ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

# Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions

Samira Sheikhi[a], Jean-Marc Odobez[a,**]

[a]*Idiap Research Institue, Rue Marconi 19, 1920 Martigny, Switzerland*

ARTICLE INFO

*Article history*:
Received 25 Feb 2014

*Keywords:*
Attention recognition
Gaze
Head pose
Non-verbal behavior
HRI
Context

ABSTRACT

The ability to recognize the Visual Focus of Attention (VFOA, i.e. what or whom a person is looking at) of people is important for robots or conversational agents interacting with multiple people, since it plays a key role in turn-taking, engagement or intention monitoring. As eye gaze estimation is often impossible to achieve, most systems currently rely on head pose as an approximation, creating ambiguities since the same head pose can be used to look at different VFOA targets. To address this challenge, we propose a dynamic Bayesian model for the VFOA recognition from head pose, where we make two main contributions. First, taking inspiration from behavioral models describing the relationships between the body, head and gaze orientations involved in gaze shifts, we propose novel gaze models that dynamically and more accurately predict the expected head orientation used for looking in a given gaze target direction. This is a neglected aspect of previous works but essential for recognition. Secondly, we propose to exploit the robot conversational state (when he speaks, objects to which he refers) as context to set appropriate priors on candidate VFOA targets and reduce the inherent VFOA ambiguities. Experiments on a public dataset where the humanoid robot NAO plays the role of an art guide and quiz master demonstrate the benefit of the two contributions.

## 1. Introduction

### *1.1. Task and motivation*

Endowing human-computer interaction (HCI) or human-robot interaction (HRI) systems with social skills relies on advances of technologies in several areas including speech recognition and synthesis, dialog and interaction modeling, and human behavior perception and situated scene analysis that go beyond people localization and speaking status determination.

In this paper, we address the recognition of gaze and more precisely its interpretation in terms of VFOA in HRI or Embodied Conversational Agent (ECA) settings. Amongst the behaviors exhibited during interactions, gaze is one of the most important. It is a non-verbal cue which has many functions in communication such as establishing relationships, expressing intimacy, or exercising social control; its role in discourse regulation has been well documented (Kendon, 1967). In particular, gaze is a good indicator of the addressee (to whom a person is speaking) which is an important information to know for robots or ECAs interacting with multiple people. Due to this important role, gaze has often been used for turn-taking management and at a higher level to recognize a user's predefined interaction state (Foster et al., 2012) or monitor people engagement and intention (Bohus and Horvitz, 2009). Besides conversation, gaze can be used as a social skill, for instance to assign importance to different people and decide how to share the robot attention on them (Bennewitz et al., 2007), or monitor the attention of people towards objects related to the task they perform or verbally referenced in the conversation (Cooper, 1974) and (re)act appropriately. For instance, in ECA based city trip planning applications, the user's gaze to different map locations can be used

[**]Corresponding author: Tel.: +41- 27 721 77 26.
*e-mail:* `odobez@idiap.ch` (Jean-Marc Odobez)

alone or in combination with other cues like nods to manage dialog content or the conversation state (Nakano et al., 2003). In another direction, several works have proposed computational model of the emergence and learning of shared attention mechanisms in human infants (Triesch et al., 2006). For instance, (Nagai et al., 2006) implemented on a robotic platform a cognitive model for the learning of visuomotor gaze following skills based on the appropriate perception of head pose images of a caregiver as well as adaptive feedback from this caregiver.

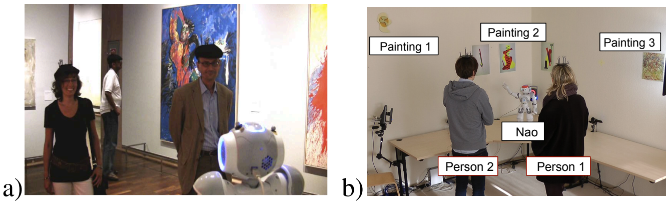

**Fig. 1. Vernissage scenario: a robot acts as an art exhibition guide, providing explanations about artworks placed around it, and giving a quiz about these artworks and more general art and culture topics. a) Real ste-up. b) Vernissage dataset considered for evaluation.**

### *1.2. Approach and related work*

Measuring and interpreting the gaze of people is however a difficult task. Eye tracking devices can be used but are usually expensive, considered as intrusive, or not applicable for natural interaction analysis. Some works successfully used simple attentional cues like torso orientation (Bennewitz et al., 2007). Nevertheless, benefiting from advances in computer vision, researchers have mainly considered head pose as an approximation of the gaze (Nakano et al., 2003; Foster et al., 2012), a trend that should increase with the new Kinect camera and API that directly delivers this information. However, while this approach is supported by both behavioral modeling (Langton et al., 2000) and empirical evidence, interpreting the head pose as looking at VFOA target remains ambiguous since in realistic scenarios, the same pose can be used to look at different targets. Two directions need to be explored in conjunction to solve this ambiguity. We discuss both of them below, presenting relevant related work and introducing then our contributions.

**Head pose-VFOA gaze direction association.** A central issue when designing recognizers like Hidden Markov Models (HMM) to decode the sequence of VFOA targets given the head pose sequence is the following: what is the expected head pose of a person who looks at a given VFOA target? Indeed, in gazing behaviors, the difference between a gaze direction and the head pose used to look in that direction, which is due to the missing eye information, can not be considered as a random noise with zero mean. Rather, it is often biased, and the bias depends on several factors related to the body, head and eye dynamics, as discussed below. In spite of its importance, the above issue has seldom been addressed in the past. Some methods rely on manual setting, potentially followed by adaptation (Otsuka et al., 2005). Others like Foster et al. (2012) use training data to directly infer VFOA from head pose without defining gaze as an intermediate step. Learned parameters, however, are then specific to the geometric configuration between the sensor (robot), person, and VFOA targets, and thus such an approach is not suitable for robot dealing with moving people.

As one of the few works addressing this problem, Ba and Odobez (2009) Exploited results on gazing behavior and head-eye dynamics involved in gaze shifts (Langton et al., 2000; Hanes and McCollum, 2006) and introduced a linear gaze model relating the head pose, gaze direction, and body orientation through a head-to-gaze ratio (see Fig. 3). While the method worked when applied to meetings, it suffered from two drawbacks: the body orientation was assumed to be fixed and set according to the setup, an assumption that might not hold in more dynamic settings from HCI or HRI where the robot is not always the main focus and people are free to move and re-orient themselves, as illustrated in Fig. 11. The second drawback, pointed out in several psychovisual works, is that the mapping not only depends on the the gaze direction and body orientation, but also on the head or gaze direction before the shift. To overcome some of the above limitations, Voit and Stiefelhagen (2008) who addressed VFOA recognition in a room from external sensors, proposed ad-hoc differential head pose indicators to select more appropriate head-to-gaze ratios and constrain the succession of gaze shifts, but without accounting for the body orientation or rely on documented human behavior models.

