Tracking the Visual Focus of Attention for a Varying Number of Wandering People

Kevin Smith, *Member*, *IEEE*, Sileye O. Ba, Jean-Marc Odobez, *Member*, *IEEE*, and Daniel Gatica-Perez, *Member*, *IEEE*

Abstract—In this paper, we define and address the problem of finding the *visual focus of attention for a varying number of wandering people* (VFOA-W), determining where a person is looking when their movement is unconstrained. The VFOA-W estimation is a new and important problem with implications in behavior understanding and cognitive science and real-world applications. One such application, presented in this paper, monitors the attention passers-by pay to an outdoor advertisement by using a single video camera. In our approach to the VFOA-W problem, we propose a multiperson tracking solution based on a dynamic Bayesian network that simultaneously infers the number of people in a scene, their body locations, their head locations, and their head pose. For efficient inference in the resulting variable-dimensional state-space, we propose a Reversible-Jump Markov Chain Monte Carlo (RJMCMC) sampling scheme and a novel global observation model, which determines the number of people in the scene and their locations. To determine if a person is looking at the advertisement or not, we propose Gaussian Mixture Model (GMM)-based and Hidden Markov Model (HMM)-based VFOA-W models, which use head pose and location information. Our models are evaluated for tracking performance and ability to recognize people looking at an outdoor advertisement, with results indicating good performance on sequences where up to three mobile observers pass in front of an advertisement.

Index Terms—Computer vision, tracking, video analysis, consumer products.

1 Introduction

As our motivation for this work, we consider the following hypothetical question: "*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 empirically judge the best placement and content of these advertisements?*" Currently, the advertising industry relies on recall surveys or traffic studies to measure the effectiveness of outdoor advertisements. However, these hand-tabulated approaches are often impractical or too expensive to be commercially viable and yield small samples of data. A tool that automatically measures the effectiveness of printed outdoor advertisements would be extremely valuable but does not currently exist.

However, in the television industry, such a tool exists. The Nielsen ratings measure media effectiveness by estimating the size of the net cumulative audience of a program via surveys and Nielsen Boxes. If one were to design a similar system for outdoor advertisements, it might *automatically determine the number of people who have actually viewed an advertisement as a percentage of the total number of people exposed to it*. This is an example of an important extension of the visual focus of attention (VFOA) problem, in which there exists a varying number of wandering people. We denote this as the *VFOA-W problem*, whose tasks are

1. to automatically detect and track a *varying* number of mobile observers and
2. to estimate their VFOA with respect to one or more fixed targets.

Solutions to the VFOA-W problem have implications for other fields (for example, human behavior and human-computer interaction (HCI)) and real-life applications. In our example of the outdoor advertisement application, the goal is to identify each person exposed to the advertisement and determine if and when they looked at it. We can also collect other useful statistics such as the amount of time that they spent looking at the advertisement.

The VFOA-W problem represents an extension of traditional VFOA problems studied in computer vision (for example, [38]) in two respects. First, for VFOA-W, the VFOA must be estimated for an unknown varying number of subjects instead of a fixed number of static subjects. Second, in VFOA-W, mobility is unconstrained. By unconstrained motion, we mean that the subjects are free to walk about the scene (or wander): They are not forced to remain seated or otherwise restrained. This complicates the task, as the subject's appearance will change as he moves about the scene and keeps his attention focused on the target.

Camera placement and the unconstrained motion of the subjects can limit the video resolution of the subjects, making the VFOA estimation from eye gaze difficult, as illustrated in Fig. 1. To address this problem, we follow the work of Stiefelhagen et al., who showed that VFOA can be deduced from head pose when the resolution is insufficient to determine eye gaze [38].

In this paper, we propose a principled probabilistic framework for estimating VFOA-W and apply our method

- K. Smith is with École Polytechnique Fédérale de Lausanne (EPFL) IC ISIM CVLAB, BC 304 (Bâtiment BC), Station 14CH-1015 Lausanne, Switzerland. E-mail: kevin.smith@epfl.ch.
- S.O. Ba, J.-M. Odobez, and D. Gatica-Perez are with IDIAP Research Institute, Centre du Parc, Av. des Prés-Beudin 20, Case Postale 592, CH-1920 Martigny, Switzerland. E-mail: sba@idiap.ch.

