Simultaneous Estimation of Gaze Direction and Visual Focus of Attention for Multi-Person-to-Robot Interaction

Benoit Massé, Silèye Ba, Radu Horaud

SIMULTANEOUS ESTIMATION OF GAZE DIRECTION AND VISUAL FOCUS OF ATTENTION FOR MULTI-PERSON-TO-ROBOT INTERACTION

Benoit Massé, Silèye Ba and Radu Horaud

INRIA Grenoble Rhône-Alpes, FRANCE

ABSTRACT

We address the problem of estimating the visual focus of attention (VFOA), *e.g*. who is looking at whom? This is of particular interest in human-robot interactive scenarios, *e.g*. when the task requires to identify targets of interest over time. The paper makes the following contributions. We propose a Bayesian temporal model that connects VFOA to gaze direction and to head pose. Model inference is then cast into a switching Kalman filter formulation, which makes it tractable. The model parameters are estimated via training based on manual annotations. The method is tested and benchmarked using a publicly available dataset. We show that both the gaze and the VFOA of several persons can be reliably and simultaneously estimated over time from observed head poses as well as from people and object locations. On average, our method compares favorably with two other methods.

1. INTRODUCTION

Whether engaged in formal meetings or in informal gatherings, people communicate via a number of verbal and non-verbal cues, such as speech, prosody, head and hand gestures, head and eye gaze, facial expressions, etc. For example in a multi-party conversation, a common behavior among the participants consists in looking either at the speaker or at the current object of interest. *e.g*. a computer screen, a painting on a wall, or an object on a table top. This enables participants to both respect social etiquette and to focus their attention onto the topic of the meeting/gathering. This is also the case in human-robot interaction (HRI) scenarios that involve both person-to-person and robot-to-person interaction. Consider, for example, the case of a robot companion whose role is to assist people. The primary task of the robot is to analyse a number of non-verbal cues in order to understand the situation and to act appropriately, *e.g*. pop into the conversation at the right moment. Among these cues, visual focus of attention (VFOA) estimation of multiple persons provides answers to: *Who is looking at whom? Who is looking at what? Who is the speaker? Who are the listeners? etc*.

Nevertheless, simultaneous estimation of VFOAs of several persons is a difficult task. It requires the estimation of object locations and of gaze directions. The former can be obtained *e.g*. using either face tracking [1] or upper-body tracking [2]; the latter depends on both head and eye orientation.[1] Many existing methods provide an accurate estimation of gaze from eye analysis, *e.g*. [3, 4, 5, 6]. These methods rely on high-quality iris detection, either from an invasive head-mounted system [6], or by constraining the user to gaze towards the camera. For unconstrained scenarios, *e.g*. informal interactions, it is generally not possible to directly observe the eyes in the sensory data. Some faces are partially occluded, not facing the cameras, or too far away. Without observing the eyes, these methods cannot infer gaze. An alternative is to use the head pose as a cue for gaze direction [7]. Indeed, gaze direction shifts are often done by moving synchronously both the head and the eyes [8], and a vast class of methods provides head orientation from visual data [9]. Many methods estimate VFOA from head pose in meetings *e.g*. [7, 10, 11, 12]. Indeed, meetings provide a natural interaction between people that do not move, where head pose is not constrained but still stays into an acceptable range. Joint use of cognitive models and of geometric information to overcome the unobserved eye direction was proposed in [11], later extended with a temporal geometric model in [13]. [14] proposed to estimate gaze direction as an intermediary step: gaze is first estimated from head pose and then VFOA is estimated from gaze.

In this paper we propose an on-line Bayesian temporal model for the simultaneous estimation of gaze direction and of visual focus of attention from observed head poses (location and orientation) and from object locations. Gaze directions, head directions and VFOAs are combined in a temporal Gaussian model in which the VFOAs provide gaze direction priors. We introduce an additional set of latent variable, namely the head reference directions, and we define their dynamics to account for long-term gaze variations. We show that the joint estimation of gaze and VFOA can be cast into a switching Kalman filter model and thus the proposed formulation is tractable. We formally derive formulas for the gaze dynamics and for the VFOA transition probabilities and we show that their parameters can be easily estimated via standard maximum-likelihood procedures.

The method is tested with the publicly available

Funding from the European Research Council through the Advanced Grand VHIA #340113 is greatly acknowledged.

[1] Throughout this paper we make a clear distinction between gaze direction and eye orientation.

Vernissage dataset [15]. The scenarios consist of two persons and of one robot that interact with each other while gazing at different objects in the scene. The dataset was recorded with a network of infra-red cameras synchronized with a camera mounted onto a robot head. In conjunction with optical markers mounted onto the persons' and robot's heads, this setup allows accurate estimation of head poses in each frame. The ground-truth VFOAs, for each frame and for each person, were carefully annotated, thus allowing quantitative evaluation and benchmarking of both gaze direction and VFOA estimation.