**Contextual recognition.** A second way of resolving head pose ambiguities is to rely on other social cues leveraging on the fact that some behaviors provide context to others. In human interactions, examples for VFOA recognition include speaker information (Stiefelhagen et al., 2002) or higher conversational states (Gorga and Otsuka, 2010), that can be complemented with group activity (Ba and Odobez, 2011). While in these cases the context have to be inferred from the data and might be noisy, in the robotic or ECA cases, the agent is fully aware of its own conversational acts, allowing them to be conveniently exploited to better interpret the non-verbal cues performed by interacting people. For instance, in Morency et al. (2005), 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 (Bohus and Horvitz, 2009; Foster et al., 2012), the use of the robot dialog context to improve the recognition of a user attention (VFOA) has not been explored in the past.

### *1.3. Contributions*

We propose a novel Input-Output HMM (IO-HMM) combining the two above approaches to improve VFOA recognition. As for the first one, our model relies on a time-varying and implicit estimation of the body orientation to implement dynamic gaze-to-head mapping and gaze shift models inspired by Hanes and McCollum (2006) which in our HRI scenario are shown to considerably improve the accuracy of the predicted head pose used to look at VFOA targets, and VFOA recognition as a consequence. As for the second one, we benefit from the HRI context by exploiting two robot dialog act types that influence VFOA expectations: communicative acts (people look more at speakers, including the robot) and verbal acts (references to scene objects). Experiments on a large dataset (140 minutes) featuring natural human robot interactions demonstrate the complementary benefit of the two modeling steps.

The paper is organized as follows. Section 2 provides an overview of the approach, while the Sections that follow describe the baseline algorithm (Sec. 3), the novel gaze dynamical mapping (Sec. 4), and the contextual model (Sec. 5). Section 6 introduces the experimental set-up, while Sec. 7 discusses in details the results. Section 8 concludes the paper.

## 2. Approach Overview

Our objective is to monitor the visual attention of people in a given environment relying on head pose. We thus assume to have a specific set of visual targets $\mathbb{F}$ which are of interest in the given context, and would like to recognize which of these targets a given person is looking at.

To illustrate the above, the main robotic setup which we have considered is based on the Vernissage scenario shown in Fig. 1. Recognizing what or whom people are looking at in this context gives useful information about their attention to the robot or paintings and hence whether they follow the explanation or not which could be used to decide how to proceed in the conversation. More specifically, our experiments will be conducted on the Vernissage dataset (Fig. 1b). In this case we define $\mathbb{F}$ as $\{Nao, partner, pai_1, pai_2, pai_3, other\}$, where $Nao$ refers to the robot, $pai_j$ refers to painting number $j$ and $other$ stands for VFOA that is not attributed to any other label (see Fig. 1b).

Fig. 2 illustrates the recognition approach. The middle part (box) shows the main recognition process, which consists of an HMM allowing to decode the sequence of head poses $H_t$ in terms of VFOA states $F_t \in \mathbb{F}$. The head pose $H_t = (H_t^{pan}, H_t^{tilt}) \in \mathbb{R}^2$ is represented by the pan and tilt angles characterizing the left-right and up-down head rotations, as illustrated in Fig. 3a. The roll angle was left-out since it does not bring information regarding the gaze direction. This process is affected in two ways. First, by the gaze-head mapping model shown at the bottom, whose goal is to dynamically predict at each instant $t$ the expected head pose $\mu_t^h = (\mu_t^{h,pan}, \mu_t^{h,tilt})$ used to look at each VFOA target, as addressed in Sections 3 and 4. It is designed to reflect the findings from studies on gazing behavior related to the coordination of the body, head and eyes in gaze shift. Secondly, as shown at the top of Fig. 2, by leveraging contextual information aiming at removing ambiguities introduced by relying on noisy head poses measurements rather than gaze. Given our robotics setup, contextual cues are extracted from the robot's conversational acts as discussed in Section 5.

## 3. Baseline: HMM with Geometrical Mapping

We build our VFOA recognition model based on the HMM illustrated in the middle part of Figure 2, without exploiting the context at this stage. In this model, the distribution of head poses associated to a given VFOA target is represented by a Gaussian distribution, whereas transitions between VFOA targets are represented by a transition matrix $A$. More specifically, let $\mu_t^h(f) \in \mathbb{R}^2$ and $\Sigma_H(f) \in \mathbb{R}^4$ denote the mean and covariance of the Gaussian associated with target $f$. The HMM equations can be written as:

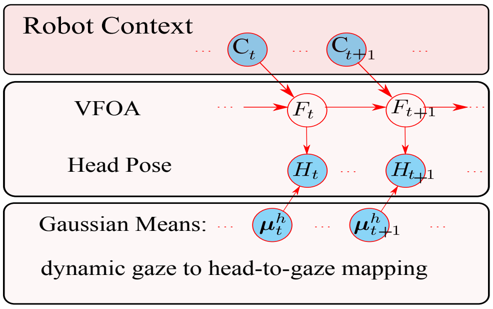

**Fig. 2. VFOA recognition from head pose. The robot conversation context $C_t$ appears as an input observation and provides expectations about which VFOA should be observed. At the bottom, a gaze-head mapping module dynamically monitors the expected head pose associated with each VFOA target.**

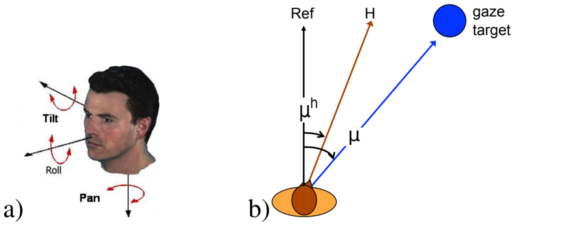

**Fig. 3. a) Head pose euler angles. b) Geometrical Gaze Model. The person is assumed to be looking at the reference direction, or midline (body orientation). Then, looking at a gaze target is accomplished by rotating both the head and eyes, the head part being a fixed fraction of the full gaze rotation.**

$$p(H_t|F_t = f, \mu_t^h) = \mathcal{N}(H_t|\mu_t^h(f), \Sigma_H(f)) \quad (1)$$

$$p(F_t = f|F_{t-1} = \hat{f}) = A_{f\hat{f}} \quad (2)$$

Parameter setting plays an important role for recognition. Following previous works, gaussian covariances can be set to reflect target sizes and/or head pose estimation variability. In absence of other information, the temporal prior $p(F_t|F_{t-1})$ can be used to satisfy our expectation of having smooth VFOA sequences by setting in $A$ large probabilities to stay in the same state, and equal low probability to transit to other states.

