IDIAP RESEARCH REPORT

TRACKING ATTENTION FOR MULTIPLE PEOPLE: WANDERING VISUAL FOCUS OF ATTENTION ESTIMATION

Kevin Smith [a] Sileye Ba [a]
Jean-Marc Odobez [a] Daniel Gatica-Perez [a]

IDIAP–RR 06-39

JULY 2006

À PARAÎTRE DANS

[a] IDIAP Research Institute, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 06-39

TRACKING ATTENTION FOR MULTIPLE PEOPLE: WANDERING VISUAL FOCUS OF ATTENTION ESTIMATION

Kevin Smith Sileye Ba Jean-Marc Odobez Daniel Gatica-Perez

JULY 2006

À PARAÎTRE DANS

Résumé. The problem of finding the visual focus of attention of multiple people free to move in an unconstrained manner is defined here as the *wandering visual focus of attention* (WVFOA) problem. Estimating the WVFOA for multiple unconstrained people is a new and important problem with implications for human behavior understanding and cognitive science, as well as real-world applications. One such application, which we present in this article, monitors the attention passers-by pay to an outdoor advertisement. In our approach to the WVFOA problem, we propose a multi-person tracking solution based on a hybrid Dynamic Bayesian Network that simultaneously infers the number of people in a scene, their body locations, their head locations, and their head pose. It is defined in a joint state-space formulation that allows for the modeling of interactions between people. For inference in the resulting high-dimensional state-space, we propose a trans-dimensional Markov Chain Monte Carlo (MCMC) sampling scheme, which not only handles a varying number of people, but also efficiently searches the state-space by allowing person-part state updates. Our model was rigorously evaluated for tracking quality and ability to recognize people looking at an outdoor advertisement, and the results indicate good performance for these tasks.

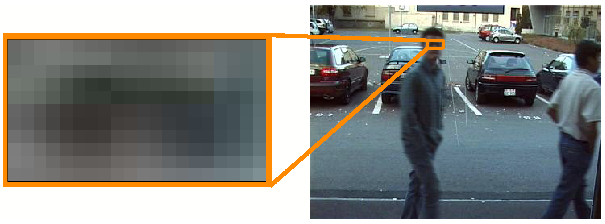


FIG. 1 – In the WVFOA problem, allowing the unconstrained motion of multiple people complicates the task of estimating a subject's *visual focus of attention* (VFOA). Here, the resolution is too low to estimate his focus of attention from eye gaze.

1 Introduction

An advertising firm has been asked to produce an outdoor display ad campaign for use in shopping malls and train stations. Internally, the firm has developed several competing designs, one of which must be chosen to present to the client. Is there some way to judge the best placement and content of these outdoor advertisements ?

Currently, the advertising industry relies on recall surveys or traffic studies to measure the effectiveness of outdoor advertisements [38, 39]. However, these hand-tabulated approaches are often too impractical or expensive to be commercially viable, and yield small data samples. A tool that automatically measures the effectiveness of printed outdoor advertisements does not exist, which leaves advertisers with few options to measure the effectiveness of such advertisements.

But in the television industry, such a system exists. The Nielsen ratings measure media effectiveness by estimating the size of the net cumulative audience of a program through surveys and Nielsen Boxes [32]. If one were to design an automatic Nielsen-like system for outdoor display advertisements, it might *automatically determine the number of people who have actually viewed the advertisement as a percentage of the total number of people exposed to it*.

This is an example of what we have termed the *wandering visual focus of attention* (WVFOA) problem, in which the tasks are :

1. to automatically detect and track an unknown, varying number of people able to move about freely,
2. and to estimate their visual focus of attention (VFOA).

The WVFOA problem is an extension of the traditional VFOA [36] problem in two respects. First, for WVFOA, the VFOA must be estimated for an unknown, varying number of subjects instead of a fixed number of static subjects. Second, in WVFOA, mobility is unconstrained. As a result, the subject's target of attention may be mobile, or as the subject moves about the scene, his appearance may change as he attempts to keep his attention on a specific target. Unconstrained motion also limits the resolution of the subject, as a wide field of view is necessary to capture multiple subjects over an area of interest. Limiting the resolution of the head makes estimating the VFOA from eye gaze more difficult (if not impossible in some cases), as seen in Figure 1.

Solutions to the WVFOA problem have implications for other scientific fields as well as real-life applications, including behavioral studies of humans and other animals, modeling human-computer interaction, modeling robot-human interaction, and security/surveillance, to name just a few. In the example of the outdoor advertisement application, the goal is to identify each person exposed to the advertisement and determine if they looked at it. Additionally, we can collect other useful statistics such as the amount of time they spent looking at the advertisement.

In this article, our goal is to propose and present a principled probabilistic framework for estimating WVFOA for multiple people. We applied our method to the advertising example to demonstrate its usefulness in real-life applications. Our solution assumes a fixed uncalibrated camera which can be placed arbitrarily, so long as the subjects appear clearly within the field of view. Our method requires a training phase in which the

appearance of people in the scene (and the orientation of their heads) is modeled. Our method consists of two parts : a Dynamic Bayesian Network, which simultaneously tracks people in the scene and estimates their head pose, and a WVFOA model, which infers a subject's VFOA from their location and head pose.

Besides defining the WVFOA problem itself, which to our knowledge is a previously unaddressed problem in the literature, we also present four key contributions in this article. First, we propose a principled probabilistic framework for solving the WVFOA problem by designing a mixed-state Dynamic Bayesian Network that jointly represents the people in the scene and their various parameters. The state-space is formulated in a true multi-person fashion, consisting of size and location parameters for the head and body, as well as head pose parameters for each person in the scene. This type of framework facilitates defining interactions between people.

Because the dimension of the state representing a single person is sizable, the multi-object state-space can grow to be quite large when several people appear together in the scene. Efficiently inferring a solution in such a framework can be a difficult problem. As our second contribution, we present an efficient method to do inference in this high-dimensional model. This is done using a trans-dimensional Markov Chain Monte Carlo (MCMC) sampling technique, an efficient global search algorithm robust to the problem of getting trapped in local minima.

Third, we demonstrate the real-world applicability of our model by applying it to the outdoor advertisement problem described earlier. We show that we are able to gather useful statistics such as the number of viewers and the total number of people exposed to the advertisement.

Finally, we thoroughly evaluate our model in the context of the outdoor advertisement application using realistic data and a detailed set of objective performance measures.

The remainder of the article is organized as follows. In the next section we will discuss related works. In Section 3, we describe our joint multi-person and head-pose tracking model. In Section 4, we present a method for modeling a person's VFOA. In Section 5 we describe our procedure for learning and parameter selection. In Section 6 we test our model on captured video sequences of people passing by an outdoor advertisement and evaluate its performance. Finally, Section 7 contains some concluding remarks.

2 Related Work

To our knowledge, this work is the first attempt to estimate the wandering visual focus of attention for multiple people as defined in Section 1. However, there is an abundance of literature concerning the three component tasks of the WVFOA problem : multi-person tracking, head pose tracking, and estimation of the visual focus of attention.

2.1 Multi-Person Tracking

Multi-person tracking is the process of locating a variable number of moving persons or objects over time. Multi-person tracking is a well studied topic, and a multitude of approaches to this problem have met with varying degrees of success. Here, we will restrict our discussion to probabilistic tracking methods which use a particle filter (PF) formulation [18, 37, 13, 21, 43]. Some computationally inexpensive methods use a single-object state-space model [21], but suffer from the inability to resolve the identities of different objects or model interactions between objects. As a result, much work has been focused on adopting a rigorous Bayesian joint state-space formulation to the problem, where object interactions can be explicitly defined [18, 37, 13, 15, 45, 43, 29]. However, sampling from a joint state-space can quickly become inefficient as the dimension of the space increases when more people or objects are added [18]. Recent work has concentrated on using MCMC sampling to track multiple people more efficiently. In [15] ants were tracked using MCMC sampling and a simple observation model. In [45], multiple humans were tracked from overhead as they crossed a college campus. In a previous work [29], we extended this model to handle varying number of people using a reversible-jump MCMC sampling technique. In this paper, we significantly extend the model of [29] by handling a much more complex object models and a larger state-space. This has necessitated the non-trivial design of new jump

types and proposal distributions, inter- and intra-personal interactions, the decoupling of MCMC move types, and the design of a complex observation model (see Section 3.4).

2.2 Head-Pose Tracking

Head-pose tracking is the process of locating a person's head and estimating its orientation in space. Head-pose tracking can be neatly categorized in two of the following ways : feature-based vs. appearance-based approaches or parallel vs. serial approaches. In feature-based approaches, a set of facial features such as the eyes, nose, and mouth are tracked. Making use of anthropometric measurements on these features, the relative positions of the tracked features can be used to estimate the head-pose [7, 12, 35]. A feature-based approach employing stereo vision was proposed [42]. The major drawback of the feature-based approach is that it requires high resolution head images, which can be difficult or impossible to acquire in some situations like the ad application in this paper. Also, occlusions and other ambiguities present difficult challenges to the feature-based approach.

In the appearance-based approach to head-pose tracking, instead of concentrating on specific facial features which require high-resolution images and may not be visible as a result of occlusion, the appearance of the entire head is modeled and learned from training data. Due to its robustness, there is an abundance of literature on appearance-based approaches. Several authors proposed using neural networks [26, 17, 44], others used principal component analysis [5, 31], and still others used multi-dimensional Gaussian distributions [40, 3].

In the serial approach to head-pose tracking, the tasks of head tracking and pose estimation are performed sequentially. This is also known as a "head tracking then pose estimation" framework, where head tracking is accomplished through a generic tracking algorithm, and features are extracted from the tracking results to perform pose estimation. This methodology has been used by several authors [35, 26, 17, 44, 34, 31, 40, 3]. In approaches relying on state-space models, the serial approach may have a lower computational cost over the parallel approach as a result of a smaller configuration space, but head-pose estimation depends on the tracking quality.

In the parallel approach, the tasks of head tracking and pose estimation are performed jointly. In this approach, knowledge of the head-pose can be used to improve localization accuracy, and vice-versa. Though the configuration space may be larger in the parallel approach, the computational cost of the two approaches may ultimately be comparable as a result of the parallel approaches improved accuracy through jointly tracking and estimating the pose. Benefits of this method can be seen in [42, 5] and [2]. In this work, we adopt an appearance-based parallel approach to head-pose tracking, where we jointly track the bodies, the heads, and estimate the poses of the heads of multiple people within a single framework.

2.3 Visual Focus of Attention

Strictly speaking, a person's VFOA is determined by their eye gaze. Estimating VFOA is of interest to several domains including advertising, psychology, and computer vision.

A persons VFOA can be most directly measured by tracking the subject's eye gaze. Various methods have been devised to accomplish this. In one such method, infrared light is shined directly into the subject's eyes, and the difference of reflection between the cornea and the pupil is used to determine the direction of the gaze. The system can be wearable [25], or the subject may be required to keep their head still on a chin-rest as advertisements are placed in front of them [8]. Other, less invasive procedures for estimating the visual focus of attention rely on the appearance of the eyes in a camera. In one such example, the aim was to automatically determine the loss of driver attention by using motion and skin color to localize the head and reconstruct the gaze direction from the eye locations [30]. In a similar example, the VFOA of a worker sitting in front of his computer in an office environment was measured [20].

However, all of the previously mentioned methods have one common drawback : they constrain the movement of the subject. Using infrared images to determine VFOA requires an expensive and invasive setup, and using eye appearance to determine VFOA requires high resolution images of the face. According to the definition of WVFOA, the subjects movement must be unconstrained. This rules out infrared methods or high

resolution images of the subjects face, (see Figure 1). Either method would fail for the advertising application described in Section 1.

In [36], Stiefelhagen et al.showed that *visual focus of attention can be reasonably approximated by head-pose* in a meeting room scenario. Others have followed this assumption, such as [6], where VFOA was found through the head-pose in an office setting. Using this important assumption, in this work we are able to simultaneously estimate the VFOA for multiple people without restricting their motion.

2.4 Other Related Work

While we believe that this work is the first attempt to estimate the WVFOA for multiple people, there exist several previous works in a similar vein. The 2002 Workshop on Performance and Evaluation of Tracking Systems (PETS) defined a number of estimation tasks on data depicting people passing in front of a shop window, including 1) determining the number of people in the scene, 2) determining the number of people in front of the window, and 3) determining the number of people looking at the window. Several methods attempted to accomplish these tasks through various means, including [19, 24, 23]. However, among these works there were no attempts to use head-pose or eye gaze to detect when people were looking at the window ; this was done using *only* body location and *assuming* that a person in front of the window is looking at it. Another previous work studied the detection and tracking of shopping groups in a store and estimation of transaction time [11]. Again, body motion was used to determine a subject's actions. Finally, a preliminary version of the work appearing in this article is reported in [27].