The remainder of the paper is organized as follows. Section 2 formulates VFOA and gaze estimation as a MAP problem and describes the associated graphical model. Section 3 describes the likelihood model and derives the gaze and the VFOA dynamics. Section 4 show how the MAP problem is cast into a switching Kalman filter formulation and describes the associated parameter learning method. Section 5 describes in detail experiments conducted with the Vernissage dataset. Finally, Section 6 draws some conclusions.

2. PROBLEM FORMULATION

We consider a scenario composed of $N+M$ *objects*, namely N persons, M targets, as well as a robot. While the persons are *active*, the targets are *passive* and without loss of generality it will be assumed that the object locations are expressed in a robot-centered coordinate frame. We also assume that the number of persons N and targets M are known and remain constant over time. The VFOA of person i at time t is denoted by the discrete variable $\mathrm{V}_t^i \in \mathcal{V}^i$, with $\mathcal{V}^i = \{0, 1, \ldots, N+M\}/\{i\}$, such that $\mathrm{V}_t^i = j$ means that person i either looks at j if $1 \leq j \leq N+M$ ($j \neq i$), or looks at "nothing" if $j = 0$. The VFOA set at time t is denoted by $\mathbf{V}_t = (\mathrm{V}_t^1, \ldots, \mathrm{V}_t^i, \ldots, \mathrm{V}_t^N)$.

The VFOA is defined in the following way. In order to infer whether person i looks at object j, the gaze direction of i as well as the relative positions of i and j are needed. Gaze directions are denoted by $\{\mathbf{G}_t^i\}_{i=1}^N \subset \mathbb{R}^2$, *i.e.* pan and tilt angles, and it is assumed in this work that they cannot be directly observed from the sensory data. Instead, we rely on observing head positions, head directions and target positions. Object positions (whether persons or targets) are denoted by $\{\mathbf{X}_t^i\}_{i=1}^{N+M} \subset \mathbb{R}^3$ (3D coordinates in a robot-centered frame) and head directions by $\{\mathbf{H}_t^i\}_{i=1}^N \subset \mathbb{R}^2$, *i.e.* pan and tilt angles. We also define the directions $\{\mathbf{D}_t^{ij}\}_{i \neq j} \subset \mathbb{R}^2$ from i to j that are computed from $\mathbf{X}_t^i$ and $\mathbf{X}_t^j$.

Because *latent* gaze directions are inferred from *observed* head directions, we need to model the relationship between these two variables. For that purpose we introduce the *head reference direction* **latent** variable $\{\mathbf{R}_t^i\}_{i=1}^N \subset \mathbb{R}^2$. This direction corresponds to a gaze direction which is likely to be equal to the head direction. We assume that the expected head direction is a convex combination of gaze and head reference:

$$\mathbb{E}[\mathbf{H}_t^i] = \boldsymbol{\alpha}\mathbf{G}_t^i + (\mathbf{I}_2 - \boldsymbol{\alpha})\mathbf{R}_t^i \tag{1}$$

where $\mathbf{I}_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix and $\boldsymbol{\alpha} = \mathrm{Diag}\,(\alpha_1, \alpha_2)$ is a diagonal matrix whose entries are mixing coefficients, $0 < \alpha_1, \alpha_2 < 1$. Fig. 1 shows a graphical representation of the observed and latent variables as well as their dependencies.

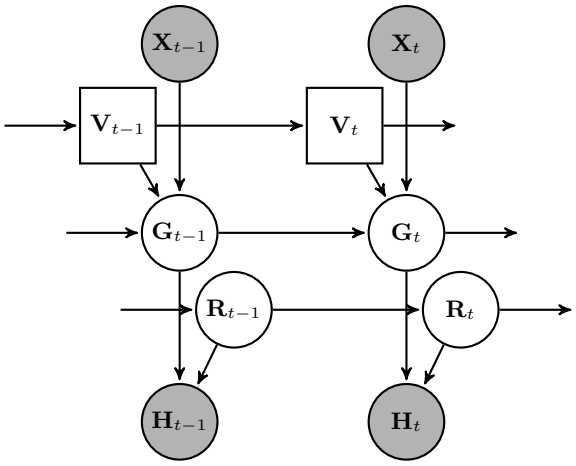

Fig. 1. Graphical representation showing the model variables and their dependencies. Squares describe discrete latent variables, circles describe continuous latent variables, and shaded circles describe observations.