However, although they play the most important role in the model, setting the Gaussians means $\mu_t^h$ is not easy. As discussed in the introduction using training data is not an option since annotation needs to be gathered for each configuration of the observer, targets and settings. This is especially problematic if people are free to move.

A solution to overcome the above difficulty is to use gaze models derived from cognitive findings in gazing behavior (Langton et al., 2000). Accordingly, gazing at a target is accomplished by rotating both the eyes ('eye-in-head' rotation) and the head as illustrated in Fig. 3b) for the pan. More precisely, as a first approximation, $\mu_t^{hb}(f)$[1], the mean of the Gaussian to look at target $f$ can be set as a fixed linear combination

[1] We will denote by $\mu^{hx}$ the mean $\mu^h$ set using the model $x$.

of the gaze and *head reference* directions:

$$\mu_t^{hb}(f) - R_0 = \alpha \star (\mu_t(f) - R_0) \Rightarrow \mu_t^{hb}(f) = \alpha \star \mu_t(f) + (1_2 - \alpha) \star R_0 \quad (3)$$

where $\star$ denotes the component wise product, $1_2 = (1, 1)$, $\alpha = (\alpha^{pan}, \alpha^{tilt})$, $R_0 \in \mathbb{R}^2$ denotes the reference direction, and $\mu_t \in (\mathbb{R}^2)^K$ the directions of the given $K$ targets which are assumed to be known. The head-to-gaze ratio for the pan, $\alpha^{pan}$, is usually set between 0.5 and 0.7, and between 0.3 and 0.5 for the tilt. Our baseline thus consists of the above HMM model where the reference $R_0$ is set to a constant value and the head pose mean for looking at target $f$ is set using Eq. 3.

## 4. Gaze to head dynamical mapping

The baseline geometrical model has been useful in static scenarios like meetings. There, since people are seated and do not move their bodies extensively, setting the reference direction $R_0$ as the middle of the target directions has been a good solution (Ba and Odobez, 2009). When the participants upper body and shoulders exhibit more dynamics, the baseline model becomes inaccurate since having a static body reference becomes an unrealistic assumption. Furthermore, this head-gaze mapping model was originally studied with discrete gaze shifts when the person gazed intentionally at a given direction from the reference (Langton et al., 2000) (cf. Fig. 3b). Therefore, it might not be sufficient when the person is continuously looking at different targets in a natural conversation. In this view, exploiting past head poses or gaze directions could be useful for obtaining dynamic and more precise predictions of the head poses used to look at a given target at the current instant. In the following, we introduce three models going in that direction.

### *4.1. Model G1: Dynamical Head Reference*

Setting the Gaussians means using the geometrical model requires the knowledge of $R_0$ and of the target directions. Eq. 3 shows the importance of the reference: using a wrong value for $R_0$ shifts the mean values $\mu_t^h(f)$ for all targets $f$ simultaneously, which can have dramatic effects for head pose interpretations. The importance of knowing the head reference (shoulder orientation) is also illustrated in Fig. 11a). Unless it is constrained by the setting (e.g. seated people), using a constant reference can be problematic. More 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.

To accounting for a dynamic situations, an estimation of the shoulder orientation is necessary. As this is difficult to extract from vision, we rely 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. Therefore we set the reference value at frame $t$ denoted by $R_t$ as the head pose average computed over the temporal window of duration $W^R$ preceeding the instant $t$:

$$R_t = \frac{1}{W^R} \sum_{i=t-W^R}^{t} H_i$$

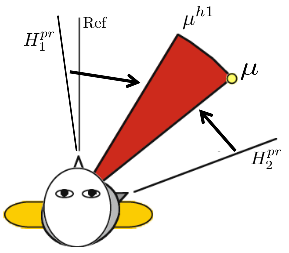

**Fig. 4. Gaze Model with Midline Effect (Hanes and McCollum, 2006) (pan superscripts are dropped for simplicity). The target direction for the shift is denoted by $\mu$. When the gaze is moved to $\mu$ from the initial head pose $H_1^{pr}$, the head is rotated to $\mu^{h1}$ according the geometrical model. However, when the gaze shift is centripetal from $H_2^{pr}$ to $\mu$, the head is moved to $\mu$. For initial head positions between $\mu^{h1}$ and $\mu$ (red zone), an eye-only saccade to $\mu$ is made (the head position remains the same).**

This value can then be used as the substitute for the static reference in the baseline model of Eq. 3, leading to:

$$\mu_t^{hg1}(f) = \alpha \star \mu_t(f) + (1_2 - \alpha) \star R_t \quad (4)$$

We will denote this gaze model by G1.

### *4.2. Model G2: Midline Effect*

Through a literature review and their own experiments, (Hanes and McCollum, 2006) studied more thourougly the relation between the reference, head, and gaze directions. The resulting model is illustrated and explained in Fig. 4[2] One important point they showed is that, for gaze rotations towards the side, while the proportion of gaze shift accomplished by the head depends on the position of the head at the start of the gaze (which in general is not aligned with the reference), the *head end direction* is actually a constant proportion of the gaze when measured *from the reference*, as given by the G1 model. This indeed shows that the G1 model is valid for most gaze shifts (except centripetal ones, cf Fig. 4).

To implement this midline effect we need to know the head pose before a potential gaze shift occurs. We thus introduced the variable $H_t^{pr,pan}$ defining the head pose pan angle prior to a shift and estimated it as the average of the head poses computed over a temporal window of size $W^p$ separated by a gap $\Delta^p$ from the current instant:

$$H_t^{pr,pan} = \frac{1}{W^p} \sum_{i=t-W^p-\Delta^p}^{t-\Delta^p} H_i^{pan} \quad (5)$$

The G2 gaze model was then implemented by setting the mean $\mu_t^{hg2,pan}(f)$ of the head pose pan angle[2] of target $f$ using the following rules (for $\mu^{pan}(f) > 0$ and omitting $f$ for simplicity):

$$\mu_t^{hg2,pan} = \begin{cases} \mu_t^{hg1,pan} & \text{if } H_t^{pr} < \mu_t^{hg1,pan} \\ \min\left(\mu_t^{pan}, \mu_t^{hg1,pan} + \alpha_H (H_t^{pr,pan} - \mu_t^{hg1,pan})\right) & \text{otherwise} \end{cases}$$

The equations to be used when $\mu^{pan}(f) \leq 0$ can be derived following the same principle. Fig. 5a shows the resulting probabilistic graphical model G2. The factor $\alpha_H$ indicates how much

[2] In (Hanes and McCollum, 2006), the reference is called midline. Note that as the model was only studied for the pan variable, in the G2 model (and G3 as well), the tilt gaussian means were set using the G1 model.