3 Joint Multi-Person and Head-Pose Tracking

In a Bayesian approach to multi-person tracking, the goal is to estimate the conditional probability for *joint multi-person configurations* of people $\mathbf{X}_t$, taking into account a sequence of observations $\mathbf{Z}_{1:t} = (\mathbf{Z}_1, ..., \mathbf{Z}_t)$. This is known as the filtering distribution $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$. In our model, a joint multi-person configuration, or joint state, is the union of the set of individual states describing each person in the scene. The observations consist of information extracted from an image sequence. The posterior distribution is expressed recursively by

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \quad = \quad C^{-1}p(\mathbf{Z}_t|\mathbf{X}_t) \times \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})d\mathbf{X}_{t-1}, \tag{1}$$

where the *motion model*, $p(\mathbf{X}_t|\mathbf{X}_{t-1})$, governs the temporal evolution of the joint state $\mathbf{X}_t$ given the previous state $\mathbf{X}_{t-1}$, and the *observation likelihood*, $p(\mathbf{Z}_t|\mathbf{X}_t)$, expresses how well the observed features $\mathbf{Z}_t$ fit the predicted state $\mathbf{X}_t$. Here C is a normalization constant.

In practice, the filtering distribution of Eq. 1 is often intractable. However, it can be approximated by applying the Monte Carlo method, in which the target distribution (Eq. 1) is represented by a set of N samples $\{\mathbf{X}_t^{(n)}, n = 1, ..., N\}$, where $\mathbf{X}_t^{(n)}$ denotes the n-th sample. For efficiency, in this work we use the Markov Chain Monte Carlo (MCMC) method, where the set of samples have equal weights and form a so-called Markov chain. Given the set of samples at $t-1$ $p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) \approx \sum_n \delta(\mathbf{X}_{t-1} - \mathbf{X}_{t-1}^{(n)})$, the Monte Carlo approximation of Eq. 1 is written

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \approx C^{-1}p(\mathbf{Z}_t|\mathbf{X}_t) \sum_n p(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)}). \tag{2}$$

Following this approach, the remainder of Section 3 is devoted to describing the details of our joint multi-person and head-pose tracking model. In the next sub-Section, we will discuss exactly how we model an individual person and the set of multiple people in the scene. In Section 3.2 we explain how to model motion and interactions between people. We describe our observation likelihood model, which estimates how well a proposed configuration fits the observed data, in Section 3.3. In Section 3.4, we discuss how to form the Markov Chain representing the distribution of Eq. 2 using the Metropolis-Hastings (MH) algorithm, and in Section 3.5 we show how to infer a point estimate solution from the posterior distribution.

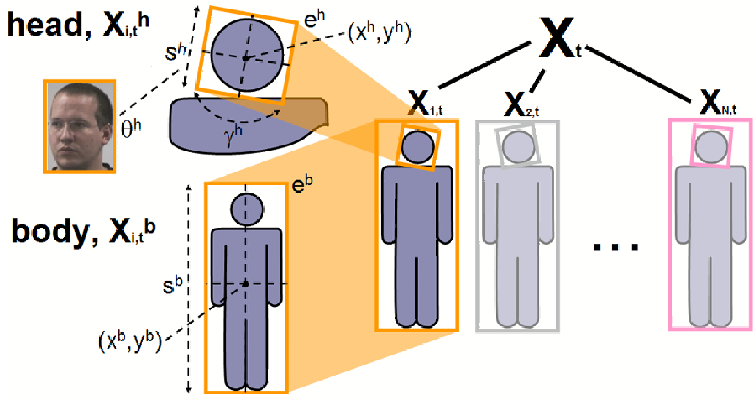


FIG. 2 – *State model for varying numbers of people*. The joint multi-person state, X_t consists of an arbitrary number of people $X_{i,t}$, each of which contain a body $\mathbf{X}_{i,t}^b$ and head $\mathbf{X}_{i,t}^h$ component. The body is modeled as a bounding box with parameters for the location (x^b, y^b), height scale s^b, and eccentricity e^b. The head location L^h has similar parameters for location (x^h, y^h), height s^h, and eccentricity e^h, as well as in-plane rotation γ^h. The head also has an associated *exemplar* θ^h, which models the out-of-plane head rotation.

3.1 State Model for a Varying Number of People

The state at time t describes the joint multi-object configuration of people in the scene. Because the amount of people in the scene may vary, we explicitly use a state model designed to accommodate changes in dimension [29], instead of a model with a fixed dimension, as in [15]. The joint state vector $\mathbf{X}_t$ is defined by $\mathbf{X}_t = \{X_{i,t} | i \in \mathcal{I}_t\}$, where $X_{i,t}$ is the state vector for person i at time t, and $\mathcal{I}_t$ is the set of all person indexes. The total number of people present in the scene is $m_t = |\mathcal{I}_t|$, where $|\cdot|$ indicates set cardinality. A special case exists when there are no people present in the scene, denoted by $\mathbf{X}_t = \emptyset$.

In our model, each person is represented by two components : a body $\mathbf{X}_{i,t}^b$, and a head $\mathbf{X}_{i,t}^h$ as seen in Figure 2. Note that we drop the i and t subscripts for the remainder of this section for simplicity. The body component is represented by a bounding box, whose state vector contains four parameters, $\mathbf{X}^b = (x^b, y^b, s^b, e^b)$. The point (x^b, y^b) is the continuous 2D location of the center of the bounding box in the image, s^b is the height scale factor of the bounding box relative to a reference height, and e^b is the eccentricity defined by the ratio of the width of the bounding box over its height.

The head component of the person model is represented by a bounding box which may rotate in the image plane, along with an associated discrete *exemplar* used to represent the head-pose (see Section 3.3.2 for more details). The state vector for the head is defined by $\mathbf{X}^h = (L^h, \theta^h)$ where $L^h = (x^h, y^h, s^h, e^h, \gamma^h)$ denotes the continuous 2D configuration of the head, including the continuous 2D location (x^h, y^h), the height scale factor s^h, the eccentricity e^h, and the in-plane rotation γ^h. A discrete variable, θ^h represents the head-pose exemplar which models the out-of-plane head rotation.

3.2 Dynamics and Interaction

The dynamic model governs the evolution of the state between time steps. It is responsible for predicting the motion of people (and their heads) as well as governing transitions between the head-pose exemplars. It is also responsible for modeling *inter-personal* interactions between the various people, as well as *intra-personal* interactions between the body and the head. The overall dynamic model for multiple people is written

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) \stackrel{def}{=} p_V(\mathbf{X}_t|\mathbf{X}_{t-1}) p_0(\mathbf{X}_t), \tag{3}$$

where p_V is the predictive distribution responsible for updating the state variables based on the previous time step, and p_0 is a prior distribution modeling interactions.

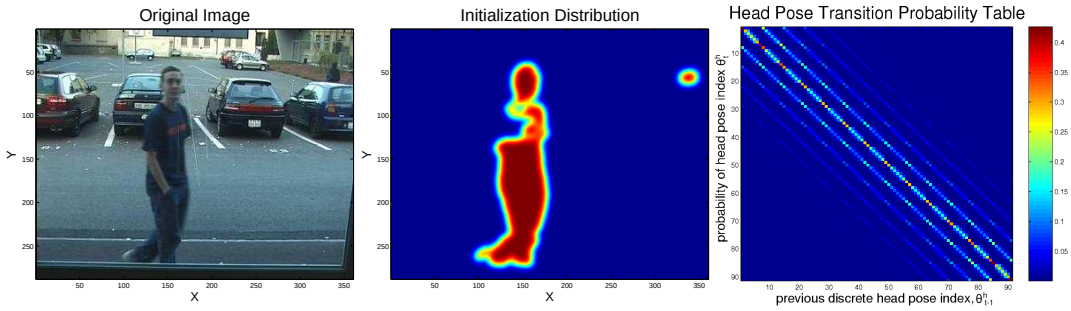


FIG. 3 – The *intitialization distribution* in *(center)* determines where to place a new person in the scene by applying a Gaussian smoothing filter to the foreground segmentation of the original image *(left)*. The *head-pose transition probability table* in *(right)* shows how the head-pose exemplars switch from the previous to current time step $p(\theta^h_t|\theta^h_{t-1})$. Visible lines show that transitions are most probable between similar poses.

To model the multi-person predictive distribution, we follow the approach of [29] where p_V is defined as

$$p_V(\mathbf{X}_t|\mathbf{X}_{t-1}) = \prod_{i\in\mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}), \tag{4}$$

when people are present in the previous time step ($\mathbf{X}_{t-1} \neq \emptyset$), and constant otherwise. However, in this work, the motion for a single person i, $p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1})$, is dependent only on its own previous state $p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1})$ if that person existed in the previous frame. If not, $p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1})$ is formed from a 2D *initialization distribution*, $p_{init}(\mathbf{X}_{i,t})$, formed by smoothing the foreground segmented image of the previous frame as seen in Figure 3.

The motion model for a single person is given by

$$p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) = p(\mathbf{X}^b_{i,t}|\mathbf{X}^b_{i,t-1})p(L^h_{i,t}|L^h_{i,t-1})p(\theta^h_{i,t}|\theta^h_{i,t-1}), \tag{5}$$

where the dynamics of the body state $\mathbf{X}^b_i$ and the head spatial state component L^h_i are modeled as 2^{nd} order auto-regressive (AR) processes. The head-pose exemplars, θ^h_i, are modeled by a discrete 1^{st} order AR process giving the transition probability table, seen in Figure 3.

The interaction model $p_0(\mathbf{X}_t)$ handles two types of interactions, *inter-personal* p_{0_1} and *intra-personal* p_{0_2} : $p_0(\mathbf{X}_t) = p_{0_1}(\mathbf{X}_t)p_{0_2}(\mathbf{X}_t)$. For modeling *inter-personal interactions*, we follow the method proposed in [15]. In this method, the inter-personal interaction model $p_{0_1}(\mathbf{X}_t)$ serves the purpose of restraining trackers from fitting the same person by penalizing overlap between trackers. This is achieved by exploiting a pair-wise Markov Random Field (MRF) whose graph nodes are defined at each time step by the people, and the links by the set $\mathcal{C}$ of pairs of proximate people. By defining an appropriate potential function $\phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \propto \exp(-g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}))$, the interaction model $p_{0_1}(\mathbf{X}_t) = \prod_{ij\in\mathcal{C}} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t})$ enforces a constraint in the multi-person dynamic model, based on locations of a person's neighbors. This constraint is defined by a non-negative penalty function, g, which is based on the spatial overlap of pairs of people (g is for no overlap, and increases as the area of overlap increases, see [29] for details).

We also introduce *intra-personal interactions* to the overall motion model. The intra-personal interaction model is meant to constrain the head model w.r.t. the body model, so that they are configured in a physically plausible way (i.e. the head is not detached from the body, or located near the waist). The intra-personal interaction model $p_{0_2}(\mathbf{X}_t)$ is defined as $p_{0_2}(\mathbf{X}_t) = \prod_{k\in\mathcal{I}_t} p(L^h_{k,t}|\mathbf{X}^b_{k,t})$, and penalizes head configurations which fall outside of an accepted domain defined by the configuration of the body. The penalization increases when the center of the head falls further than a distance outside of the top third of the body bounding box.

With these terms defined, the Monte Carlo approximation of the overall tracking filtering distribution in Eq.

2 can now be expressed

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \approx C^{-1}p(\mathbf{Z}_t|\mathbf{X}_t)p_0(\mathbf{X}_t)\sum_n p_V(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)}) \quad (6)$$

$$= C^{-1}p(\mathbf{Z}_t|\mathbf{X}_t)\prod_{ij\in\mathcal{C}}\phi(\mathbf{X}_{i,t},\mathbf{X}_{j,t})\prod_{k\in\mathcal{I}_t}p(L_{k,t}^h|\mathbf{X}_{k,t}^b)\sum_n p_V(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)}). \quad (7)$$

3.3 Observation Model

The observation model estimates the likelihood of a proposed configuration, or how well the proposed configuration is supported by evidence from the observed features. Our observation model consists of a *body model* and a *head model*, formed from a set of five total features. The body model consists of *binary* and *color* features, which are global in that they are defined pixel-wise over the entire image. The binary features ($\mathbf{Z}_t^{bin}$) make use of a foreground segmented image, while color features ($\mathbf{Z}_t^{col}$) exploit histograms in hue-saturation (HS) space. The head model is local in that its features (Z^h) are gathered independently for each person. They are responsible for the localization of the head and estimation of the head-pose, and include *texture* $\mathbf{Z}_t^{tex}$, *skin color* $\mathbf{Z}_t^{sk}$, and *silhouette* $\mathbf{Z}_t^{sil}$ features. For the remainder of this section, the time index (t) has been omitted to simplify notation. Assuming conditional independence of body and head observations, the overall likelihood is given by

$$p(\mathbf{Z}|\mathbf{X}) \triangleq p(\mathbf{Z}^{col}|\mathbf{Z}^{bin},\mathbf{X})p(\mathbf{Z}^{bin}|\mathbf{X})p(\mathbf{Z}^h|\mathbf{X}), \quad (8)$$

where the first two terms constitute the body model and the third term represents the head model. The body model, the head model, and each of the five component features are detailed in the following subsections.