Within a Bayesian temporal formulation, the objective is to estimate the VFOA filtering distribution given the observation history, namely $P(\mathbf{V}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. This distribution captures VFOA information available in the observed variables. VFOA estimation is cast into a MAP formulation:

$$\hat{\mathbf{V}}_t = \underset{\mathbf{V}_t}{\mathrm{argmax}}\, P(\mathbf{V}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}). \tag{2}$$

This filtering distribution is the marginal distribution of the joint VFOA, gaze direction, and head reference direction filtering distribution $P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t})$:

$$P(\mathbf{V}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}) = \int P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}) d\mathbf{G}_t d\mathbf{R}_t$$

which allows us to make use of the relationship between head direction, gaze direction, and head reference direction defined in (1). Using variable independency assumptions, *i.e.* Fig. 1, the joint filtering distribution can be expanded as:

$$\begin{aligned} &P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \\ &= \frac{P(\mathbf{H}_t|\mathbf{G}_t, \mathbf{R}_t) P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{H}_{1:t-1}, \mathbf{X}_{1:t})}{P(\mathbf{H}_t|\mathbf{H}_{1:t-1}, \mathbf{X}_{1:t})} \end{aligned} \tag{3}$$

which is composed of three terms: the observation likelihood $P(\mathbf{H}_t|\mathbf{G}_t, \mathbf{R}_t)$, the state predictive distribution $P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{H}_{1:t-1}, \mathbf{X}_{1:t})$, and the observation predictive distribution $P(\mathbf{H}_t|\mathbf{H}_{1:t-1}, \mathbf{X}_{1:t})$.

3. LIKELIHOOD AND STATE DYNAMICS

Observation Likelihood. Assuming the model in (1) allows to predict head direction from gaze direction and from head reference direction up to Gaussian noise with covariance matrix $\mathbf{\Sigma_H}$, and that head direction observations are conditionally independent given gaze and head reference direction, the observation likelihood writes, where the mean is given by (1):

$$P(\mathbf{H}_t|\mathbf{G}_t,\mathbf{R}_t) = \prod_i \mathcal{N}(\mathbf{H}_t^i; \mathbb{E}[\mathbf{H}_t^i], \mathbf{\Sigma_H}), \tag{4}$$

State Dynamics. The gaze, head reference, and VFOA dynamics can be factorized as

$$P(\mathbf{V}_t, \mathbf{G}_t, \mathbf{R}_t|\mathbf{V}_{t-1}, \mathbf{G}_{t-1}, \mathbf{R}_{t-1}, \mathbf{X}_t) = P(\mathbf{G}_t, \mathbf{R}_t|\mathbf{G}_{t-1}, \mathbf{R}_{t-1}, \mathbf{V}_t, \mathbf{X}_t) P(\mathbf{V}_t|\mathbf{V}_{t-1}).$$

Assuming that the dynamics of $\mathbf{G}_t$ and $\mathbf{R}_t$ are conditionally independent yields the factorization $P(\mathbf{G}_t, \mathbf{R}_t|\mathbf{G}_{t-1}, \mathbf{R}_{t-1}, \mathbf{V}_t, \mathbf{X}_t) = P(\mathbf{G}_t|\mathbf{G}_{t-1}, \mathbf{V}_t, \mathbf{X}_t)\ P(\mathbf{R}_t|\mathbf{R}_{t-1})$. Furthermore, we assume that there is no pairwise dependencies between gaze directions and head reference directions, and that the predictions are corrupted with Gaussian noise. This leads to the following first order Markov model for the head reference directions:

$$P(\mathbf{R}_t|\mathbf{R}_{t-1}) = \prod_i \mathcal{N}(\mathbf{R}_t^i; \mathbf{R}_{t-1}^i, \mathbf{\Gamma_R}) \tag{5}$$

where $\mathbf{\Gamma_R}$ is a covariance matrix. The gaze dynamics involves two input variables: the VFOA state $\mathbf{V}_t$ and the objects positions $\mathbf{X}_t$. We define the prior about the gaze dynamics as follows:

$$P(\mathbf{G}_t|\mathbf{G}_{t-1}, \mathbf{V}_t, \mathbf{X}_t) = \prod_i P(\mathbf{G}_t^i|\mathbf{G}_{t-1}^i, \mathbf{V}_t^i, \mathbf{X}_t) \tag{6}$$

where the gaze dynamics of person i is defined as

$$P(\mathbf{G}_t^i|\mathbf{G}_{t-1}^i, \mathbf{V}_t^i, \mathbf{X}_t) = \mathcal{N}(\mathbf{G}_t^i; \mathbf{G}_{t-1}^i, \mathbf{\Gamma_G})^{\delta_0(\mathbf{V}_t^i)} \times \prod_{j\neq 0} \mathcal{N}(\mathbf{G}_t^i; \boldsymbol{\beta}\mathbf{G}_{t-1}^i + (\mathbf{I}_2 - \boldsymbol{\beta})\mathbf{D}_t^{ij}, \mathbf{\Gamma_G})^{\delta_j(\mathbf{V}_t^i)} \tag{7}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{2\times 2}$ is a diagonal matrix whose entries are mixing coefficients, $0 \leq \beta_{11}, \beta_{22} \leq 1$, and δ_j is the Kronecker symbol such that $\delta_j(\mathrm{V}_t^i) = 1$ when $\mathrm{V}_t^i = j$, and $\delta_j(\mathrm{V}_t^i) = 0$ otherwise.