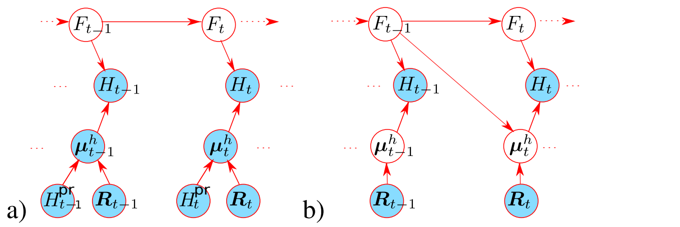

**Fig. 5. Probabilistic graphical models. (a) Model G2. The head reference direction $R_t$ and the mean head pose of the Gaussians $\mu_t^h$ are time dependent variables, and the recent head pose $H_t^{pr}$ can be exploited. (b) Model G3. The mean head pose for looking at a target ($\mu_t^h$) depends on the gaze target at the previous time step ($F_{t-1}$). Shaded nodes indicate that the corresponding random variables are set directly from observation, whereas unshaded nodes denote hidden variables that need to be inferred.**

we take into account the previous head pose in the estimate. When $\alpha_H = 0$, we always have $\mu_t^{hg2,pan} = \mu_t^{hg1,pan}$, which means that the head pose means are set using the G1 model. When $\alpha_H = 1$, the implemented model is exactly the one proposed by (Hanes and McCollum, 2006).

### 4.3. *Model G3: implementing gaze shifts*

When implementing the midline effect, the previous model has one drawback: at each time step, a gaze shift is somehow assumed. In other words, even if the person is focusing on target $f$, the previous head pose $H_t^{pr,pan}$, estimated through recursion over a short temporal window, evolves and as a consequence may lead to an evolution of what the head pose for looking at target $f$ should be, especially when $H_t^{pr,pan}$ is close to the expected head pose $\mu_t^{hg1,pan}$, which might not be appropriate.

As alternative to the model G2, we define the gaze situation prior to a visual attention shift by the actual gaze direction defined by the (discrete) VFOA at the previous instant. We then propose to define the mean of the head pan angle[2] to look at target $f$ at time $t$, given the previous focus $F_{t-1} = \hat{f}$, by:

$$\mu_t^{hg3,pan}(f) = \alpha_1 \mu_t^{pan}(f) + \alpha_2 \mu_t^{pan}(\hat{f}) + (1 - \alpha_1 - \alpha_2) R_t^{pan} \quad (6)$$

Thus, in absence of gaze shift ($F_{t-1} = F_t = f$), the head pose mean is simply given by the geometrical model with $\alpha^{pan} = \alpha_1 + \alpha_2$ and therefore the problematic pose evolution during fixation described above does not exist. In case of a gaze shift ($F_{t-1} \neq f$) the head pose pan angle is not only affected by the reference and new gaze direction $\mu_t^{pan}(f)$ as in G1, but also by the direction towards the VFOA target at previous instant (the head will be closer to direction of the previous VFOA target than what would be predicted by the model G1).

Fig. 5b) shows the new graphical model G3. The link between the hidden states $F_{t-1}$ and $\mu_t^h$ 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 $t$.

## 5. Context Modeling

We aim to leverage context cues to improve VFOA recognition from head pose. Contextual information could potentially help in removing some of the ambiguities due to the limitations of our head pose based VFOA recognition models and to compensate for noisy estimations of the head poses.

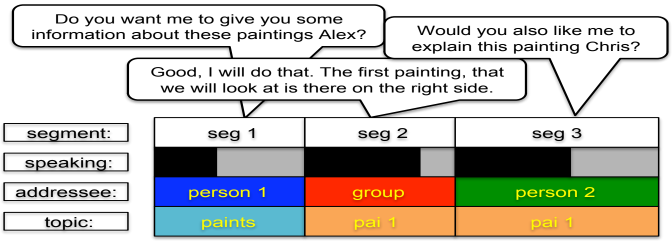
Do you want me to give you some information about these paintings Alex?
Would you also like me to explain this painting Chris?
Good, I will do that. The first painting, that we will look at is there on the right side.
segment:
speaking:
addressee:
topic:
seg 1
seg 2
seg 3
person 1
group
person 2
paints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

**Fig. 6. Illustration of the context assignments.**

When interacting with a robot, its actions influence what people would do in certain situations. Therefore, this information, which the robot is aware of, can be used to predict and better interpret people behavior. In the following, prior to describing more precisely the recognition model, we first introduce in the next section the contextual features that we have exploited.

### 5.1. *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 (Kendon, 1967). However, it is also known that objects playing a central role in conversation may attract the attention and whereby overrule the natural communication patterns (van Turnhout et al., 2005). In our art guide scenario this corresponds to physical locations in the room and particularly paintings. We thus defined the robot interaction context, illustrated in Fig. 6, as described below.

**Speaking context.** Listeners are known to gaze more at speakers than a non-speakers to show attention towards them. Thus 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 (Kendon, 1967). We thus defined the addressee context $a \in AC = \{pers_1, pers_2, group\}$ of a speech segment as the cases when the robot addresses the first, the second, or both persons. 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. 7, 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, assuming a constant model for $x > 0$ is reasonable. We defined the addressee context state $a_t$ at $t$ as

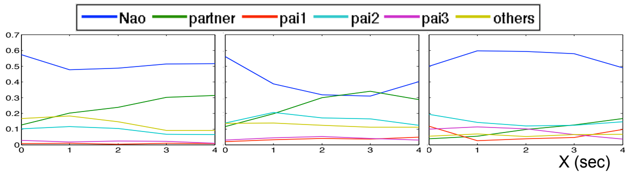

**Fig. 7. VFOA statistics of an individually addressed person (left), a non-addressed person (middle); an addressed person, when both persons are addressed (right). The x axis denotes the time (in second) since the end of the robot utterance. The statistics for $x = 0$ are collected during the robot's utterance.**

the addressee context derived from the current (if $s_t = 1$) or preceding (if $s_t = 0$) robot utterance.

**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, to two or all paintings, or none of them. The topic context state $o_t \in OC$ at time $t$ is thus defined as topic context of the robot utterance that precedes $t$.

**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. 6. The final context state $C_t$ is then defined as the Cartesian product of all contexts, ie $C_t = (s_t, a_t, o_t)$, and will influence the VFOA recognition as explained in the next Section.