3.3.1 Body Model

The body observation model is responsible for detecting and tracking people, adding or removing people from the scene, and maintaining consistent identities. It is comprised of a binary feature and a color feature.

Body Binary Feature

The binary feature is responsible for tracking bodies, and adding and removing people from the scene. We introduced the binary feature in a previous work [29], which relies on an adaptive foreground segmentation technique described in [33]. At each time step, the image is segmented into sets of foreground pixels F and background pixels B from the images ($I = F \cup B$), which form the foreground and background observations ($\mathbf{Z}^{bin,F}$ and $\mathbf{Z}^{bin,B}$)

For a given multi-person configuration and foreground segmentation, the binary feature computes the distance between the observed overlap (between the area of the multi-person configuration $S^{\mathbf{X}}$ obtained by projecting $\mathbf{X}$ onto the image plane and the segmented image) and a learned value. Qualitatively, we are following the intuition of a statement such as : "We have observed that two well-placed trackers (tracking two people) should contain approximately 65% foreground and 35% background." The overlap is measured for F and B in terms of precision ν and recall ρ : $\nu^F = \frac{S^{\mathbf{X}}\cap F}{S^{\mathbf{X}}}$, $\rho^F = \frac{S^{\mathbf{X}}\cap F}{F}$, $\nu^B = \frac{S^{\mathbf{X}}\cap B}{S^{\mathbf{X}}}$, and $\rho^B = \frac{S^{\mathbf{X}}\cap B}{B}$. An incorrect location or person count will result in ν and ρ values that do not match the learned values well, resulting in a lower likelihood and encouraging the model to choose better multi-person configurations.

The binary likelihood is computed for the foreground and background case $p(\mathbf{Z}^{bin}|\mathbf{X}) \triangleq p(\mathbf{Z}^{bin,F}|\mathbf{X})p(\mathbf{Z}^{bin,B}|\mathbf{X})$ where the definition of the binary foreground term, $p(\mathbf{Z}^{bin,F}|\mathbf{X})$, for all non-zero person counts ($m \neq 0$) is a single Gaussian distribution in precision-recall space (ν^F,ρ^F). The binary background term, on the other hand, is defined as a set of Gaussian mixture models (GMM) learned for each possible person count ($m \in \mathcal{M}$). For example, if the multi-person state hypothesizes that two people are present in the scene, the binary background likelihood term is the GMM density of the the observed ν^B and ρ^B values from the GMM learned for $m = 2$. For details on the learning procedure, see Section 5.

Body Color Feature

The color feature is responsible for maintaining the identities of people over time, as well as assisting the binary feature in localization of the body. The color feature uses HS color observations from the segmented foreground and background regions ($\mathbf{Z}^{col,F}$ and $\mathbf{Z}^{col,B}$) . Assuming conditional independence between foreground

and background, the color likelihood is written $p(\mathbf{Z}^{col}|\mathbf{Z}^{bin},\mathbf{X}) = p(\mathbf{Z}^{col,F}|\mathbf{Z}^{bin,F},\mathbf{X})p(\mathbf{Z}^{col,B}|\mathbf{Z}^{bin,B},\mathbf{X})$. The first term (foreground color likelihood) determines how well the color of each measured person matches online learned models, and the second term (background color likelihood) determines how well the background matches an off-line learned background model.

The *foreground color likelihood* compares an extracted 4D multi-person color histogram to an adaptive learned model, by $p(\mathbf{Z}^{col,F}|\mathbf{Z}^{bin,F},\mathbf{X}) \propto e^{\lambda_F d_F^2}$, where d_F is the Bhattacharya distance between the learned model and observed histogram and λ_F is a hyper-parameter [4]. The adaptive model is chosen from a small set of adaptive foreground color models (multiple models allow slightly different color models to compete). The model histograms are defined over a person index i, spatial segment (roughly corresponding to the head, torso, and legs), hue (H), and saturation (S) color spaces (quantized into 8 bins). Every frame, a vote is cast for the model that best matches the extracted data, and a 'winning' adaptive model is chosen from the set based on the number of 'votes' it has collected (the 'winner' is used in the above likelihood expression).

The *background color likelihood* helps reject configurations with untracked people by penalizing unexpected colors (i.e. those found on a person). The background model is a static 2D HS color histogram, learned from empty training images. The background color likelihood is defined as $p(\mathbf{Z}_t^{col,B}|\mathbf{Z}_t^{bin,B},\mathbf{X}_t) \propto e^{\lambda_B d_B^2}$, where λ_B and d_B^2 are defined as in the foreground case.

3.3.2 Head Model

The head model is responsible for localizing the head and estimating the head-pose. The head likelihood is defined as

$$p(\mathbf{Z}^h|\mathbf{X}) = \left[\prod_{i\in\mathcal{I}} p(\mathbf{Z}_i^{tex}|\mathbf{X}_i)p(\mathbf{Z}_i^{sk}|\mathbf{X}_i)p(\mathbf{Z}_i^{sil}|\mathbf{X}_i).\right]^{\frac{1}{m}}. \quad (9)$$

The individual head likelihood terms are geometrically averaged by $\frac{1}{m}$ to balance the overall likelihood as the number of people, m, varies. This non-standard likelihood modeling has been used previously [36], and is needed given the person-based head observation models.

The head model consists of three features : *texture* $\mathbf{Z}_t^{tex}$, *skin color* $\mathbf{Z}_t^{sk}$, and *silhouette* $\mathbf{Z}_t^{sil}$. The silhouette feature, proposed in this work, helps localize the head using foreground segmentation. The texture and skin color features, which have appeared in previous works including our own [2, 40], use appearance dependent observations to determine the head-pose of the subject.

We represent the head-pose as the angles resulting from the Euler decomposition of the head rotation w.r.t. the camera frame, known as pan α^h, tilt β^h, and roll γ^h (see Figure 4). The parameter γ^h models in-plane rotation and is modeled in the head spatial component (bounding box), L_i^h. To model out-of-plane rotations (pan α^h and tilt β^h), a *pointing vector* represents head-pose models constructed for each of the 93 discrete head-poses $\theta^h \in \Theta = \{\theta_j^h = (\alpha_j^h, \beta_j^h), j = 1, ..., 93\}$ from the Prima-Pointing Database [9] as seen in Figure 4.

Head-Pose Texture Feature

The head-pose texture feature reports how well the texture of an extracted image patch matches the texture of the head-pose hypothesized by the tracker. We represent texture using three filters : a coarse scale Gaussian filter, a fine Gabor filter, and a course Gabor filter (see Fig. 5).

Texture models were learned for each of the discrete head-pose values θ^h. Training was done on head patch images from the Prima Pointing Database resized to a reference size (64×64). The training images were preprocessed by histogram equalization to reduce light variation effects, and the filters were applied on a sub-sampled grid (to reduce computation) and concatenated into a single feature vector. Then, for each head-pose θ ($\theta = \theta^h$ here, for simplicity), the mean $e^\theta = (e_j^\theta)$ and diagonal covariance matrix $\sigma_\theta = (\sigma_j^\theta)$ of the corresponding training feature vectors were computed and used to define the person texture likelihood model from Eq.9 as

$$p(\mathbf{Z}_i^{tex}|\mathbf{X}_i) = \prod_j \frac{1}{\sigma_j^{\theta_i}} \max(\exp -\frac{1}{2}\left(\frac{\mathbf{Z}_{i,j}^{tex} - e_j^{\theta_i}}{\sigma_j^{\theta_i}}\right)^2, T_{tex}), \quad (10)$$

where T_{tex} is a threshold used to reduce the impact of outlier measurements.

Head-Pose Skin Feature

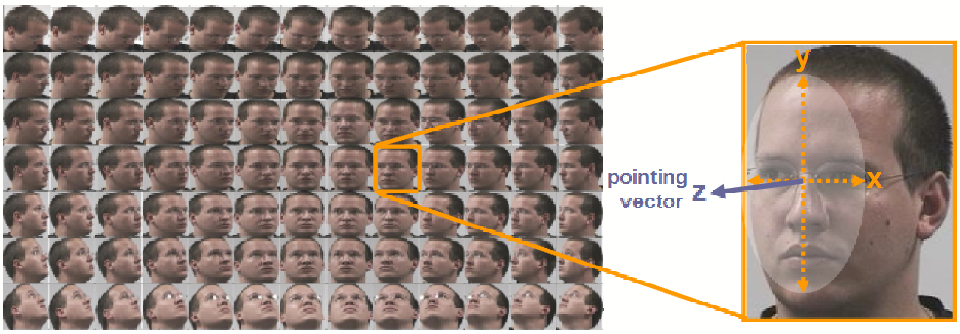
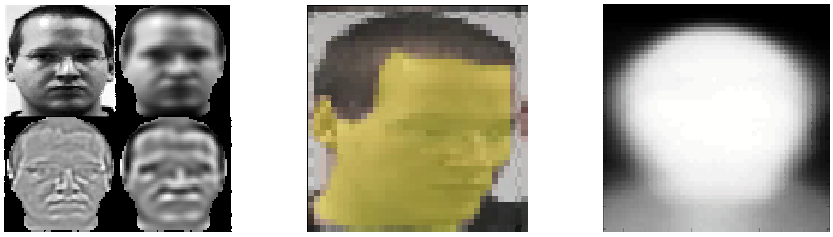


FIG. 4 – *Head-Pose Modeling*. (Left) Discrete head appearances from the Prima-Pointing Database [9]. Each appearance is represented by a discrete head-pose index θ^h. (Right) Head-pose is represented by the Euler angle decomposition of the head rotation w.r.t. the camera frame (angles pan α^h, tilt β^h, and roll γ^h). The *pointing vector* z^h is defined by pan α^h and tilt β^h.



(a) *texture* (b) *skin color* (c) *silhouette*

FIG. 5 – *(a)* Texture is used to estimate the head-pose by applying three filters to the original image (upper left). These filters include a coarse scale Gaussian filter (upper right), a fine scale Gabor filter (lower left), and a coarse scale Gabor filter (lower right). *(b)* Skin color models help to keep the head-pose robust in presence of background clutter. *(c)* A silhouette model is responsible for localizing the head's spatial component.

The texture feature is a powerful tool for modeling the head-pose, but prone to confusion due to background clutter. To help make our head model more robust, we have defined a skin color binary model (or mask), M^θ, for each head-pose, θ, in which the value at a given location indicates a skin pixel (1), or a non-skin pixel (0). An example of a skin color mask can be seen in Figure 5. The skin color binary models were learned from skin color masks extracted from the same training images used in the texture model using a Gaussian skin-color distribution modeled in normalized RG space [41].

The head-pose skin color likelihood compares the learned model with a measurement extracted from the image $\mathbf{Z}_i^{sk}$ (skin color pixels are extracted from the image using a temporally adaptive skin color distribution model). The skin color likelihood of a measurement $\mathbf{Z}_i^{sk}$ belonging to the head of person i is defined as

$$p(\mathbf{Z}_i^{sk}|\mathbf{X}_i) \propto \exp -\lambda_{sk}||\mathbf{Z}_i^{sk} - M^{\theta_i}||_1, \tag{11}$$

where $||.||_1$ denotes the L_1 norm and λ_{sk} is a hyper parameter learned on training data.

Head-Pose Silhouette Feature

In addition to the pose dependent head model, we propose to add a head silhouette likelihood model to aid in localizing the head by taking advantage of foreground segmentation information. We built a head silhouette model, H^{sil} (see Figure 5) by averaging head silhouette patches extracted from binary foreground segmentation images in the training set (note that a single model is used, unlike the pose-dependent models for texture and skin color).

The silhouette likelihood works by comparing the model H^{sil} to an extracted binary image patch (from the foreground segmentation) corresponding to the hypothesized location of the head, $\mathbf{Z}_i^{sil}$. A poor match indicates foreground pixels in unexpected locations, probably due to poor placement of the head model. The head sil-

houette likelihood term is defined as :

$$p(\mathbf{Z}_i^{sil}|\mathbf{X}_i) \propto \exp -\lambda_{sil}||\mathbf{Z}_i^{sil} - H^{sil}||_1, \tag{12}$$

where λ_{sil} is an hyper-parameter learned on training sequences.