Equation (7) should be interpreted as follows. The gaze dynamics of person i is a switching dynamical model having the VFOA state $\mathbf{V}_t^i$ as a switching variable. When person i gazes at none of the $N + M$ objects, namely $\mathbf{V}_t^i = 0$, then his/her gaze direction follows a random walk. Otherwise, when he/she gazes at object $j \neq 0$, $\mathbf{V}_t^i = j$, then his/her gaze follows a first order dynamics leaning towards $\mathbf{D}_t^{ij}$ (the direction from person i to object j) at a rate defined by $\boldsymbol{\beta}$. The proposed gaze and head reference dynamics assume that gaze dynamics is *faster* than head reference dynamics. This assumption is enforced by the constraint $\mathrm{Tr}(\mathbf{\Gamma_G}) \gg \mathrm{Tr}(\mathbf{\Gamma_R})$.

Moreover, velocities $\dot{\mathbf{G}}_t^i$ and $\dot{\mathbf{R}}_t^i$ can be added to the Gaussian dynamics. In practice, $\mathbf{R}_t^i$ in (5) and $\mathbf{G}_t^i$ in (7) are replaced with $\mathbf{R}_t^i + dt\dot{\mathbf{R}}_t^i$ and $\mathbf{G}_t^i + dt\dot{\mathbf{G}}_t^i$, respectively. The velocity dynamics are:

$$P(\dot{\mathbf{R}}_t^i|\dot{\mathbf{R}}_{t-1}^i) = \mathcal{N}(\dot{\mathbf{R}}_t^i; \dot{\mathbf{R}}_{t-1}^i, \mathbf{\Gamma}_{\dot{\mathbf{R}}}) \tag{8}$$

$$P(\dot{\mathbf{G}}_t^i|\dot{\mathbf{G}}_{t-1}^i) = \mathcal{N}(\dot{\mathbf{G}}_t^i; \dot{\mathbf{G}}_{t-1}^i, \mathbf{\Gamma}_{\dot{\mathbf{G}}}). \tag{9}$$

VFOA Dynamics. VFOA are discrete variables and hence their prior dynamics are modeled by transition matrices. Assuming that VFOA variables at t are conditionally independent given the past, the VFOA transition priors can be factorized as:

$$P(\mathbf{V}_t|\mathbf{V}_{t-1}) = \prod_i P(\mathrm{V}_t^i|\mathbf{V}_{t-1}) \tag{10}$$

The set $\mathbf{V}_{t-1}$ can be further reduced either to V_{t-1}^i alone, if the VFOA of person i is a passive object k, or to the pair (V_{t-1}^i, V_{t-1}^k) if the VFOA of person i is person k. This yields the following expression:

$$P(\mathrm{V}_t^i = j|\mathbf{V}_{t-1}) = P(\mathrm{V}_t^i = j|\mathrm{V}_{t-1}^i = k)^{1-\delta_{\mathcal{A}}(k)} \times P(\mathrm{V}_t^i = j|\mathrm{V}_{t-1}^i = k, \mathrm{V}_{t-1}^k = l)^{\delta_{\mathcal{A}}(k)} \tag{11}$$

where $\mathcal{A}$ denotes the set of persons. This model allows to account for situations where person i focuses on person k who is in turn focusing on l, leading person i to eventually focus on l. Therefore, this accounts for persons jointly focusing on the same object, and this is done in a dynamic fashion.

4. INFERENCE AND LEARNING

Let $\mathbf{L}_t = [\mathbf{G}_t; \dot{\mathbf{G}}_t; \mathbf{R}_t; \dot{\mathbf{R}}_t]$ where $[\cdot;\cdot]$ denotes vertical vector concatenation. Both $\mathbf{L}_t$ and $\mathbf{H}_t$ follow a linear Gaussian model, given the discrete state variables $\mathbf{V}_t$, *i.e.* (4)–(7).