### 5.2. *Conversation Aware VFOA Recognition*

To address VFOA recognition using head pose and context information, we use the IOHMM graphical model of Fig. 2. In this model, the VFOA is inferred by maximizing the posterior probability 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. 2 is expressed as:

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

$$\text{with } p(H_t|F_t = f, \mu^h_t) = \mathcal{N}(H_t|\mu^h_t(f), \Sigma_H(f)) \qquad (7)$$

$$\text{and } p(F_t|F_{t-1}, C_t) \propto p(F_t|F_{t-1}) p(F_t|C_t) \qquad (8)$$

where the different terms are explained below.

**Data likelihood.** The term in Eq. 7 represents the likelihood of an observed head pose for a given focus, and is modeled as in Section 3, but here we integrate dynamic means $\mu^h_t$ which play a crucial role for VFOA recognition as in Section 4.

**Contextual prior.** Eq. 8 denotes the prior on the focus, which we assumed can be decomposed in two parts. The first one is the temporal prior $p(F_t|F_{t-1})$ modeled as in Section 3 to allow temporal smoothing. The second one $p(F_t = f|C_t = c) = B_{cf}$ denotes our robot context prior which affects recognition by altering the expectations about what people look at depending on the context. It is parameterized by the probability tables $B$.

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

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

| Context | Nao | partner | $pai_1$ | $pai_2$ | $pai_3$ | others |
|---|---|---|---|---|---|---|
| $pai_1$ | 0.33 | 0.03 | 0.53 | 0.04 | 0.04 | 0.03 |
| $pai_2$ | 0.33 | 0.03 | 0.04 | 0.53 | 0.04 | 0.03 |
| $pai_3$ | 0.33 | 0.03 | 0.04 | 0.04 | 0.53 | 0.03 |
| paints | 0.32 | 0.14 | 0.16 | 0.16 | 0.16 | 0.06 |
| none | 0.58 | 0.17 | 0.04 | 0.04 | 0.04 | 0.12 |

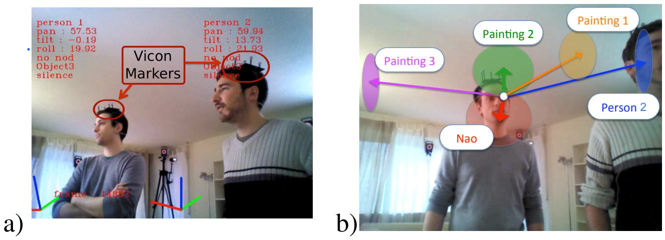

**Fig. 8. a) For the person on the right, there is a potential head pose ambiguity between looking at painting 3 or at the partner. Note as well the Vicon markers. b) VFOA targets to be recognized.**

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 occurences of the focus $f$ in $D_c$, and the Dirichlet prior parameters are set as $\alpha_f = 0.1N_f/(K \times N_C)$, 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 amongst table entries and amounting to 10% of the total number of real observations.

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.

## 6. Experimental set-up and details

In this section we describe our experimental protocol: the data we used, how we obtain the head pose and gaze directions, the performance measures and our parameter setting strategy.

**Set-up and scenario:** We used the Vernissage dataset (D. Jayagopi et al, 2013) to conduct our experiments, whose set-up can be seen in Fig 1b. It contains 10 natural interactions with a humanoid robot "Nao", realized using a Wizard of Oz approach. Each interaction involves two participants standing in front of Nao and free to walk around and look at different objects. In each recording (10 minutes on average), Nao first engages with the participants and explains them three paintings.

**Table 2. VFOA statistics obtained from annotations**

| Label | NAO | Ptr | Pai1 | Pai2 | Pai3 | OT |
|---|---|---|---|---|---|---|
| Freq | 0.43 | 0.11 | 0.06 | 0.14 | 0.06 | 0.20 |

Then, he gives them a quiz in which participants can discuss before the person to whom a question was addressed gives the answer. Both parts are approximately of equal duration. Some questions (4 out of 10) referred to paintings in the room.

**VFOA statistics:** The VFOA labels, given in Section 2, are recapitulated in Fig. 8. The ground truth was annoted by several people and the resulting statistics are shown in Table 2. As can be seen, looking at Nao is dominating, while the remaining gaze are relatively well spread on all other targets[3].

**Recorded Data.** Different synchronized information streams were recorded for each interaction. This comprised the robot state (including the dialog information which was useful for defining the context) and the video stream at VGA resolution captured by a camera located in Nao's head (Fig. 8 and 11 show sample frames). In addition, a Vicon motion capturing system was deployed, with markers placed on the participant's heads (they are visible in Fig. 8a), Nao's head, and near painting locations. To allow 3D reasoning from Nao's perspective, all 3D location and poses (in Vicon reference system) were transformed into location and head pose measurements defined in the local coordinate system of the Nao camera view, which is time-dependent since Nao sometimes rotates its head.

**Head poses.** As input head poses, we used both poses derived from the Vicon system and pose estimates obtained by applying to Nao's camera view (see above) a particle filter tracker performing the joint head tracking and pose estimation with appearance head pose modeling (Khalidov and Odobez, 2013). After inspection of the data, the head pose Vicon measures of one sequence happened to be inconsistent in time (the headbands attaching the Vicon markers to people head might have moved), and we dropped it. Furthermore, since 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, lighting changes, etc.) video tracking 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, and results with too large average pose errors or no sufficient tracker recall were removed. Ultimately, this resulted in a dataset of 14 persons, amounting to around 140 minutes of data for our experiments. On these sequences, the tracker could achieved an average recall (percentage of frames with an estimate) of 80.7%, (min: 48 and max:92), with average pose errors shown in Fig. 9.

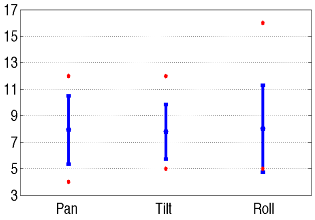

**Fig. 9. Tracker head pose errors on the 14 interactions: minimum, maximum, mean and standard deviations.**

**Gaze directions:** VFOA recognition algorithms also require to know for each participant the gazing directions for different targets in terms of pan and tilt angles. As people are free to move in our recordings, these directions are not fixed and change over time. These values were obtained from the Vicon sensors placed on Nao's head, participant's head and on each of the paintings. However, for a more general application, we assume that Nao knows the room's geometry and can localize itself in the room (Fojtu et al., 2012). By tracking the participants and knowing its location regarding to the other objects in the room, it is capable of measuring these directions and using them for the recognition task.