In practice, we found that introducing this term (not defined in our previous work [2] or in others' like [40]) greatly improved the head localization in the combined body-head optimization process.

3.4 Trans-Dimensional MCMC

In the introduction to Section 3, we showed that the Bayesian tracking distribution (Equation 1) can be approximated by Equation 2 using Monte Carlo Methods. In the previous sub-sections, we defined the components of Equation 2 : we defined a state model, a state evolution/motion model, and a likelihood function for evaluating hypotheses. This lead us to restate Equation 1 as Equation 7. In this sub-section, we describe how Reversible-Jump Markov Chain Monte Carlo (RJMCMC) can be used to efficiently generate a Markov Chain which represents the distribution of Equation 7. In Section 3.5, we will show how to infer a solution to the tracking problem from the Markov Chain, and in Section 3.6 we give an overview of the algorithm.

The multi-person state-space can quickly become very large when we allow for an arbitrary number of people. The state vector for a single person is ten-dimensional. Traditional Sequential Importance Resampling (SIR) particle filters are inefficient in such high-dimensional spaces [1]. Markov Chain Monte Carlo (MCMC) particle filters are more efficient, but do not allow for the dimensionality of the state-space to vary (fixing the number of people). To solve this problem, we have adopted the RJMCMC sampling scheme, originally proposed by [45] and also used in our own previous work [29], which retains the efficiency of MCMC sampling but allows for 'reversible jumps' which can vary the dimensionality of the state-space. However, instead of updating the state of an entire person simultaneously as in [45] and [29], in this work we propose to generalize the RJMCMC approach to update individual components of the state of a single person. The benefits of this approach include an improved robustness to becoming trapped in local minima, and a way to handle what we refer to here as the *likelihood balancing problem*.

Constructing a Markov Chain with Metropolis-Hastings

As previously mentioned, the stationary distribution of the Markov Chain must be defined over the configuration space $\mathbf{X}_t$, it must be able to vary in dimension, and it must approximate the filtering distribution defined in Eq. 7 (please note that for the remainder of this section, we omit the time subscript t for simplicity). For the task of constructing the Markov Chain, we turn to the Metropolis-Hastings (MH) algorithm [1].

Starting from an arbitrary initial configuration[1] $\mathbf{X}$, the MH algorithm samples a new configuration $\mathbf{X}^*$ from a proposal distribution $q(\mathbf{X}^*|\mathbf{X})$, and adds the proposed sample to the Markov Chain with probability

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)q(\mathbf{X}|\mathbf{X}^*)}{p(\mathbf{X})q(\mathbf{X}^*|\mathbf{X})}\right), \tag{13}$$

otherwise, the current configuration $\mathbf{X}$ is added to the Markov Chain (with probability $1-\alpha$). This is known as the *acceptance test*. A Markov Chain is constructed by repeatedly adding samples to the chain in this fashion.

In practice, a new configuration $\mathbf{X}^*$ is chosen by first selecting a *move type*, υ^* from a set of reversible moves Υ (defined in the next sub-section) with prior probability p_{υ^*}. Moves must be defined in such a way that every move type which changes the dimensionality of the state has a corresponding reverse move type [10]. The acceptance ratio α can be re-expressed through *dimension-matching* [10] as

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)p_\upsilon q_\upsilon(\mathbf{X})}{p(\mathbf{X})p_{\upsilon^*}q_{\upsilon^*}(\mathbf{X}^*)}\right), \tag{14}$$

where q_υ is a move-specific distribution, defined for all move types (*except swap*) as

$$q_\upsilon(\mathbf{X}_t^*) = \sum_i q_\upsilon(i)q_\upsilon(\mathbf{X}_t^*|i), \tag{15}$$

over all people i, in such a way that the move is applied to a *target index* i^*, while the rest of the multi-person configuration is fixed.

[1] In our work, we initialize the Markov Chain at time t by sampling uniformly the Markov Chain at $t-1$.

Splitting the Six Reversible Move Types

We define six move types in our model : *birth*, *death*, *swap*, *body update*, *head update*, and *pose update*. In previous works using RJMCMC [29, 16], a single update move was defined in which *all* the parameters of a randomly selected person were updated simultaneously. This was sufficient for less complex object models, but a problem arises when multiple features are used to evaluate different aspects of the state. For example, let us imagine an update move designed to apply motion to the body, apply motion to the head, and adjust the head-pose simultaneously. Even if applying such a move results in a better overall likelihood, there is no guarantee that the individual body, head, and pose configurations have improved. Some may have remained the same, or even worsened. Because the ranges of the terms of the overall likelihood vary, some will dominate. The result might be a model which tracks bodies well, but poorly estimates the head-pose (which we observed in practice). We refer to this as the *likelihood balancing problem*.

To overcome this problem, we propose to decouple the update move into three separate moves : body update, head update, and pose update. In this way, we divide the task of finding a good configuration for an entire person into three smaller problems : finding a good configuration for the body, for the head location, and for the head-pose.

We now define the six move types. For each move type, we explain how to choose a *target index*, and define the move-specific proposal distributions and acceptance ratios. Due to space limitations, some details appear in a technical report available at http ://www.idiap.ch/∼smith/.

(1) Birth. A new person i^* is added to the new configuration $\mathbf{X}^*$ which was not present in the old configuration $\mathbf{X}$: ($\mathcal{I}_t^* = \mathcal{I}_t \cup \{i^*\}$). This implies a dimension change from m_t to $m_t + 1$.

To add a new person, a birth location, x_b^*, is sampled from the birth proposal distribution described in Figure 3 using a stratified sampling strategy. Next, a new person index, or *target index*, must be chosen, which may be a previously 'dead' person, or an 'unborn' person. The target index i^* is sampled from the *birth* target proposal model, $q_{birth}(i)$, which is defined in such a way that the probability of choosing a 'dead' person depends on their temporal distance and spatial distance to the sampled birth location (recently killed, nearby people are most likely to be reborn). The probability of choosing an 'unborn' person is unity minus the sum of 'dead' probabilities.

Having chosen a target index, the birth move is applied to i^* while the rest of the multi-person configuration is fixed ; in Eq. 15 this is done by defining $q_v(\mathbf{X}_t^*|i)$ as

$$q_{birth}(\mathbf{X}_t^*|i) = \frac{1}{N}\sum_n p(\mathbf{X}_{i,t}^*|\mathbf{X}_{t-1}^{(n)}) \prod_{l\in\mathcal{I}_t} p(\mathbf{X}_{l,t}|\mathbf{X}_{t-1}^{(n)})\delta(\mathbf{X}_{l,t}^* - \mathbf{X}_{l,t}). \tag{16}$$

Initial body parameters of a new object (born or reborn) are sampled from learned Gaussian distributions. Initial head and pose parameters are chosen to maximize the head likelihood.

Now that the identity and parameters of the new person i^* have been determined, it can be shown that the acceptance ratio for the new multi-person configuration $\mathbf{X}_t^*$ is given by

$$\alpha_{birth} = \min\left(1, \frac{p(\mathbf{Z}_t|\mathbf{X}_t^*)\prod_{j\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t}^*, \mathbf{X}_{j,t}^*)}{p(\mathbf{Z}_t|\mathbf{X}_t)} \frac{p_{death}}{p_{birth}} \frac{q_{death}(i^*)}{q_{birth}(i^*)}\right), \tag{17}$$

where $\mathcal{C}_{i^*}$ and ϕ are pairs of proximate objects and the interaction potential defined in Section 3.2, and $q_{death}(i^*)$ is the reverse-move *death* target proposal model. Note that the interaction model helps discourage births that overlap existing people and complexity is reduced as many terms in $p(\mathbf{Z}_t|\mathbf{X}_t^*)$ and $p(\mathbf{Z}_t|\mathbf{X}_t)$ cancel.

(2) Death. An existing person i^* is removed from the new configuration $\mathbf{X}^*$ which was present in the old configuration $\mathbf{X}$ ($\mathcal{I}_t^* = \mathcal{I}_t \setminus \{i^*\}$) where $\setminus$ is the difference between sets. This implies a dimension change from m_t to $m_t - 1$. The target index i^* is chosen with probability $q_{death}(i)$ by uniformly sampling from the set of 'live' people (i.e. $q_{death}(i) = \frac{1}{m_t}$). Person i^* is removed keeping the rest of the multi-person configuration fixed, with mixture components defined as

$$q_{death}(\mathbf{X}_t^*|i) = \frac{1}{N}\sum_n \prod_{l\in\mathcal{I}_t, l\neq i} p(\mathbf{X}_{l,t}|\mathbf{X}_{t-1}^{(n)})\delta(\mathbf{X}_{l,t}^* - \mathbf{X}_{l,t}), \tag{18}$$

and the acceptance probability can shown to simplify to

$$\alpha_{death} = \min\left(1, \frac{p(\mathbf{Z}_t|\mathbf{X}_t^*)}{p(\mathbf{Z}_t|\mathbf{X}_t)\prod_{j\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t},\mathbf{X}_{j,t})}\frac{p_{birth}}{p_{death}}\frac{q_{birth}(i^*)}{q_{death}(i^*)}\right). \tag{19}$$

(3) Swap. The configurations of a pair of objects, i^* and j^* are swapped. The proposal is a mixture model over pairs of objects $q_s(\mathbf{X}_t^*) = \sum_{i,j} q_{swap}(i,j)q_{swap}(\mathbf{X}_t^*|i,j)$. Candidates are chosen with probability $q_{swap}(i,j)$, which is defined such that the probability a pair of people are chosen is a function of their proximity (nearby pairs are more likely to be selected). When the move is applied, the mixture component $q_{swap}(\mathbf{X}_t^*|i,j)$ swaps the configurations of objects i^* and j^*. It can be shown that the acceptance ratio for the swap move is reduced to

$$\alpha_{swap} = \min\left(1, \frac{p(\mathbf{Z}_t|\mathbf{X}_t^*)}{p(\mathbf{Z}_t|\mathbf{X}_t)}\right). \tag{20}$$

(4) Body update. The body parameters $\mathbf{X}_{i,t}^b$: including its location (x^b, y^b), height s^b, and eccentricity e^b are updated. The body update move proposal is defined as $q_{body}(\mathbf{X}^*) = \sum_i \frac{1}{m_t} q_{body}(\mathbf{X}^*|i)$ with

$$q_{body}(\mathbf{X}^*|i) = \frac{1}{N}\sum_n p(\mathbf{X}_{i^*,t}^{b,*}|\mathbf{X}_{t-1}^{(n)})p(\overline{\mathbf{X}_{i^*,t}^{b,*}}|\mathbf{X}_{t-1}^{(n)})\delta(\overline{\mathbf{X}_{i^*,t}^{b,*}} - \overline{\mathbf{X}_{i^*,t}^{b}})\prod_{l\in\mathcal{I}_t\setminus i^*} p(\mathbf{X}_{l,t}|\mathbf{X}_{t-1}^{(n)})\delta(\mathbf{X}_{l,t}^* - \mathbf{X}_{l,t}), \tag{21}$$

where $\overline{\mathbf{X}_{i^*,t}^{b}}$ denotes all state parameters except $\mathbf{X}_{i^*,t}^{b}$, and $\mathbf{X}_{i^*,t}^{b,*}$ denotes the proposed body configuration for target i^*. In practice, this implies first selecting a person randomly, i^*, and sampling a new body configuration for this person from $p(\mathbf{X}_{i^*,t}^{b,*}|\mathbf{X}_{t-1}^{b,n^*})$, using an appropriately sample n^* from the previous time and keeping all the other parameters unchanged. With this proposal, the acceptance probability α_{body} can then be shown to reduce to :

$$\alpha_{body} = \min\left(1, \frac{p(\mathbf{Z}_t^b|\mathbf{X}_{i^*,t}^{b,*})p(L_{i^*,t}^{h,*}|\mathbf{X}_{i^*,t}^{b,*})\prod_{j\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t}^*,\mathbf{X}_{j,t}^*)}{p(\mathbf{Z}_t^b|\mathbf{X}_{i^*,t}^{b})p(L_{i^*,t}^{h}|\mathbf{X}_{i^*,t}^{b})\prod_{j\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t},\mathbf{X}_{j,t})}\right). \tag{22}$$

(5) Head update. This move implies sampling the new head spatial configuration of person i^* in a similar fashion according to $p(L_{i^*,t}^*|L_{t-1}^{n^*})$. The acceptance ratio α_{head} simplifies to

$$\alpha_{head} = \min\left(1, \frac{p(\mathbf{Z}_{i^*,t}^{h,*}|\mathbf{X}_{i^*,t}^{h,*})p(L_{i^*,t}^{h,*}|\mathbf{X}_{i^*,t}^{b,*})}{p(\mathbf{Z}_{i^*,t}^{h}|\mathbf{X}_{i^*,t}^{h})p(L_{i^*,t}^{h}|\mathbf{X}_{i^*,t}^{b})}\right). \tag{23}$$

(6) Pose update. The last move consists of simply sampling the new head-pose from the proposal function $p(\theta_{i^*,t}^*|\theta_{t-1}^{n^*})$ (see Fig. 3) and accepting with probability α_{pose} :

$$\alpha_{pose} = \min\left(1, \frac{p(\mathbf{Z}_{i^*,t}^{h}|\mathbf{X}_{i^*,t}^{h,*})}{p(\mathbf{Z}_{i^*,t}^{h}|\mathbf{X}_{i^*,t}^{h})}\right). \tag{24}$$

3.5 Inferring a Solution

In RJMCMC, the first N_b samples added to the Markov Chain (using the MH algorithm) are part of the *burn-in* cycle, which allows the Markov Chain to reach the target density. The *filtering distribution* is approximated by the N_p samples taken after the burn-in point. The Markov Chain, however, does not provide a single answer to the tracking problem.