Inference. As stated in eq. (2), we want to find the MAP over $\mathbf{V}_t$. However, the number of states is exponential in the number of people. Even if the posterior distribution is tractable for simple scenarios with few people, it requires a lot of parameters that must be learn for each value of N and of M. Instead, we approximate the joint filtering distribution as

$$P(\mathbf{V}_t|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \approx \prod_i P(\mathrm{V}_t^i|\mathbf{H}_{1:t}, \mathbf{X}_{1:t}). \tag{12}$$

The inference problem is then reduced to evaluating $c_t^{ij} = P(\mathrm{V}_t^i = j|\mathbf{H}_{1:t}, \mathbf{X}_{1:t})$ for every person i and every object j. A propagation formulation is now derived to obtain c_t^{ij} recursively: $c_t^{ij} = \sum_k c_{t-1,t}^{ijk}$ where $c_{t-1,t}^{ijk} = P(\mathrm{V}_t^i = j, \mathrm{V}_{t-1}^i =$

$k \mid \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. Bayes formula yields:

$$c_{t-1,t}^{ijk} \propto P(\mathbf{H}_t|\mathrm{V}_t^i = j, \mathrm{V}_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) \times c_{t-1}^{ik} \sum_l c_{t-1}^{kl} P(\mathrm{V}_t^i = j|\mathbf{V}_{t-1}). \quad (13)$$

This provides a recursive formulation for c_t^{ij} where the dependency on the last factor in (13) w.r.t. to l appears from (11). The first factor in (13), the observation component, can be factorized as $P(\mathbf{H}_t^i|\mathrm{V}_t^i = j, \mathrm{V}_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) \times \prod_{n \neq i} \sum_m \sum_p P(\mathbf{H}_t^n|\mathrm{V}_t^n = m, \mathrm{V}_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1})$. by introducing the latent variable $\mathbf{L}_t$ we obtain:

$$\begin{aligned} &P(\mathbf{H}_t^n|\mathrm{V}_t^n = m, \mathrm{V}_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) \\ &= \int P(\mathbf{H}_t^n|\mathbf{L}_t^n)\, P(\mathbf{L}_t^n|\mathbf{L}_{t-1}^n, \mathrm{V}_t^n = m) \\ &\times P(\mathbf{L}_{t-1}^n|\mathrm{V}_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) d\mathbf{L}_{t-1}^n d\mathbf{L}_t^n. \quad (14) \end{aligned}$$

While $P(\mathbf{H}_t^n|\mathbf{L}_t^n)$ is known from (4) and $P(\mathbf{L}_t^n|\mathbf{L}_{t-1}^n, \mathrm{V}_t^n)$ from (5) and (7), $P(\mathbf{L}_{t-1}^n|\mathrm{V}_{t-1}^n, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1})$ must be evaluated. $\mathbf{L}_{t-1}^i$ follows a linear Gaussian dynamics, whose parameters depend on the value of V_{t-1}^i. This exactly fits the switching Kalman filter (SKF) formulation [16] where V_{t-1}^i is the switch variable. Specifically, $P(\mathbf{L}_{t-1}^i|\mathrm{V}_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1})$ follows the distribution $\mathcal{N}(\mathbf{L}_{t-1}^i; \boldsymbol{\mu}_{t-1}^{ik}, \boldsymbol{\Sigma}_{t-1}^{ik})$. Then (14) and then (13) can be solved in closed form. Finally, we need a recursive formulation to obtain $\boldsymbol{\mu}_t^{ij}$ and $\boldsymbol{\Sigma}_t^{ij}$ from their values at $t-1$. This is done using the GPB2 algorithm [16]. The idea is to compute the filtering step $\boldsymbol{\mu}_t^{ijk}$ and $\boldsymbol{\Sigma}_t^{ijk}$ for each possible transition path $P(\mathbf{L}_t^i|\mathrm{V}_{t-1}^{ik}, \mathrm{V}_t^{ij}, \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. Then, the resulting mixture of Gaussians $P(\mathbf{L}_t^i|\mathrm{V}_t^{ij}, \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$ is approximated by a single Gaussian. This collapsing process is weighted with $c_{t-1,t}^{ijk}$.

Based on this formalism we devised a procedure that alternates between evaluating the VFOA distribution and evaluating the gaze and head reference variables. The proposed procedure propagates forward the information about past observations and allows to infer the VFOA MAP of each person in an online fashion.