**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.

**Parameter Setting:** For both Vicon and tracked head pose data, 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 20 and 10 for pan and tilt. The remaining parameters (including context tables) were adjusted by leave-one-out cross-validation separately for each of the models i.e. considering the rest of the all participants as the training set while testing on each participant. Table 3 summarizes the parameters of the gaze-head pose mappings obtained in majority for each of the dynamic model. We can notice that the selected value of $\alpha^{pan}$ (amongs values ranging from 0.4 to 0.9) corresponds to numbers reported in the literature[4]. W.r.t. the size $W^R$ of the window used to average head poses and an approximation of the body orientation, we can see that a rather short size of 20 second was selected (amongs values ranging from 20s to 50s). Indeed, while larger windows provide more stable results, they also introduce more lag to adapt to new situations in case of strong body shifts which occurs for instance when people look at painting $pai_3$ (Fig. 8a) and then switch to painting $pai_1$ (Fig. 8b).

## 7. Results

### 7.1. Head pose-gaze correspondence models

As first experiments, we evaluate and compare the different head pose-gaze dynamical mapping approaches (Baseline, G1,

[3] Note: the statistic for "other" comprises 6% of frames when the people were not visible, and 5% where the annotator could not tell the VFOA label.

[4] A value of $\alpha^{tilt}$ = 0.5 was used in all experiments.

Table 3. Parameters of the dynamical model obtained in majority through cross-validation on Vicon data. $W^R$, $W^p$ and $\Delta^p$ are expressed in seconds.

| Parameters | $\alpha^{pan}$ | $W^R$ | $W^p$ | $\Delta^p$ | $\alpha_H$ | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|---|---|---|
| Baseline | 0.7 | - | - | - | - | - | - |
| G1 | 0.6 | 20 | - | - | - | - | - |
| G2 | 0.6 | 20 | 1 | 0.4 | 1 | - | - |
| G3 | 0.7 | 20 | - | - | - | 0.22 | 0.07 |

Table 4. Recognition rates of head-gaze mappings methods.

| | Vicon head poses | | | Tracker head poses | | |
|---|---|---|---|---|---|---|
| | Full | Explain | Quiz | Full | Explain | Quiz |
| Baseline | 53.8 | 52.4 | 54.6 | 57.3 | 59.3 | 57.4 |
| G1 | 65.5 | 68.8 | 64.2 | 59.1 | 61.7 | 58.7 |
| G2 | 66.6 | 69.9 | 65.3 | 59.8 | 62.3 | 59.3 |
| G3 | 64.3 | 66.7 | 63.3 | 56.7 | 60.2 | 56.0 |

G2 and G3), leaving aside the context part. We first consider the results obtained using head poses given by the Vicon system, and then using poses estimated from the video tracker (see the corresponding paragraphs in Sec. 6). Table 4 summarizes the obtained results.

**Vicon head poses.** The baseline relying on the geometrical model to set the head pose means has only a 53.8% recognition accuracy, which is mainly due to the wrong predictions of the Gaussian means (head directions). In particular, as can be seen from typical confusion matrices[5] of the baseline (left matrices in Fig. 10a) and 10b)) for a person located on the right or left in Fig. 1b, the main source of confusion is between *Nao* and the painting $pai_2$. This is not surprising given their proximity in the gaze space, where they mainly differ in the tilt angular space. Similarly, as expected given the setup, confusion between looking at the third painting ($pai_3$) and *partner* can be seen for the VFOA of person 2 (see Fig. 1b or Fig. 8a) and between the first painting and *partner* for person 1. Moreover, although the Gaussians standards deviations in the HMM are relatively large, several labels are wrongly recognized as looking at *other*.

Among the different dynamic models, G2 which implements the midline model is the best, leading to an average gain of 13% over the baseline. Notice that the gain is more important in the explanation part (17.5%) where people do not face the robot all the time, but orient their bodies towards the paintings (see Fig. 8a) for instance), rather than in the quiz part (10.7%) where people mainly stay oriented towards the robot. The confusion matrices on the right of Fig. 10a) and 10b) obtained with G2, compared to those from the baseline clearly show that the gain is due to a reduced confusion between *Nao* and painting $pai_2$, a reduction of the misclassifications between *partner* and the confusing painting (either painting $pai_3$ for person 2, or painting $pai_1$ for person 1), and less recognition as *others*.

The VFOA recognition is due to a better prediction of the expected head pose for looking at the different targets. To quantify this, we used the ground truth VFOA, and compared at any given instant the participants' head poses used for looking at the targets with their predicted value as given by the baseline of the model G2. Mean errors are shown in Table 5, where the smaller the error, the better the predicted head pose is. As can be seen, the pan angle errors are smaller for all VFOA targets when the dynamic model G2 is used, and in all but one cases for the tilt angle. This is particularly important for *Nao* and $pai_2$ which differ only slightly in their tilt angle.

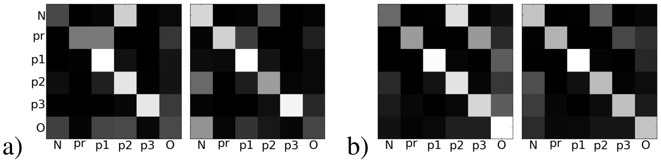

Fig. 10. Confusion matrices (rows are ground truth, columns denote the recognized labels) for (a) a person located in position 'person 1' in Fig. 1b) and (b) a person located in position 'person 2'. In (a) and (b), the matrices on the left result from the Baseline model, whereas the matrices on the right are computed from the G2 results.

Table 5. For each target, means of the errors in degrees between the head pose actually used to look at the target, and the prediction made either by the baseline or the G2 models.

| | Pan angle | | Tilt angle | |
|---|---|---|---|---|
| Target | Baseline | Model G2 | Baseline | Model G2 |
| *Nao* | 7.5 | 4.4 | 5.8 | 3.8 |
| *partner* | 10.6 | 9.9 | 5.0 | 5.0 |
| $pai_1$ | 21.6 | 14.1 | 11.9 | 13.1 |
| $pai_2$ | 38.6 | 30.3 | 7.2 | 4.6 |
| $pai_3$ | 47.4 | 39.5 | 8.0 | 4.8 |

Given the small gain obtained by G2 over G1, we can conclude that the dynamic mapping (through the estimated body orientation) is what contributes the most to the improvement. Qualitatively, its effect is illustrated in Fig. 11. Nevertheless, the midline effect (centripetal movement) is also useful as it provides better recognition results in 13 out of 14 sequences. However, since this effect is happening rarely in the data its effect on performance is also small.