For this reason, we compute a *point estimate*, which is a single multi-person configuration calculated from the stationary distribution that serves as the tracking output. To determine the (discrete) configuration of people in the scene, we search for the the most common configuration of people, taking into account swapped identities, births, and deaths. Using these samples, we determine the (continuous) body $\mathbf{X}_{i,t}^b$ and head spatial configurations $L_{i,t}^h$ for the various people in the scene (including head roll $\gamma_{i,t}$) by taking the Marginal Mean of each parameter. For the out-of-plane head rotations represented by the discrete exemplar θ_i, we take the Marginal Mean of the corresponding Euler angles for pan and tilt.

Algorithm 1 : Multi-Person Body/Head Tracking and WVFOA Estimation with RJMCMC
At each time step t, the posterior distribution of Eq. 7 is represented by a Markov Chain consisting of a set of $N = N_b + N_p$ samples $\{\mathbf{X}_t^{(n)}, n = N_b, ..., N\}$. Body, head, and pose parameters are inferred from the Markov Chain after it reaches the *burn-in point*. Using these values, the WVFOA model determines if the persons attention is *focused* on the advertisement or *unfocused*.

1. `Initialize` the MH sampler by choosing a sample from the $t-1$ Markov Chain with the MPM number of people (m_{t-1}^{MPM}). Apply the motion model and accept it as sample $n = 1$.
2. `Metropolis-Hastings Sampling`. Draw $N = N_b + N_p$ samples according to the following schedule (where N_b is the *burn-in point*) :
   - `Begin` with the state of the previous sample $X_t^{(n)} = X_t^{(n-1)}$.
   - `Choose Move Type` by sampling from the set of moves $\Upsilon = \{$*birth, death, swap, body update, head update, pose update*$\}$ with prior probability p_{υ^*}.
   - `Select a Target` i^* (or set of targets i^*, j^* for swap) according to the target proposal $q_v(i)$ for the selected move type.
   - `Sample New Configuration` $\mathbf{X}^*{}_t$ from the move-specific proposal distribution q_{υ^*}. For the various move types, this implies :
     - *Birth* - add a new person i^* $m_t^{(n)*} = m_t^{(n)} + 1$ according to Eq. 16.
     - *Death* - remove an existing person i^* $m_t^{(n)*} = m_t^{(n)} - 1$ according to Eq. 18.
     - *Swap* - swap the parameters of two existing people i^*,j^* $\mathbf{X}_{i,t}^{(n)} \rightarrow \mathbf{X}_{j,t}^{(n)*}$, $\mathbf{X}_{j,t}^{(n)} \rightarrow \mathbf{X}_{i,t}^{(n)*}$.
     - *Body Update* - update the body parameters $X_{i,t}^{b,(n)*}$ of an existing person i^* (Eq. 21).
     - *Head Update* - update the head parameters $L_{i,t}^{h,(n)*}$ of an existing person i^*.
     - *Pose Update* - update the pose parameter $\theta_{i,t}^{(n)*}$ of an existing person i^*.
   - `Compute Acceptance Ratio` α according to Equations 17, 19, 20, 22 23, and 24.
   - `Add` n^{th} `Sample` to the Markov Chain : If $\alpha \geq 1$, then add the proposed configuration ($\mathbf{X}^*$). Otherwise, add the proposed $\mathbf{X}^*$ with probability α. If the proposed configuration is rejected, add the previous $\mathbf{X}$ (i.e. $\mathbf{X}_t^{(n-1)}$).
3. `Compute a Point Estimate Solution` from the Markov Chain (as in Section 3.5) :
   - determine the most common multi-person configuration $\hat{X}_t$ accounting for births, deaths, and swaps. Collect the samples of this configuration into a set W.
   - determine the body $\hat{X}_t^b$ and head $\hat{L}_t^h$ spatial configurations, and the out-of-plane head rotations for the various people in the scene by computing the Marginal Mean of the parameters over the set W (using Euler decompositions for pan $\hat{\alpha}^h$ and tilt $\hat{\beta}^h$).
4. `Determine the WVFOA` for each person in the scene (as in Section 4) :
   - determine the likelihood each person is in a focused state from their horizontal head location $\hat{x}_t^h$ and pointing vector $\hat{z}_t^h$ according to Equation 25.
   - if the likelihood is greater than a threshold $p(z^h) > T_{wvfoa}$, that person is *focused*, otherwise he/she is *unfocused*.

FIG. 6 – Algorithm for joint multi-person body and head tracking and WVFOA estimation with RJMCMC.

3.6 Pseudo-Code

The detailed steps of our joint multi-person body-head tracking and WVFOA estimation model is summarized in Figure 6.

4 Wandering Visual Focus of Attention (WVFOA) Modeling

The WVFOA task is to automatically detect and track a varying number of people able to move about freely, and to estimate their VFOA. The WVFOA problem is significantly more complex than the traditional VFOA problem because it allows for a variable number of moving people instead of a single stationary person.

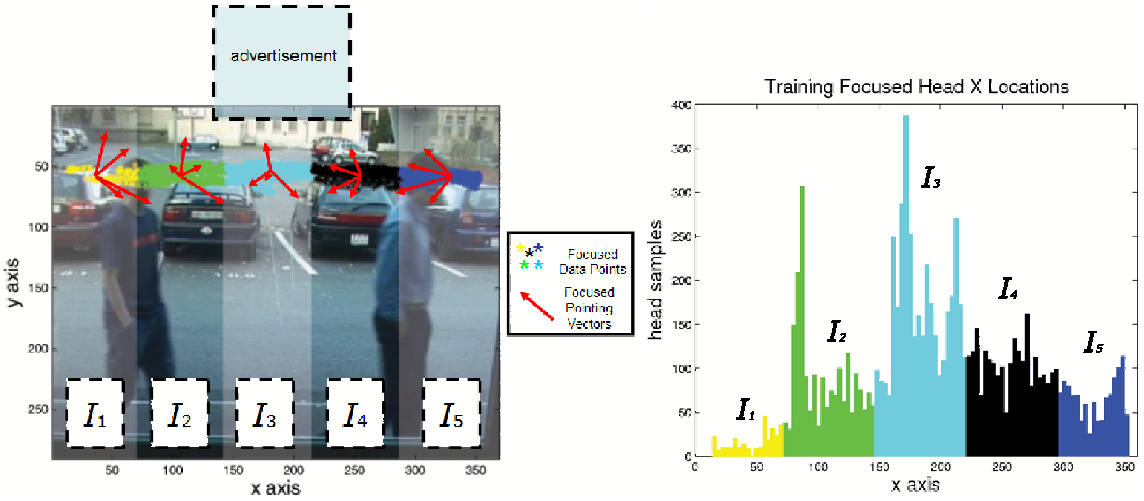


FIG. 7 – *WVFOA Modeling.* (Left) WVFOA is determined by *head-pose* and horizontal *position* in the image. The horizontal axis is split into 5 regions ($I_1, ..., I_5$), and a WVFOA model is defined for each of these regions. Yellow, green, cyan, black, and blue data points represent *focused* head locations used for training and red arrows represent 2D projections of typical samples of *focused* pointing vectors z^h. Note that the advertisement is affixed to a window and appears just above the image frame. (Right) Over 9400 training points representing a person in a *focused* state (also seen in the left pane) were split into 5 regions along the horizontal axis and used to train a Gaussian model for each region.

The advertising application we have chosen as an introduction to WVFOA represents a relatively simple instance of the problem because we are only attempting to measure the focus of attention on a single target : the advertisement. More complex WVFOA scenarios could have several targets, moving targets, or both.

For the advertising application, a person's WVFOA is defined as being in one of two states :

– *focused* - looking at the advertisement, or

– *unfocused* - not looking at the advertisement.

Note that this is just one of many ways in which the WVFOA can be represented, but it is sufficient to solve the tasks set forth in Section 1. A persons state of focus depends both on their *location* and on their *head-pose* as seen in Figure 7. For head location and head-pose information, we rely on the output of the RJMCMC tracker described in Section 3.

To model the WVFOA, we chose to check for only the *focused* state, though this method could be easily extended to model both focused and unfocused states. To determine if a person is in a focused state, we extract the pointing vector z^h from the pose estimate output by the RJMCMC tracker (see Fig. 4), which is characterized by the pan and tilt angles, as well as the horizontal head position x^h (see Figure 7). Because the target advertisement is stationary, the ranges of z^h corresponding to the *focused* state are directly dependent on the location of the head in the image. For this reason, we chose to split the image into $K = 5$ horizontal regions $I_k, k = \{1, ..., 5\}$, and modeled the likelihood of a focused state as

$$p(z^h) = \sum_{k=1}^{K} p(x^h \in I_k, z^h) = \sum_{k=1}^{K} p(x^h \in I_k) p(z^h | x^h \in I_k) \tag{25}$$

where the first term $p(x^h \in I_k)$ models the likelihood a person's head location belongs to region I_k, and the second term $p(z^h | x^h \in I_k)$ models the likelihood of *focused* head-pose given the region the head belongs to. The inclusion of the head location in modeling the WVFOA allowed us to solve an issue not previously addressed [22, 30, 36] : resolving the WVFOA of a person whose focused state depends on their location.

The terms of the WVFOA model in Equation 25 are defined as follows. The image horizontal axis, x, is divided into K regions I_k whose centers and width are denoted by x_{I_k} and σ_{I_k}, respectively. The probability of

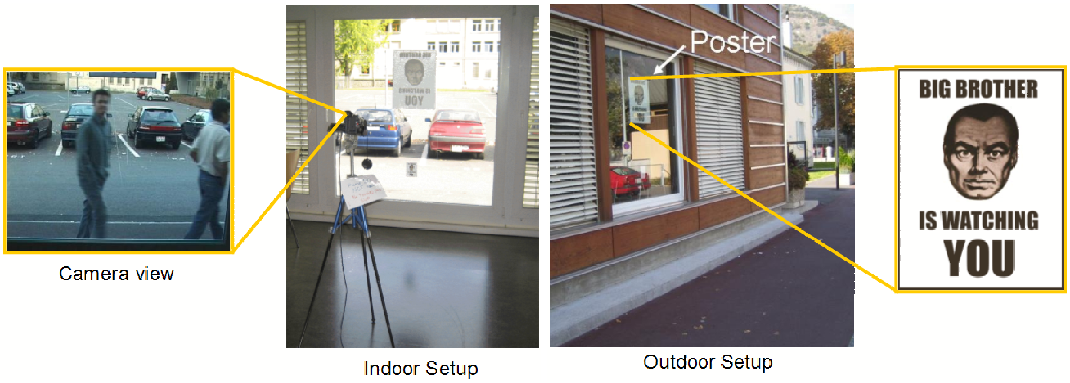


FIG. 8 – *Experimental Setup.* (Left) Inside the building, a camera is placed so that it is facing out a window. The view from the camera can be seen in the insert. (Right) Outside, the advertisement in the window is noticeable to people walking along the footpath. The fake advertisement poster can be seen in the insert.

a head location x^h belonging to region I_k is modeled by a Gaussian distribution $p(x^h \in I_k) = \mathcal{N}(x^h; x_{I_k}, \sigma_{I_k})$. For each region, the distribution of pointing vectors representing a *focused state* was modeled using a Gaussian distribution. Typical pointing vectors for each region are seen in Figure 7.