Learning. The parameters of the proposed model are the covariance matrices $\boldsymbol{\Sigma}_\mathbf{H}$, $\boldsymbol{\Gamma}_\mathbf{R}$ and $\boldsymbol{\Gamma}_\mathbf{G}$ in (4), (5) and (7), and the VFOA transition probabilities (11). Notice that the mean vectors are provided by the matrices $\boldsymbol{\alpha}$ in (1) and $\boldsymbol{\beta}$ in (7) whose diagonal entries are mixing coefficients acting as hyper-parameters. Since it is assumed that VFOA annotation is available for training, one can estimate the model parameters via maximum likelihood. The VFOA transition probability $P(\mathrm{V}_t^i = j|\mathbf{V}_{t-1})$ does not depend on the specific persons but, instead, on whether the VFOA changes and how it changes. Given the dependency chosen in (11), one can enumerate 15 cases. A reliable maximum-likelihood estimator simply consists in counting the transitions in the training set and normalizing with respect to the previous state. The covariance matrices are estimated via a closed-form EM. The hyper-parameters are estimated using a cross-validation protocol, namely the values that best match the expected VFOAs.

Fig. 2. The Vernissage setup. Left: Global view of the "exhibition" scene showing wall painting, two persons and the NAO robot. Right: Top view representation of the room.

5. EXPERIMENTS

In order to evaluate the proposed method, we used the Vernissage dataset [15], that consists of ten recordings of people in an exhibition. Each recording is composed of two people, denoted *Left* and *Right*, one robot, denoted NAO (N=3) as well as three wall paintings, denoted o_1, o_2, and o_3 (M=3), *e.g.* Fig. 2. The dataset is composed of ten-minute recordings involving 20 different persons. The recorded scenario is the following: first, the robot presents the paintings to the public (lasting four minutes) and second, the two visitors talk to each other and to the robot in order to solve a quiz (lasting six minutes). The experiments described below only used the second part of the recordings.

The scene was recorded with a camera mounted onto the robot head and with a network of infrared cameras placed on the walls. These cameras are used in conjunction with optical markers, placed onto both the robot and person heads, to provide accurate head positions $\mathbf{X}_{1:t}$ and head orientations $\mathbf{H}_{1:t}$ in a common reference frame. The robot-head camera is synchronized with the infrared cameras at 25 FPS, hence there is a total of $10 \times 360 \times 25 = 90,000$ frames. The VFOAs $\mathbf{V}_{1:t}$ of the two persons were manually annotated in each frame, thus providing ground-truth VFOA for each person.

To evaluate the method, we used head and painting positions, and head orientations provided by a motion capture system that uses the camera network in conjunction with the optical markers, while the robot-head camera was used only for visualization purposes. A VFOA estimation method based on HMMs was proposed in [11]. We implemented this method and used it as a baseline for comparison purposes.

The latent state of the proposed model is composed of gaze and head-reference direction variables $\mathbf{G}_{1:t}$ and $\mathbf{R}_{1:t}$; the observed head direction is a convex combination of these variables (1). Whenever velocity dynamics is being consid-

	Ba [11]		Sheikhi [13]		Proposed	
Video	Left	Right	Left	Right	Left	Right
09	54.6	58.8	51.2	**59.6**	**57.9**	55.4
10	64.9	**77.1**	-	-	**70.7**	66.9
12	**49.9**	**70.0**	-	-	45.7	59.9
15	66.3	46.1	-	-	**70.4**	**68.0**
18	36.3	25.5	-	-	**66.9**	**56.4**
19	54.3	49.6	-	-	**54.6**	**69.8**
24	33.9	48.7	-	-	**35.7**	**56.1**
26	39.0	28.0	-	-	**47.4**	**42.9**
27	70.6	**74.0**	-	-	**71.3**	73.5
30	75.0	48.6	-	-	**76.3**	**66.2**
Overall	53.6		*55.4*		**60.6**	

Table 1. FRR scores for VFOA estimation with the Vernissage dataset.

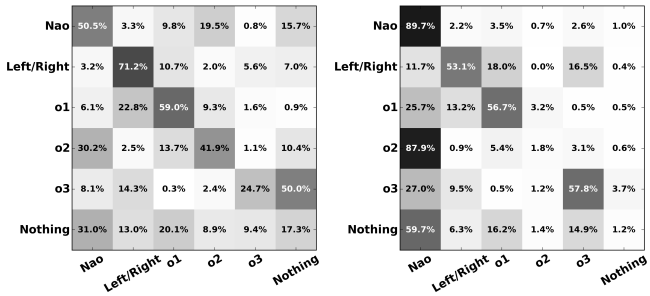

Fig. 3. Confusion matrix for [11] (left) and for the proposed method (right). Rows: ground-truth VFOA. Columns: estimated VFOA.

ered, the expected latent state may diverge while the emission distribution is correctly evaluated. This problem is addressed by restricting the optimal Kalman gain, as proposed in [17]. This is implemented with the constraints $|G_{1,t} - H_{1,t}| < 25^\circ$ and $|G_{2,t} - H_{2,t}| < 25^\circ$.