Finally, the model G3 performs much better than the baseline, but a little worse than G1 and G2. However, applied to meeting data, G3 was shown to outperform them (Sheikhi and Odobez, 2012), indicating that it might be more appropriate in presence of more frequent and shorter gaze shifts.

**Head Pose tracker data:** With these data, the main conclusions (ranking of the dynamical models) drawn using Vicon head poses hold. However, here the baseline already gives good recognitions as compared to the Vicon data, and the improvement is smaller (2.5%). This situation can be understood by looking at the average confusion matrices shown in Fig. 13, comparing them with those of the Vicon data. As can be seen from the diagonal elements, the higher accuracy in the baseline is mainly due to a higher recognition for the Nao class, which, given its predominance in the data, results in a higher frame-recognition rate. A potential explanation for the bias towards Nao can be understood by looking at the tracker estimation results in Fig. 12 which displays estimated values for ground truth

[5] For space reasons, VFOA targets in the legend of confusion matrices are denoted by *N* for *Nao*, *pr* for *partner*, *pi* for painting $pai_i$, and *O* for *other*.

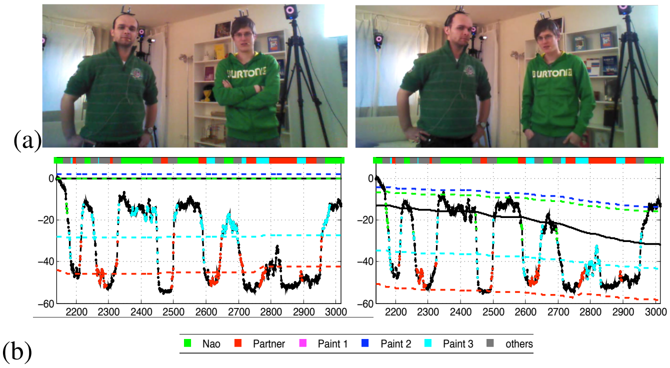

**Fig. 11. (a) Left: during frames 1700-2200, Nao is the main speaker, participants tend to look straight at him. Right: afterwards (quiz part) participants discuss together, alternatively look at the robot and the second person (amongst others). Their reference direction is thus different, and so are the poses 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 pan data is displayed in the graph. It is 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 geometric model (left), or dynamic model G1 (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 (pan near 0 at frame 2150, near -17 at frame 2550).**

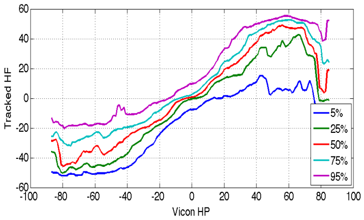

**Fig. 12. Tracker vs Vicon head poses (the pan angle is shown). Different quantiles of the distribution of estimated pose pan values for a given ground truth pose (as given by the Vicon measurements). For instance, the 50% quantile corresponds to the median value. The tracker is relatively accurate up to 40 degrees, but with a tendency for underestimation. This is accentuated for poses beyond 40 degrees.**

head poses given by the Vicon. These curves suggests an underestimation of the pose in general, with the effect of favoring the recognition of Nao as compared to painting 2 for instance. In addition, the underestimation for larger poses leads to head poses that do not match well any of the predicted VFOA targets, and result in a higher recognition of the *other* label (right column). The dynamical model G2 (most right matrix of Fig. 13) tends to reduce the later aspect in certain situations, and to increase the recognition of some targets like painting $pai_3$, including looking at *Nao*.

### 7.2. Exploiting dialog context

To evaluate the contribution of the different contexts, we considered different settings: No context, one single context cue (speaking, addressee, or topic), and all cues together. Furthermore, we experimented the use of the context with both the baseline (static geometrical model) and the best dynamic gaze prediction models (G2) to investigate whether the context is still useful when more accurate gaze-to-head pose predictions are exploited. Tables 6 and 7 show the results when using Vicon and tracked head poses.

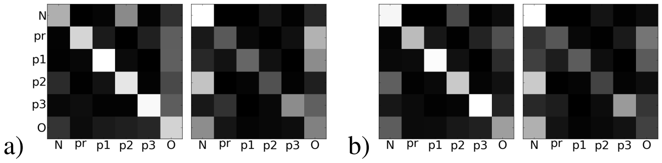

**Fig. 13. Vicon vs Tracker data. (a) average confusion matrices obtained using either the Vicon (left) or tracker data (right). (b) confusion matrices for Vicon (left) vs tracker (right) data using the dynamic model G2.**

**Table 6. Recognition rates using dialog act contexts - Vicon head poses**

| | Baseline Model | | | Model G2 | | |
|---|---|---|---|---|---|---|
| Context | Full | Explain | Quiz | Full | Explain | Quiz |
| None | 53.8 | 52.4 | 54.6 | 66.6 | 69.9 | 65.3 |
| Speak. | 60.9 | 58.3 | 62.1 | 70.2 | 72.3 | 69.4 |
| Addr. | 61.4 | 59.8 | 62.2 | 70.8 | 73.1 | 69.9 |
| Topic | 63.4 | 62.2 | 64.0 | 72.1 | 75.3 | 70.9 |
| All | 64.2 | 63.3 | 64.7 | 72.6 | 75.9 | 71.3 |

When using Vicon data and the baseline dynamical model, we see that the performance improves whatever individual cue we consider. The increase is larger when we use the topic context. Altogether, the use of all context cues brings a considerable improvement of more than 10%. This improvement is valid for all of the 14 sequences, and is illustrated for 2 persons in Fig. 14. As suggested by the shown confusion matrices, the context improves the recognition of all targets simultaneously, and is particularly helpful for removing ambiguities between *Nao* and $pai_2$, *partner* and $pai_1$ and *other* for most cases.

Looking at the combination of context with the dynamical model G2, we can first notice that the context alone (i.e. with the geometric model and static body reference) does not reach the accuracy of the dynamical setting (64.2% with context vs 66.6% with G2). Still, the effects of both approaches are complementary, as the addition of context improves the results of G2 with a gain of 6% when using all cues, and further decreases the confusion between VFOA targets similarly to what is explained above (i.e. between *Nao* and $pai_2$, *partner* and $pai_1$ or $pai_3$). The improvement due to context is observed for 12 out of 14 sequences, and the degradation for the other 2 sequences is very small (2.0% and 0.1%).

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 make a bigger increase in performance during the quiz, which is more interactive, and lower increase during the painting explanations, whereas the topic context improves almost equally on both parts. Finally, using all cues, the performance is higher in all situations.

Considering the results on the tracked head poses, shown in Table 7, we can see that the main conclusions still hold. Individual cues are all useful, the topic cue is more beneficial especially on the explanation part. Combined with the baseline, the context and dynamical model lead to a total improvement of 5%, a gain that is smaller than with Vicon due less accurate head poses and thus more ambiguous situations.