The parameters of the WVFOA model (Gaussian mean and covariance matrix) were learned from the training data described in the next section. Though our WVFOA model does not make use of the vertical head location, it is straightforward to generalize the models by allowing other partitions $\{I_k\}$ of the image plane. Finally, a person is determined to be *focused* when the corresponding likelihood $p(z^h)$ in Eq. 25 is greater than a threshold, T_{wvfoa}.

As an alternative, one might attempt to bypass the tracking model and find the location and head-pose using a face detector. However, *a face detector alone might not be sufficient to solve the WVFOA problem* for several reasons : (1) the WVFOA problem allows for a range of head-poses beyond that of typical face detectors (including situations where part or none of the face is visible - partially visible in our case) (2) unless they include an additional tracking stage, existing state-of-the-art face detectors such as that described in [14] have no mechanism to maintain identity between time steps or recover from occlusions. Properties of face detection *and* tracking are necessary to solve the WVFOA problem, and indeed elements of our head model share commonalities with face detectors.

5 Training and Parameter Selection

In this section we will describe our training procedure and how we determine the key parameters of our model. We begin by describing the setup of the experiment.

5.1 Experimental Setup

To simulate the advertising application, an experiment was set up as seen in Figure 8. A fake advertisement was placed in an exposed window with a camera set behind. The camera view can be seen in the left-hand insert, with the bottom edge of the poster appearing at the top of the image above the heads of the subjects.

In our experiments, actors were used due to privacy concerns for actual passers-by. The actors were instructed to pass in front of the window with the freedom to look at the advertisement (or not) as they would naturally. A recording of 10-minute duration (360×288 resolution, 25 fps) was made in which up to three people appear in the scene simultaneously. The recorded data includes several difficult tracking events such as people passing and occluding each other. Though simulated, every effort was made to ensure that the data was as fair a representation of a real-life scenario as possible.

TAB. 1 – Symbols, values, and descriptions for key parameters of our model.

Parameter	Value	Set by	Description
α_{scale}	0.01	learned	*motion model* body and head scale variance (AR2 process)
$\alpha_{position}$	2.4	learned	*motion model* body and head position variance (AR2 process)
K_{bf}	1	learned	*observation model* body binary model number of Gaussians (foreground)
K_{bb}	3	learned	*observation model* body binary model number of Gaussians (background)
λ_F	20	learned	*observation model* body color foreground hyper-parameter
λ_{sil}	200	learned	*observation model* head silhouette hyper-parameter
λ_{tex}	0.5	learned	*observation model* head texture hyper-parameter
T_{tex}	$exp(\frac{-9}{2})$	learned	*observation model* head texture threshold
λ_{sk}	0.5	learned	*observation model* head skin color hyper-parameter
p_{birth}	0.05	hand	*RJMCMC* prior probability of choosing a *birth* move
p_{death}	0.05	hand	*RJMCMC* prior probability of choosing a *death* move
p_{swap}	0.05	hand	*RJMCMC* prior probability of choosing a *swap* move
p_{body}	0.283	hand	*RJMCMC* prior probability of choosing a *body update* move
p_{head}	0.283	hand	*RJMCMC* prior probability of choosing a *head update* move
p_{pose}	0.283	hand	*RJMCMC* prior probability of choosing a *pose update* move
N_p	300,600,800	learned	*RJMCMC* number of samples in chain for 1,2,3 simultaneous people, resp.
N_b	0.25*N_p	hand	*RJMCMC* number of *burn-in* samples
K_{wvfoa}	5	hand	*WVFOA model* number of Gaussians
T_{wvfoa}	*0.00095*	learned	*WVFOA model* likelihood threshold

5.2 Training

The recorded video data was organized into a training and test set of equal size and disjoint from each other. The training set, consisting of nine sequences for a total of 1929 frames, was manually annotated for body location, head location, and focused/unfocused state.

The parameters for the foreground segmentation were tuned by hand by observing results on the training set. The binary body feature model was trained with the annotated body locations and foreground segmented binary images of the training set. Using this information, GMMs were were trained for precision and recall for the foreground and the background. Head annotations were used to learn the parameters of the Gaussian skin-color distribution in the head-pose skin feature. The silhouette mask was also trained using the head annotations (1929 frames), by averaging the binary patches corresponding to head annotations. Parameters for the WVFOA model, including T_{wvfoa}, were optimized on the training data (bootstrapped to 9400 training points, see Figure 7) to achieve the highest WVFOA event recognition performance (see Section 6 for details on event recognition performance). The training set was also used to learn prior sizes (scale and eccentricity) for the person models. Texture models and the skin color masks were learned from the Prima-Pointing Database, which consists of 30 sets of images of 15 people, each containing 93 frontal images of the same person in a different pose ranging from -90 degrees to 90 degrees (see Figure 4).

5.3 Parameter Selection

In addition to the trained models, the rest of the parameters our algorithm were chosen by hand. Some were selected using the training set without exhaustive tuning. Others (e.g. single-person dynamic model parameters) were assigned standard values. Unless explicitly stated, all parameters remain fixed for the evaluation described in the next section. In Table 1, a description of the key parameters mentioned in the text and their values are provided.

TAB. 2 – Test set data summary.

sequence	length (s)	# people total	# people simultaneous	# looks at ad	description
a	15	3	1	2	person from right (no look), person from left (looks), person from right (looks)
b	13	3	1	3	person from left (looks), person from right (looks), person from right (looks)
c	10	3	1	3	person from right (looks), person from left (looks), person from right (looks)
d	5	2	2	2	2 people cross from the right, both look at ad
e	6	2	2	3	2 people cross from the left, both look at ad (1^{st} looks twice)
f	4	2	2	2	2 people cross from the left, both look at ad
g	4	2	2	1	2 people cross from the right, 2^{nd} looks at ad
h	4	2	2	2	1 person from right (looks at ad), another from left (no look)
i	11	3	3	4	3 people appear from right, all look at ad (1^{st} looks twice)

6 Evaluation

As mentioned in the introduction, we applied our model to a hypothetical Nielsen-like outdoor advertisement application. The task was to determine the number of people who actually look at an advertisement as a percentage of the total number of people exposed to it.

In order to evaluate the performance of our application, a ground truth for the test set was hand annotated in a similar manner to the training set. The test set consists of nine sequences, a through i. Sequences a, b, and c contain three people (appearing sequentially) passing in front of the window. Sequences d through h contain two people appearing simultaneously. Sequence i contains three people appearing simultaneously. The details of the test set are summarized in Table 2. Our evaluation compared our results with the ground truth over 180 experiments on the 9 test sequences (as our method is a stochastic process, we ran 20 runs per sequence). The length of the Markov Chain was chosen such that there was a sufficient number of samples for good quality tracking according to the number of people in the scene (see Table 1). Experimental results are illustrated in Figures 9 and 13, and fully shown in companion videos, available at http ://www.idiap.ch/∼smith/.

In the remainder of this section, we will discuss the performance of the multi-person body and head tracking (Section 6.1), the advertisement application (Section 6.2), and the effect of varying the length of the Markov Chain (Section 6.3).

6.1 Multi-Person Body and Head Tracking Performance

To evaluate the multi-person body and head tracking performance we adopt a set of measures proposed in our previous work [28], with some minor changes to names and notation. These measures evaluate three tracking features : the ability to estimate the number and placement of people in the scene (*detection*), the ability to persistently track a particular person over time (*tracking*), and how tightly the estimated bounding boxes fit the ground truth (*spatial fitting*).

To evaluate detection, we rely on the rates of *False Positive* and *False Negative* errors (normalized per person, per frame) denoted by $\overline{FP}$ and $\overline{FN}$. The *Counting Distance* $\overline{CD}$ measures how close the estimated number of people is to the actual number (normalized per person per frame). A $\overline{CD}$ value of zero indicates a perfect match. To evaluate tracking, we report the *Tracker Purity* $\overline{TP}$ and *Object Purity* $\overline{OP}$, which estimate the degree of consistency with which the estimates and ground truths were properly identified ($\overline{TP}$ and $\overline{OP}$ near 1 indicate well maintained identity, near 0 indicate poor performance). For spatial fitting, the *F-measure* measures the overlap between the estimate and the ground truth for the body and head from recall ρ and precision ν, ($F = \frac{2\nu\rho}{\nu+\rho}$). A perfect fit is indicated by $F = 1$, no overlap by $F = 0$. For further details on these measures, omitted here for space reasons, see [28].

Per-sequence results appear in Fig. 10 with illustrations for sequence f in Fig. 9 and for sequences b, e, h, and i in Fig. 13. For detection (Fig. 10a), the FP and FN rates are reasonably low, averaging a total of 2.0 FN errors and 4.2 FP errors per sequence. These errors usually correspond to problems detecting exactly when a person enters or leaves the scene. The overall $\overline{CD}$, which indicates the average error in the estimation of the number of people in the scene, was 0.018 (where zero is ideal).

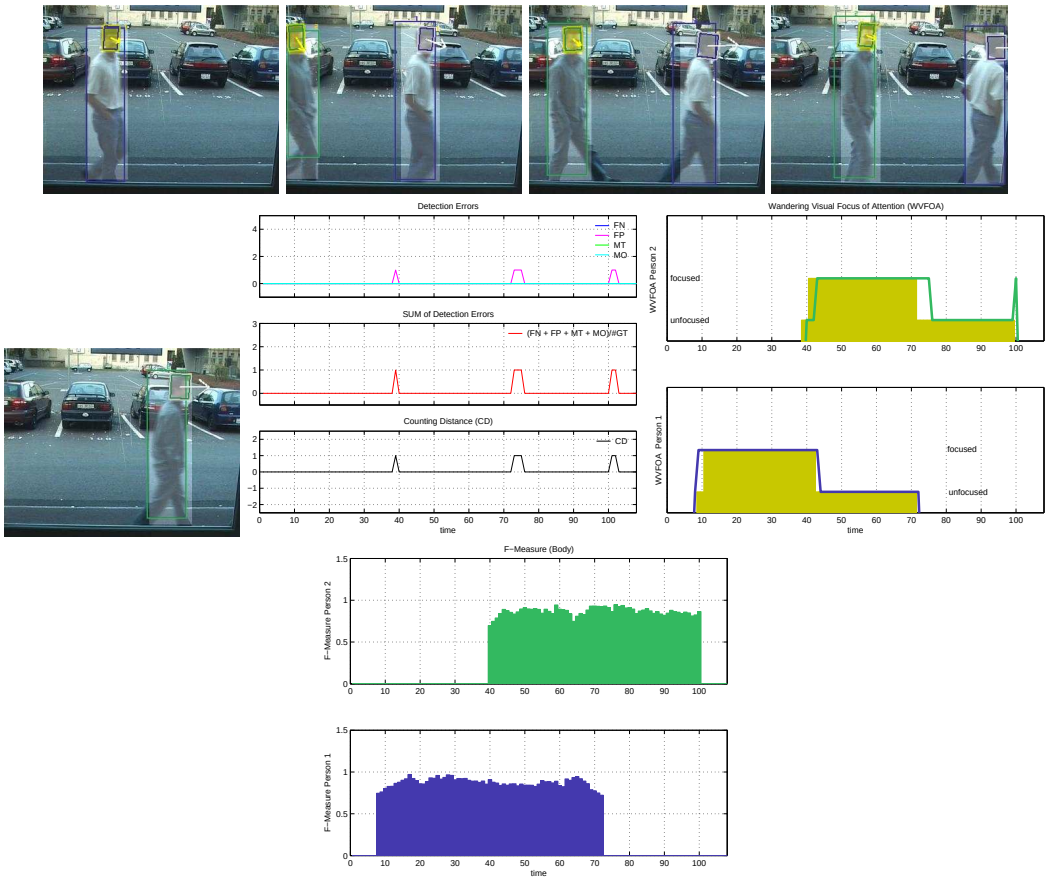


FIG. 9 – *Experimental Results*. Upper Row : Frames from Sequence f in which two people cross the scene from left to right, looking at the advertisement once each. Tracking results appear as green and blue boxes around the body and head (with an associated pointing vector). A yellow pointing vector/head border indicates a *focused* state, a white pointing vector/head border indicates an *unfocused* state. The ground truth appears as shaded boxes for the head and the body (the head area is shaded yellow when labeled as *focused* and grey when labeled as *unfocused*). Bottom Row, Left : The top plot contains a history of person detection errors over the course of the sequence, the middle plot contains a summation over all the errors, the bottom plot shows CD (see text for descriptions of these measures). Center : WVFOA results for both people over the duration of the sequence. The ground truth appears as yellow bars (raised indicates a *focused* state, lowered when *unfocused*, and non-existent when the object does not appear in the scene). The tracking results appear as blue and green lines. Right : F measures how tightly the bounding boxes fit the ground truth for each person.