The model parameters were estimated based on the learning method described at the end of section 4 and using the manual VFOA annotations. Based on cross-validation, the diagonal entries of the mixing matrices were set to $\boldsymbol{\alpha} = \mathrm{Diag}\,(0.7, 0.3)$ and to $\boldsymbol{\beta} = \mathrm{Diag}\,(0.5, 0.5)$. The VFOA transition probabilities were estimated via maximum likelihood. Referring to (11), the transition probabilities vary between 0.89 and 0.97 if $j = k$ (the probability that the VFOA is the same at $t-1$ and at t) and between 0.005 and 0.05 otherwise. The covariances are first initialized as isotropic covariances, namely $\mathbf{\Sigma_H} = \sigma_H^2\mathbf{I}_2$, $\mathbf{\Gamma_G} = \gamma_G^2\mathbf{I}_2$ and $\mathbf{\Gamma_R} = \gamma_R^2\mathbf{I}_2$ with $\sigma_H = 15^\circ$, $\gamma_G = 5^\circ$, and $\gamma_R = 0.5^\circ$, and second they are estimated via a standard Kalman EM algorithm.

We use the frame recognition rate (FRR) to measure the performance. FRR is the percentage of frames for which the VFOA is correctly estimated. Since there are $90,000$ annotated frames in the Vernissage dataset, FRR is a statistically meaningful score. Table 1 summarizes the results obtained with the proposed method and with [11] and [13]. The results show that our method performs better than the other two methods, on an average. Notice that the performance of our method, *i.e*. percentage of correct VFOA estimates, varies from 31% to 76%. This variability is mainly due to differences in people behavior in terms of gaze. For some of the persons in the dataset, the proposed relationship between head direction, gaze direction, and head reference direction is valid. In other terms, our formulation is well suited for people who move their heads while they gaze to an object.

It should be noted that FRR is biased. Indeed, in the Vernissage dataset people look at NAO half of the time. Since the VFOA probability transition matrix favors continuity (the probability to gaze at the same object over time is high), our implementation performs very well when the VFOA is either NAO or the paintings o_1 and o_3, and performs less well when the VFOA is painting o_2 which is behind the robot. The method of [11] uses a fixed head reference direction, which is defined by the user. Hence, the results obtained with [11] strongly depend on the reference direction prior. This is illustrated with the confusion matrices shown on Fig. 3. Examples obtained with our method are shown on Fig. 4.

6. CONCLUSION

We proposed a method for the joint estimation of gaze directions and VFOAs in multi-person-to-robot interactive scenarios. The main novely of the proposed model is that direct estimation of eye gaze from the data is not required. Instead, a generative model is proposed that treats both gaze and VFOA as latent variables in a Bayesian temporal formulation. We showed that the proposed model can be cast into an SKF formalism, thus insuring tractability in terms of inference and learning. The method was thoroughly trained and tested using a publicly available dataset. The results were compared with two state-of-the-art methods.

The experiments use observations from a motion capture system (infrared cameras and optical markers) to estimate head poses and a camera mounted onto a robot head for visualization of the results. In the near future we plan to use the robot-head camera instead of the motion capture system in order to fully demonstrate the robustness of the method in less constrained human-robot interaction scenarios. We also plan to extend our method such that it can deal with moving persons that may be partially occluded, and with objects that are not visible. Indeed, we believe that our approach is particularly well suited in such challenging, yet realistic, situations because the method does not need direct observation of gaze from eye detection, localization and orientation.

