing to handle the lack of data for some contexts, and further modeling assumption to avoid data overfitting and better capture the generalization capabilities of the model.

- Given a training dataset, we gather the VFOA data $D_c = \{f_i\}$ observed under each given context c. Then, using a Maximum A Posteriori approach with a conjugate Dirichlet prior (i.e. maximizing $p(B_{c\cdot}|D_c) \propto p(D_c|B_{c\cdot})Dir(B_{c\cdot}|\alpha)$), the table entries are defined as $B_{cf} \propto n_f + \alpha_f$, where n_f denotes the number of occurrences of the focus f in D_c, and the Dirichlet prior parameters are set as $\alpha_f = 0.1 N_f/(K \times N_C)$, where N_f, K and N_C denote the number of observation in the whole training set, the number of VFOA targets, and the number of contexts, respectively. In other words, the prior corresponded to the addition of virtual observations equally spread among table entries and amounting to 10% of the total number of real observations.

Table 1: Sample context probability priors (using only the topic context) showing parameter tyings.

Context	Nao	partner	pai_1	pai_2	pai_3	others
pai1	0.33	0.03	0.53	0.04	0.04	0.03
pai2	0.33	0.03	0.04	0.53	0.04	0.03
none	0.58	0.17	0.04	0.04	0.04	0.12

- Priors learned using the above scheme might overfit the specific setup. In particular, the painting positions or the duration of references and explanations about each of them lead to the gathering of different statistics for each painting. To be more general, we applied parameter tying, enforcing that all table entries involving paintings which play the same role should be the same, as illustrated in Table 1 with some sample context probability priors.

3. EXPERIMENTS AND RESULTS

Data. We used the Vernissage dataset [3], containing natural interactions recorded using a Wizard of Oz approach. In each recording, that lasts around 10 minutes, Nao first engages with two participants and explains them three paintings. He then gives them a quiz in which participants could discuss before the person to whom a question was addressed gave the answer. Both parts are approximately of equal duration. Note that people were free to walk around, and that some of the questions (4 out of 10) referred to paintings in the room. This dataset consists of 10 sequences and VFOA is fully annotated for all participants.

As head poses, we used both measures derived from Vicon (a motion capturing system) data and estimates obtained by applying a joint head tracking and pose estimation algorithm [7]. After inspection, the head pose Vicon measures of one sequence happened to be inconsistent in time (the head-bands attaching the Vicon markers to people head might have moved), and we dropped it. Pose estimated from video were obtained by applying a particle filter tracker with appearance head pose modeling [7]. However, since in this dataset Nao is performing head gestures -pointing to paintings, rotating the head to address people, nodding- that greatly affects the video quality (with people disappearing from the field of view, frequent lighting changes, etc.) results were not very accurate. Since our goal is to evaluate VFOA performance under reasonable head pose estimation, the tracker output was filtered by keeping only track segments that matched the (sparse) ground truth location available in the dataset [3], and results with too large average pose errors or no sufficient tracker recall were removed. This resulted in a dataset of 13 persons, amounting to around 100 minutes of data.

Experimental setup. To evaluate the contribution of the different contexts, we considered different settings: No context (baseline), one single context cue (speaking, addressee, or topic), and all cues together. In addition, we experimented without or with (Static and Dynamic settings, respectively) the dynamic model (as explained in Sec. 2.2) for head pose prediction [9]. The static case was obtained by using the reasonable assumption that people were facing the robot (body orientation reference set to 0). The dynamic setting allows us to investigate whether the con-

Table 2: VFOA recognition with Vicon head poses.

	Static Setting			Dynamic Setting		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	52.6	54.9	51.7	65.8	67.8	64.9
Speak.	61.2	62.3	60.8	67.3	63.2	68.9
Addr.	61.6	63.8	60.8	68.1	64.3	69.5
Topic	63.9	67.4	62.5	71.3	73.5	70.2
All.	65.0	68.0	63.8	72.3	74.2	71.4

Table 3: VFOA recognition with tracker head pose estimates.