For tracking (Fig. 10b), $\overline{TP}$ and $\overline{OP}$ are both of high quality. Combining $\overline{TP}$ and $\overline{OP}$ using the F-measure as for spatial fitting ($\frac{2\overline{TP}\,\overline{OP}}{\overline{TP}+\overline{OP}}$), we find that overall our model produced a high value (0.93). The main source of error in tracking was due to extraneous trackers appearing when people enter or leave the scene. A second source of error occurred when a person exited the scene followed by another person entering from the same place in a short period of time : the second person was often misinterpreted as the first. Sequence h, in which people crossed paths and occluded one another, saw a slight drop in performance compared to the other sequences, but we were still able to maintain 81.3% purity (other sequences ranged from 80.5% to 98.3%). These numbers indicate that our model was mostly successful in maintaining personal identity through occlusion, as seen in Fig. 13

Finally, for spatial fitting, the bounding boxes generally fit the ground truths tightly, as evidenced by Figures

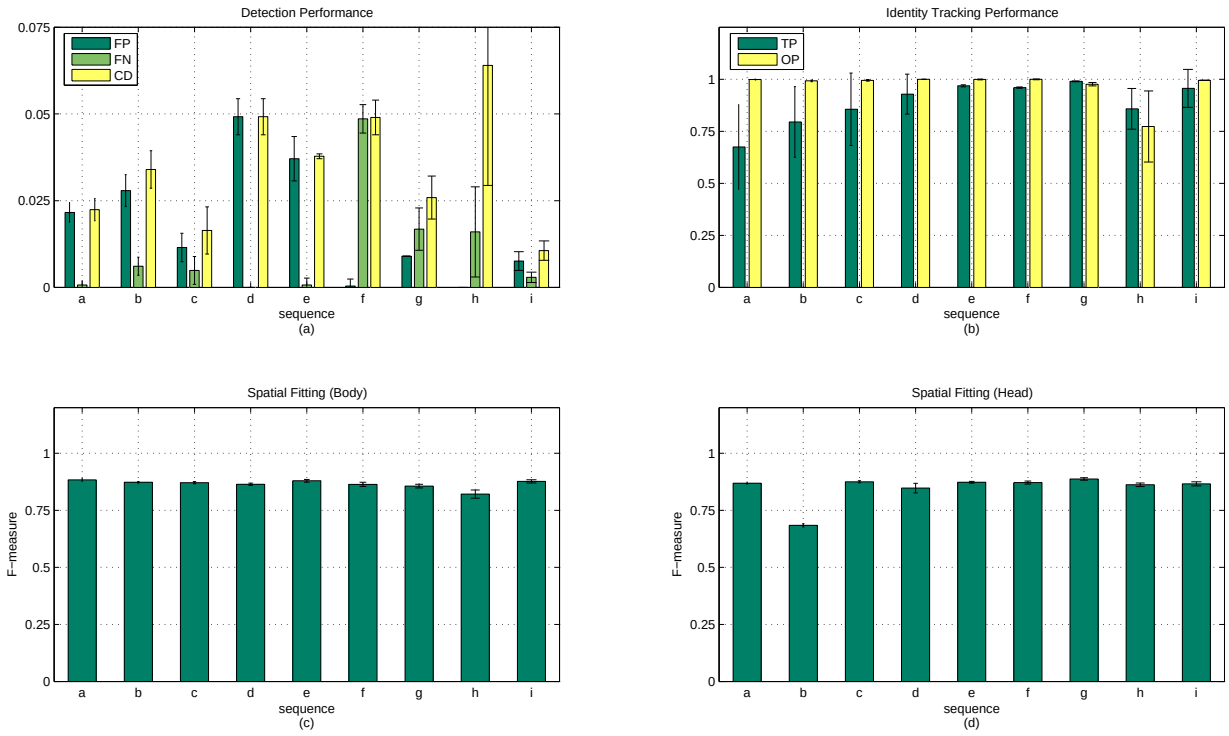


FIG. 10 – *Multi-Person Head and Body Tracking Results*. The *detection performance* plot in *(a)* measures the ability of the model to estimate the correct number and placement of people in the scene. Measures shown include the normalized *false positive* ($\overline{FP}$) and *false negative* ($\overline{FN}$) error rates (per person, per frame), and the *counting distance* ($\overline{CD}$) (near-zero values are good, see text). The *tracking performance* plot in *(b)* measures the ability of the model to persistently track people over time. *Tracker purity* ($\overline{TP}$) and *object purity* ($\overline{OP}$) measure the consistency with which the ground truths and estimates were properly identified. TP and OP values near 1 indicate good performance. Plots *(c)* and *(d)* show the *spatial fitting* results (how well the tracker bounding boxes fit the ground truth) for the body and the head over the nine sequences. Overlap between the estimate and ground truth are measured using the F-measure. A value of 1 indicates a perfect fit, a value of zero indicates no overlap. In each plot, the standard deviation is represented by error bars (cases where no error bar is visible indicates $std = 0$).

9, 10c and 10d. Both the body and head had a mean fit of 0.87 (1 being optimal). As seen in Figure 9, the fit often suffered from partially visible bodies and heads that occurred when people entered and exited the scene.

6.2 Advertisement Application Performance

To evaluate the performance of the advertisement application, the results from our model were compared with a ground truth where the WVFOA was labeled for each person as either *focused* or *unfocused*. In our evaluation, we considered the following criteria : (1) the number of people exposed to the advertisement, (2) the number of people who looked, or *focused*, at the advertisement, (3) the number of events where someone *focused* on the advertisement (look-events), and (4) the frame-based and (5) event-based recognition rates of the WVFOA. Results for the ad application evaluation appear in Figure 11 and in the companion videos in the website.

Regarding criterion 1, over the entire test set, 22 people passed the advertisement, while our model estima-

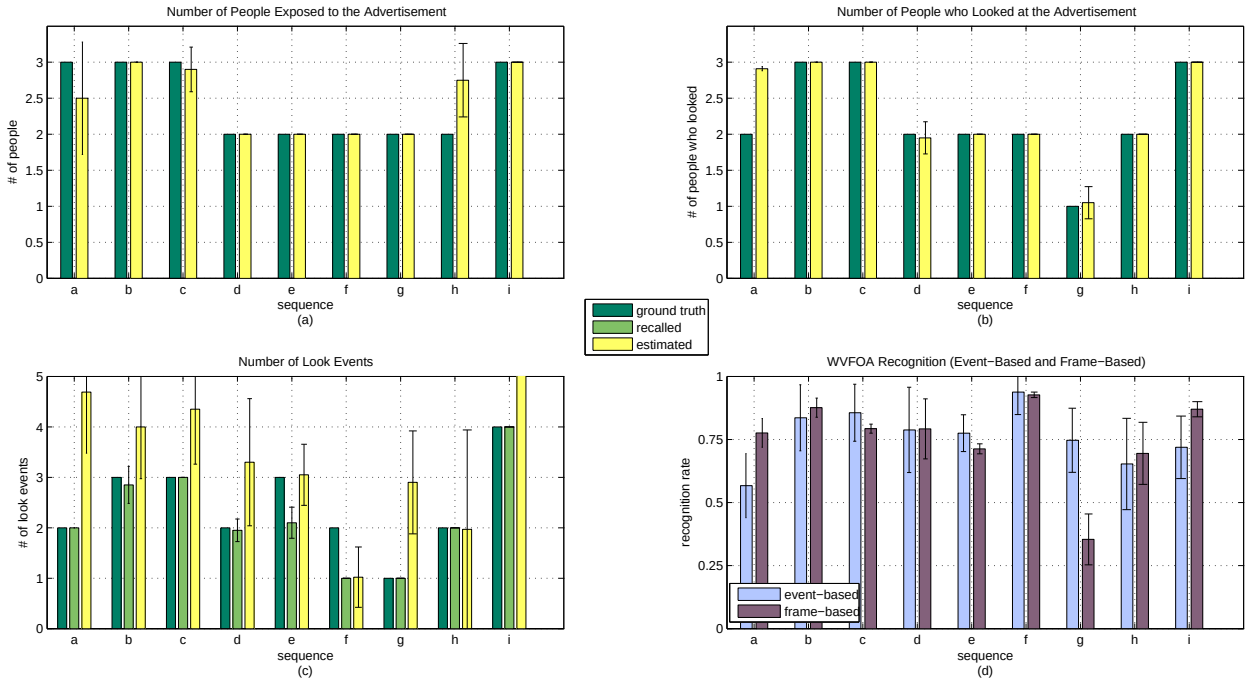


FIG. 11 – *Ad Application Results*. The first three plots show *(a)* the number of people exposed to the advertisement, *(b)* the number of people who looked at the advertisement, and *(c)* the number of "look events" for the nine sequences. The dark green bars represent the ground truth, while the yellow bars represent our model's estimates. In *(c)*, the light green bars represent the number of actual look events detected by our system. In each plot, the standard deviation is represented by error bars (cases where no error bar is visible indicates $std = 0$). Plot *(d)* shows the overall recognition rate of *focused* and *unfocused* states (calculated based on events and based on frame counts).

ted a value of 22.15 (average for all runs, std $= .17$), In Figure 11a we can see that the number of people was correctly estimated for all sequences except a, c, and h.

With respect to criterion 2, of the 22 total people, 20 actually *focused* on the advertisement. Our model estimated a value of 20.75 (std $= .09$). Figure 11b shows perfect results for all sequences except a, d, and h.

For criterion 3, we defined a look-event as a *focused* state for a continuous period of time of 3 frames or more. The total number of look-events in the test data set was 22, 21 of which our system recognized on average (std = $.89$). This result was determined through a standard symbol matching technique (see below). However, our model estimated 37 total look-events on average (std $= 1.1$). This disparity can be partially attributed to problems in head-pose estimation for heads partially outside the image as people enter or leave. The look-event estimation results would improve if we did not consider WVFOA in these cases. Also, the look event duration of 3 frames is quite strict, and some erroneous looks were generated by noise.

Finally, to evaluate the overall quality of WVFOA estimation, we compute recognition rates for event-based WVFOA and frame-based WVFOA using the aforementioned *F-measure* (criteria 4 and 5 from the previous page). To compute the event-based F, the ground truth and estimated WVFOA are segmented over the entire sequence into focused and unfocused events, symbol matching is performed accounting for temporal alignment, and F is computed on matched segments. Results are show in Figure 11d. The overall event-based F is 0.76 (std = $.13$). The frame-based F is computed by matching the estimated WVFOA for each frame to the ground truth. The overall frame-based F-measure is 0.76 (std $= .06$). Poor frame-based results in sequence g occurred because the subject *focused* for a very short time as he entered the field of view (0.3s), during which time his head was only partially visible. However, our model still managed to detect at the *event* level with $F = .75$.

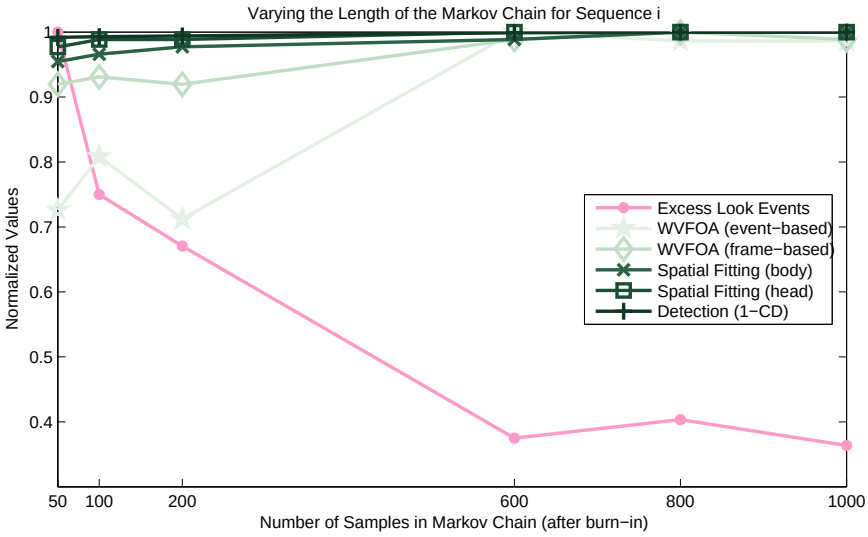


FIG. 12 – *Varying the Number of Samples in the Markov Chain*. As more samples used, various performance gauges increase, as seen here for sequence i. Excess (false alarm) look events (pink) detected by the system drop as more samples are added, the WVFOA recognition improves (both for event and frame based), spatial fitting for the body and head improves, and *detection* performance increases (as measured by $1 - \overline{CD}$). Note that the measures have been normalized to appear on the same axis.