**Table 7. Recognition rates using dialog act contexts - tracked head poses**

| | Baseline Model | | | Model G2 | | |
|---|---|---|---|---|---|---|
| Context | Full | Explain | Quiz | Full | Explain | Quiz |
| None | 57.3 | 59.3 | 57.4 | 59.8 | 62.4 | 59.3 |
| Speak. | 59.1 | 61.5 | 59.1 | 61.0 | 63.1 | 60.9 |
| Addr. | 59.5 | 62.2 | 59.3 | 61.3 | 63.7 | 61.0 |
| Topic | 60.1 | 64.2 | 59.5 | 62.0 | 65.6 | 61.4 |
| All | 60.6 | 65.4 | 59.8 | 62.4 | 66.4 | 61.7 |

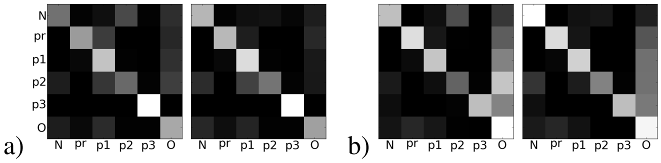

**Fig. 14. Context effect (Vicon data) (a) the image on the left shows the confusion matrix for a given participant when context information is not used while the right one shows the matrix when using the context. (b) shows the same matrices for another participant.**

## 8. Conclusion

In this paper we addressed improving VFOA recognition from head poses in an HRI context using two different solutions. First, we proposed algorithms inspired from body, head and gaze behavioral models to improve the dynamic prediction of the head pose used to look at different VFOA targets. Our experiments on a challenging dataset showed that these models indeed generated more accurate predictions, improving head pose-gaze direction association for all VFOA targets, resulting in a performance increase of more than 10%. Secondly, we proposed a contextual VFOA recognition approach to exploit the robot's gaze-related conversational context (communicative cues, topical cues). It was shown to greatly improve results, and to be complementary to the head-pose dynamical model. Altogether, the combination of the two approaches led to an increase close to 20% in VFOA recognition.

The experiments also showed that obtaining unbiased and accurate head pose is important, as the improvement was smaller using head poses derived from our vision tracker than with the Vicon ones. Such pose estimation improvements come from advances in sensing, and in particular the use of RGB-Depth camera like Kinect. In practice, given the availability of real-time head pose tracking with such device[6], we expect our model to be directly usable by researchers and developers in the HRI or ECA field. Furthermore, the most effective part in our dynamical gaze-to-head prediction approach relies on the use of the body orientation. Hence it would be interesting in the future to test our method on a dataset with available RGB-D dataset that would provide a more direct and more accurate way of estimating it than what we propose. In another direction, with higher definition images, using image-based gaze directions (Gorga and Otsuka, 2010) would be beneficial, and could be combined with our approach. Our prediction model could provide priors on the gaze and be fused with actual image measurements even in noisy conditions.

On the context side, since the dialog act information required by the method is directly incorporated in the dialog system and used at runtime, the model can be exploited for any interactions and in any other scenarios implying objects with the robot is aware of. Finding more systematic ways of setting appropriate VFOA statistics is an avenue for future work, as well as the addition of timing information (how long is a dialog act active?) as well as the use of other cues that can affect the attention of interacting people, like the robot's gestures.

[6] http://msdn.microsoft.com/en-us/library/jj130970.aspx

## References

Ba, S., Odobez, J.M., 2011. Multi-person visual focus of attention from head pose and meeting contextual cues. IEEE PAMI 33, 101–116.

Ba, S.O., Odobez, J.M., 2009. Recognizing visual focus of attention from head pose in natural meetings. Trans. Sys. Man Cyber. Part B 39, 16–33.

Bennewitz, M., Faber, F., Joho, D., Behnke, S., 2007. Fritz – a humanoid communication robot, in: Proc. of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1072–1077.

Bohus, D., Horvitz, E., 2009. Models for multiparty engagement in open-world dialog, in: Proc. of the SIGDIAL Conference, pp. 225–234.

Cooper, R.M., 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology 6.

D. Jayagopi et al, 2013. The vernissage corpus: A conversational human-robot interaction dataset, in: Int. Conf. on Human-Robot Interaction.

Fojtu, S., Havlena, M., Pajdla, T., 2012. Nao robot localization and navigation using fusion of odometry and visual sensor data, in: Intelligent Robotics and Applications. volume 7507, pp. 427–438.

Foster, M.E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., Petrick, R.P., 2012. Two people walk into a bar: dynamic multi-party social interaction with a robot agent, in: Proc. of the Int. Con. on Multimodal Interfaces, ACM.

Gorga, S., Otsuka, K., 2010. Conversation scene analysis based on dynamic bayesian network and image-based gaze detection., in: in Int. Conf. on Multimodal Interfaces, p. 54.

Hanes, D.A., McCollum, G., 2006. Variables contributing to the coordination of rapid eye/head gaze shifts. Biol. Cybern. 94, 300–324.

Kendon, A., 1967. Some functions of gaze-direction in social interaction. Acta Psychol (Amst) 26, 22–63.

Khalidov, V., Odobez, J., 2013. Real-time multiple head tracking using texture and colour cues, in: Internal report, Idiap.

Langton, S.R., Watt, R.J., Bruce, I., 2000. Do the eyes have it? cues to the direction of social attention. Trends Cogn Sci 4, 50–59.

Morency, L.P., Sidner, C.L., Lee, C., Darrell, T., 2005. Contextual recognition of head gestures, in: Int. Conf. on Multimodal Interfaces, pp. 18–24.

Nagai, Y., Asada, M., Hosoda, K., 2006. Learning for joint attention helped by functional development. Advanced Robotics 20, 1165–1181.

Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J., 2003. Towards a model of face-to-face grounding, in: Proc. of the Annual Meeting on Association for Computational Linguistics, pp. 553–561.

Otsuka, K., Takemae, Y., Yamato, J., Murase, H., 2005. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances, in: Int. Conf. on Multimodal Interfaces.

Sheikhi, S., Odobez, J.M., 2012. Investigating the midline effect for visual focus of attention recognition, in: Int Conf. on Multimodal Interactions.

Stiefelhagen, R., Yang, J., Waibel, A., 2002. Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans. on Neural Networks 13(4), 928–938.

Triesch, J., Teuscher, C., Deák, G.O., Carlson, E., 2006. Gaze following: why (not) learn it? Developmental Science 9, 125–147.

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

Voit, M., Stiefelhagen, R., 2008. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios, in: ICMI.