	Static Setting			Dynamic Setting		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	61.7	56.3	63.2	62.0	58.9	62.8
Speak.	63.9	56.7	65.9	63.6	58.0	65.0
Addr.	64.6	57.9	66.5	64.1	59.1	65.4
Topic	66.7	66.0	66.9	66.4	67.5	66.0
All.	67.0	66.8	67.2	66.8	68.1	66.3

text is still useful when more accurate gaze-to-head pose predictions are made. In all experiments, the parameters of the HMM model (variances, temporal transition priors) were the same and set as in [9]. Context tables were learnt for each participant through leave-one-out cross-validation on data from all the other participants. Finally, we used as performance measure the Frame based Recognition Rate (FRR), defined as the percentage of frames during which the VFOA has been correctly recognized. Performance was reported for the full recordings, or separately on the first part (explanation part) of the recording, or on the quiz part.

Results and Discussion. Table 2 shows the results obtained when using the head poses derived from the Vicon.

Looking first at the static setting commonly used by researchers, we can notice the following. Despite accurate head poses, the result of the baseline is only of 52%, showing the difficulty of the task. Most confusion comes from looking at Nao and at the group of paintings above him, as well as looking at the partner versus at painting in a similar direction (like in Fig. 1). The performance improves whatever individual cue we consider. The increase is larger when using the topic context, in particular during the first part of the interaction when Nao makes regular reference to the paintings. Alltogether, the use of all context cues brings a considerable improvement of more than 12%.

With the dynamic setting the baseline already produces a recognition rate of 65.8%, which is more than in the static case with context. Still, even in this case the context improves the results with a gain of 6.5% when using all cues. Interestingly, the results with individual cues exhibit different behaviors depending on the interaction phase. As can be seen, the communication cues (speaking, addressee) which emphasize Nao or people as VFOA prior increase performance during the quiz, which is more interactive, but decrease by around 4% the performance during the painting explanations. However, when using all cues, the performance is higher in all situations, demonstrating their complementarity for the VFOA recognition task.

Finally, Table 2 presents the results obtained with automatic head pose tracker estimates. They are not comparable to those obtained with Vicon pose since the set of sequences are slightly different, and the tracker misses around 15% of the heads (particularly those with profile poses). Nevertheless, we can notice that i) resultats are quite good overall, despite the task complexity, and comparable to using the Vicon poses; ii) the dynamic pose modeling does not help much, which can be partly explained by the missed profile poses; iii) the context (esp. the topic one) improves results in all situations and the conclusions made in the Vicon case hold true as well here.

4. CONCLUSION.

We proposed a contextual VFOA recognition model exploiting a robot's (or ECA's) gaze-related conversational context. As context, we relied on communicative cues –the robot's speaking status, addressee– as well as topical cues referring to object in the scene. Experiments with Vicon and estimated head pose data on a challenging dataset containing natural interaction between people and a humanoid robot acting as an art guide demonstrated the usefulness of the approach and the cue complementarity to remove pose ambiguities. Note that the method can be used with a different number of artworks: topic context statistics can still be used appropriately. Performance will probably decrease with more artworks since it increases the VFOA confusion, but the drop would be even more important without our approach. More generally, the method can be used with any other scenarios implying objects which the robot is aware of and that he can reference to in the conversation.

5. ACKNOWLEDGMENTS

The work was supported by the EU HUMAVIPS project and Swiss National Science Foundation through the IM2 NCCR on Interactive Multimodal Information Management.

6. REFERENCES

[1] S. Ba and J.-M. Odobez. Multi-person visual focus of attention from head pose and meeting contextual cues. *IEEE PAMI*, 2011.

[2] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *Proc. of the SIGDIAL 2009 Conference*, 2009.

[3] D. Jayagopi et. al. The vernissage corpus: A conversational human-robot interaction dataset. In *Int. Conf. on Human-Robot Interaction*, march 2013.

[4] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. de Ruiter, and A. Knoll. Modelling state of interaction from head poses for social human-robot interaction. In *Proc. of the Gaze in Human-Robot Interaction Workshop, HRI*, 2012.

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

[6] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)*, 26(1):22–63, 1967.

[7] V. Khalidov and J. M. Odobez. Real-time multiple head tracking using texture and colour cues. In *Idiap research report*, 2013.

[8] L.-P. Morency, C. L. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Int. Conf. on Multimodal Interfaces*, pages 18–24, 2005.

[9] S. Sheikhi and J.-M. Odobez. Investigating the midline effect for visual focus of attention recognition. In *Int Conf. on Multimodal Interfaces*, oct 2012.

[10] 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.

[11] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Int. Conf. on Multimodal Interfaces*, 2005.