6.3 Varying the Number of Particles

To study the model's dependency on the number of samples, we conducted experiments on sequence i (the most complex in terms of number of simultaneous people), varying the number of samples $N = \{50, 100, 200, 600, 800, 1000\}$. The results are shown in Fig. 12. For all N, the model correctly estimated the number of people who passed and the number of people who looked. With less samples, the spatial fitting and detection (as measured by $1 - \overline{CD}$) suffered. The head tracking and head-pose estimation was noticeably shakier with less samples, and the WVFOA estimation suffered as a consequence. This is shown by the increased error in the number of estimated looks for low sample counts. The model stabilized around approximately $N = 600$. The computational complexity was roughly linear to N, with a cost ranging from < 1 second ($N = 50$) to ≈ 5 seconds per frame ($N = 600$), non-optimized in Matlab.

7 Conclusion

In this article, we have introduced the WVFOA problem and presented a principled probabilistic approach to solving it. Our work thus contributes in terms of both problem definition and statistical vision modeling. Our approach expands on state-of-the-art RJMCMC tracking models, with novel contributions to object modeling, likelihood modeling, and the sampling scheme. We applied our model to a real-world application and provided a rigorous objective evaluation of its performance. From these results we have shown that our proposed model is able to track a varying number of moving people and determine their WVFOA with good quality. Our model is general, and can be adapted for other applications with similar tasks.

For future work, investigating the usefulness of using a spatially dependent face/pose detector as an additional feature is one possible avenue. Other work might include modeling multiple human-to-human interaction using WVFOA, or coupling head-pose to head motion.

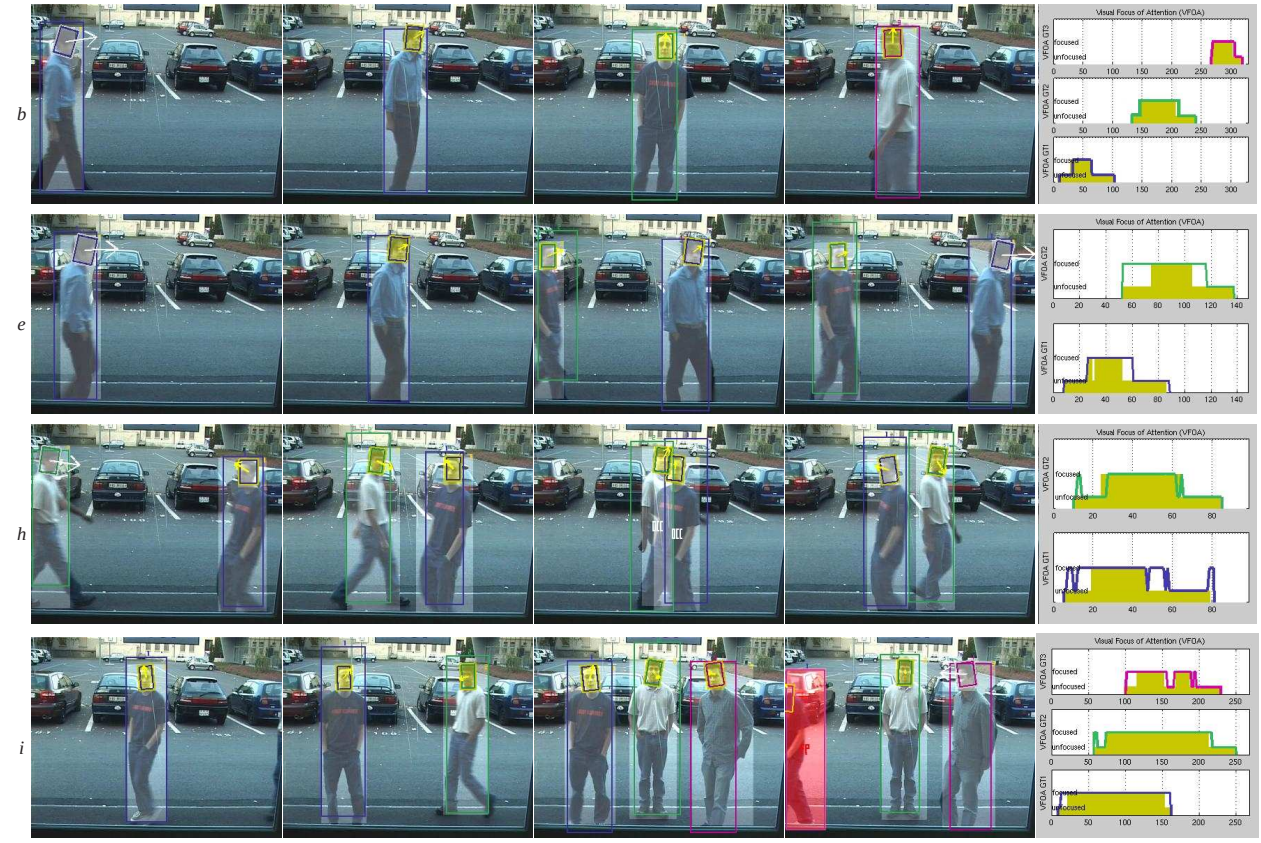


FIG. 13 – *Tracking and WVFOA Results.* Results for four sequences, b, e, h, and i. A summary plot of the WVFOA performance is provided in the last pane of each row. For details on interpreting the plots and symbols, refer to Fig 9. Here, we can see the WVFOA performance was nearly perfect for sequence b and exhibited slight errors in sequence i. The 2^{nd} person (green) in sequence e suffered from prematurely estimating a *focused* state. Sequence h suffered some ambiguities due to the loss of head tracking as people crossed paths. The last frame of sequence i shows a FP error generated as a tracker was placed where no ground truth was present (though the subject is half visible as he exits the scene). Such situations can cause ambiguity problems.

Références

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to mcmc for machine learning, 2003.

[2] S. Ba and J. Odobez. Evaluation of multiple cues head-pose tracking algorithms in indoor environments. In *Proc. of Int. Conf. on Multimedia and Expo (ICME)*, Amsterdam, July 2005.

[3] L. Brown and Y. Tian. A study of coarse head-pose estimation. In *Workshop on motion and video computing*, Orlando, Dec. 2002.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, June 2000.

[5] T. Cootes, G. Edwards, and J. Taylor. Active appearance model. In *Proc. of European Conference on Computer Vision (ECCV)*, Freibourg, June 1998.

[6] M. Danninger, R. Vertegaal, D. Siewiorek, and A. Mamuji. Using social geometry to manage interruptions and co-worker attention in office environments. In *Proc. of Conference on Graphics Interface*, Victoria BC, 2005.

[7] A. Gee and R. Cipolla. Estimating gaze from a single view of a face. In *Proc. International Conference on Pattern Recognition (ICPR)*, Jerusalem, Oct. 1994.

[8] L. Gerald. Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising*, 26 :61–73, 1997.

[9] N. Gourier, D. Hall, and J.L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Work. on Visual Observation of Deictic Gestures*, Aug. 2004.

[10] P. Green. Reversible jump mcmc computation and bayesian model determination. *Biometrika*, 82 :711–732, 1995.

[11] I. Haritaoglu and M. Flickner. Detection and tracking of shopping groups in stores. In *Proc. of Computer Vision and Patter Recognition (CVPR)*, Hawaii, Dec. 2001.

[12] A.T. Horprasert, Y. Yacoob, and L.S. Davis. Computing 3d head orientation from a monocular image sequence. In *Proc. of Intl. Society of Optical Engineering (SPIE)*, Killington, VT, 1996.

[13] M. Isard and J. MacCormick. Bramble : a bayesian multi-blob tracker. In *Proc. Intl. Conference on Computer Vision (ICCV)*, Vancouver, Jul. 2001.

[14] M.J. Jones and P. Viola. Fast multi-view face detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, June 2003.

[15] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *Proc. European Conference on Computer Vision (ECCV)*, Prague, May 2004.

[16] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 27 :1805–1819, 2005.

[17] V. Kruger, S. Bruns, and G. Sommer. Efficient head-pose estimation with gabor wavelet networks. In *Proc. of the British Machine Vision Conference (BMVC)*, Bristol, Sep. 2000.

[18] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. Intl. Conference on Computer Vision (ICCV)*, Kerkyra, Greece, Sep. 1999.

[19] L. Marcenaro, L. Marchesotti, and C. Regazzoni. Tracking and counting multiple interacting people in indoor scenes. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, June 2002.

[20] Y. Matsumoto, T. Ogasawara, and A. Zelinsky. Behavior recognition based on head-pose and gaze direction measurement. In *Proc. of Conference on Intelligent Robots and Systems*, 2002.

[21] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe. A boosted particle filter : multi-target detection and tracking. In *Proc. European Conference on Computer Vision (ECCV)*, Prague, May 2004.

[22] K. Otsuka, J. Takemae, and H. Murase. A probabilistic inference of multi party-conversation structure based on markov switching models of gaze patterns, head direction and utterance. In *Intl. Conference on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[23] A.E.C. Pece. From cluster tracking to people counting. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, June 2002.

[24] J. Piater, S. Richetto, and J. Crowley. Event-based activity analysis in live video using a generic object tracker. In *Performance Evaluation of Tracking and Surveillance (PETS) Workshop*, Copenhagen, June 2002.

[25] G.M. Pieters, E. Rosbergen, and M. Hartog. Visual attention to advertising : the impact of motivation and repetition. In *Proc. Conference on Advances in Consumer Research*, Provo, UT, 1995.

[26] R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Trans. on Neural Networks*, 9(2) :257–265, 1998.

[27] K. Smith, S. Ba, D. Gatica-Perez, and J.M. Odobez. Tracking the multi-person wandering visual focus of attention. In *Intl. Conference on Multimodal Interfaces (ICMI)*, Banff, Canada, Nov. 2006.

[28] K. Smith, D. Gatica-Perez, S. Ba, and J.M. Odobez. Evaluating multi-object tracking. In *CVPR Workshop on Empirical Evaluation Methods in Computer Vision*, San Diego, June 2005.

[29] K. Smith, D. Gatica-Perez, and J.M. Odobez. Using particles to track varying numbers of objects. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, San Diego, June 2005.

[30] P. Smith, M. Shah, and N. da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE Trans. on Intelligent Transporation Systems*, 4(4) :205–218, 2004.

[31] S. Srinivasan and K. Boyer. Head-pose estimation using view based eigenspaces. In *Proc. of Intl. Conference on Pattern Recognition (ICPR)*, Quebec, Aug. 2002.

[32] V. Starr and C.A. Lowe. The influence of program context and order of ad presentation on immediate and delayed responses to television ads. *Advances in Consumer Research*, 22 :184–190, 1995.

[33] C. Stauffer and E. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Ft. Collins, CO, June 1999.

[34] R. Stiefelhagen. Estimating head-pose with neural networks-results on the pointing04 icpr workshop evaluation data. In *Proc. of Pointing 04 ICPR Workshop*, Cambridge, Aug. 2004.

[35] R. Stiefelhagen, M. Finke, and A. Waibel. A model-based gaze tracking system. In *Proc. of IEEE Intl. Joint Symposia on Intelligence and Systems*, 1996.

[36] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *Visual Information and Information Systems*, pages 761–768, 1999.

[37] H. Tao, H. Sawhney, and R. Kumar. A sampling algorithm for detection and tracking multiple objects. In *ICCV Workshop on Vision Algorithms*, Kerkyra, Sept. 1999.

[38] W. Thoretz. Press release : Nielsen to test electronic ratings service for outdoor advertising, 2002.

[39] E.M. Tucker. The power of posters. Technical report, University of Texas at Austin, 1999.

[40] Y. Wu and K. Toyama. Wide range illumination insensitive head orientation estimation. In *Automatic Face and Gesture Recognition (AFGR)*, Grenoble France, Apr. 2001.

[41] J. Yang, W. Lu, and A. Weibel. Skin color modeling and adaptation. In *Asian Conference on Computer Vision*, Oct 1998.

[42] R. Yang and Z. Zhang. Model-based head-pose tracking with stereo vision. Technical Report MSR-TR-2001-102, Microsoft Research, 2001.

[43] T. Yu and Y. Wu. Collaborative tracking of multiple targets. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Washington DC, June 2004.

[44] L. Zhao, G. Pingali, and I. Carlbom. Real-time head orientation estimation using neural networks. In *Proc. of the Intl. Conference on Image Processing (ICIP)*, Rochester, NY, Sep. 2002.

[45] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Washington DC, June 2004.