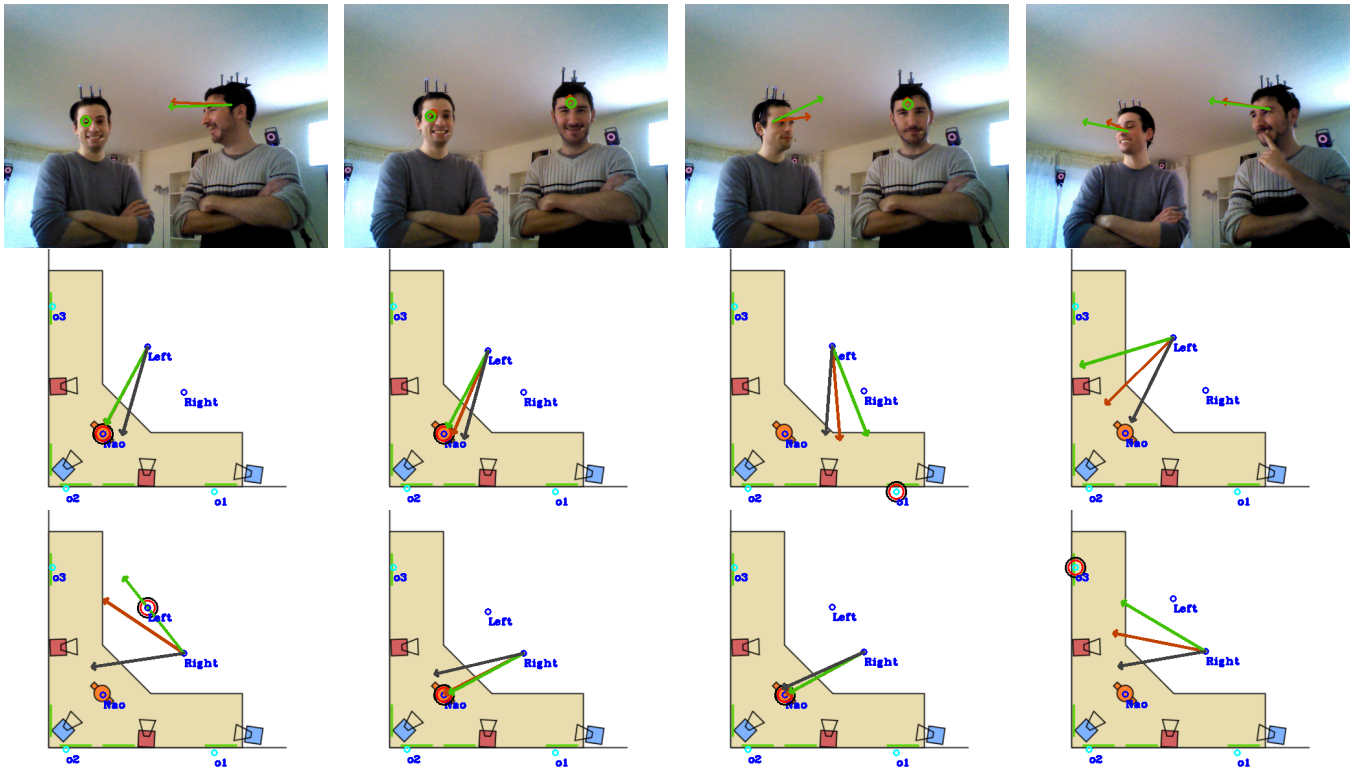

Fig. 4. Results obtained with the proposed method. Gaze directions are shown with green arrows, head reference directions with dark-grey arrows and observed head directions with red arrows. The ground-truth VFOA is shown with a black circle. The top row displays the image of the robot-head camera. Top views of the room show and results obtained for the Left (middle row) and Right (bottom row) persons. In the last example the Left person gazes at "nothing".

7. REFERENCES

[1] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modified census transformation," *Image and Vision Computing*, vol. 24, 2006.

[2] R. Poppe, "Vision-based human motion analysis: An overview," *CVIU*, vol. 108, 2007.

[3] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head pose and gaze direction measurement," in *IROS Proceedings*, 2000, vol. 3.

[4] P. Smith, M. Shah, and N. Da Vitoria Lobo, "Determining driver visual attention with one camera," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 4, 2003.

[5] T. Ohno and N. Mukawa, "A free-head, simple calibration, gaze tracking system that enables gaze-based interaction," in *Proceedings of the ETRA symposium*. ACM, 2004.

[6] A. K. A. Hong, J. Pelz, and J. Cockburn, "Lightweight, low-cost, side-mounted mobile eye tracking system," in *Western New York Image Processing Workshop*. IEEE, 2012.

[7] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *Human Factors in Computing Systems*, 2002.

[8] E. G. Freedman and D. L. Sparks, "Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys," *Journal of Neurophysiology*, 1997.

[9] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE TPAMI*, vol. 31, 2009.

[10] K. Otsuka, J. Yamato, and Y. Takemae, "Conversation scene analysis with dynamic bayesian network based on visual head tracking," in *IEEE ICME*, 2006.

[11] S.O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE TSMC-B*, 2009.

[12] S. Duffner and C. Garcia, "Visual focus of attention estimation with unsupervised incremental learning," *IEEE TCSVT*, 2015.

[13] S. Sheikhi and J-M. Odobez, "Recognizing the visual focus of attention for human robot interaction," in *International Conference on Human Behavior Understanding*, 2012.

[14] Z. Yucel, A. A. Salah, C. Mericli, T. Mericli, R. Valenti, and T. Gevers, "Joint attention by gaze interpolation and saliency," *IEEE TSMC-B*, 2013.

[15] D. B. Jayagopi et al., "The vernissage corpus: A multimodal human-robot-interaction dataset," Tech. Rep., IDIAP, 2012.

[16] K. P. Murphy, "Switching Kalman filters," Tech. Rep., UC Berkeley, 1998.

[17] D. Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *Control Theory Applications, IET*, 2010.