Manuscript received 11 Aug. 2006; revised 5 Mar. 2007; accepted 17 July 2007; published online 5 Sept. 2007.
Recommended for acceptance by T. Darrell.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0596-0806.
Digital Object Identifier no. 10.1109/TPAMI.2007.70773.

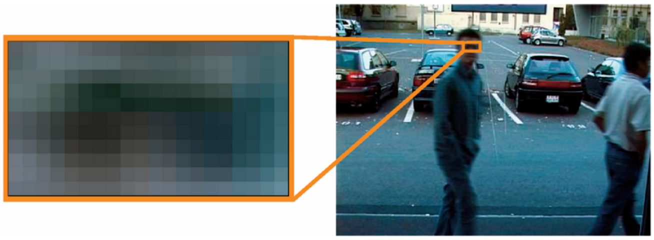


Fig. 1. **Determining VFOA from eye gaze**. In the VFOA-W problem, allowing an unknown number of people to move about the scene (as well as enter and exit the scene) complicates the task of estimating each subject's *visual focus of attention* (VFOA). Because a large field of view is necessary, the resolution is often too low to estimate the VFOA by using eye gaze (as seen above). In our work, VFOA is inferred from a person's location and head pose.

to the advertising example to demonstrate its usefulness in a real-life application. Our method consists of two components: a dynamic Bayesian network, which simultaneously tracks people in the scene and estimates their head pose, and two VFOA-W models based on Gaussian mixture models (GMMs) and hidden Markov models (HMMs), which infer a subject's VFOA from their location and head pose. We assume a fixed uncalibrated camera, which can be placed arbitrarily, with the condition that subjects appear vertical, with their face in view of the camera when they look at the target, as in Fig. 1.

Besides defining the VFOA-W problem itself, which, to our knowledge, is a previously unaddressed problem in the literature, we also make several contributions toward a solution. First, we propose a probabilistic framework for solving the VFOA-W 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 multiperson fashion, consisting of size and location parameters for the head and body, and head pose parameters for each person in the scene. This type of framework facilitates defining interactions between people.

Second, because the dimension of the state representing a single person is sizeable, the multiobject state-space can grow to be quite large when several people appear together in the scene. The dimension of the state-space also changes as people enter or leave the scene. Efficiently inferring a solution in a large variable-dimensional space is a challenging problem. To address this issue, we designed a Reversible-Jump Markov Chain Monte Carlo (RJMCMC) sampling method to do inference in this large variable-dimensional space.

Third, in order to localize, identify, and determine the correct number of people present, we propose a novel global observation model. This model uses color and binary measurements taken from a background subtraction model and allows for the direct comparison of observations containing different numbers of objects.

Finally, we demonstrate the applicability of our model by applying it to the outdoor advertisement problem. We show that we are able to gather useful statistics such as the number of people who looked at the advertisement and the total number of people exposed to it on a set of video sequences in which people walk past a simulated advertisement. We provide an evaluation of our approach on this data by using a comprehensive set of objective performance measures.

The remainder of this paper is organized as follows: In Section 2, we discuss related works. In Section 3, we describe our joint multiperson head-pose tracking model. In Section 4, we propose the GMM and HMM methods for modeling VFOA-W. In Section 5, we describe our parameter setting procedure. In Section 6, we evaluate our models on captured video sequences of people passing by an outdoor advertisement. Some limitations of our approach are discussed in Section 7. Finally, Section 8 contains some concluding remarks.

2 Related Work

To our knowledge, our work is the first attempt to estimate the VFOA-W. However, there is an abundance of literature concerning the three component tasks of the VFOA-W problem: multiperson tracking, head-pose tracking, and VFOA estimation.

2.1 Multiperson Tracking

Multiperson tracking is the process of locating a variable number of moving people or objects in a video over time. Multiperson tracking is a well-studied topic with a variety of different approaches. We restrict our discussion to probabilistic tracking methods, which use a particle filter (PF) formulation [20], [39], [15], [23]. Some computationally inexpensive methods use a single-object state-space model [23] 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 [20], [39], [15], [17], [44], [32]. However, sampling from a joint state-space can quickly become inefficient, as the dimension of the space increases when more people are added [20]. Recent work has concentrated on using MCMC sampling to track multiple people more efficiently [17], [44]. In a previous work [32], we proposed to generalize this model to handle a varying number of people using RJMCMC, which allows for a formal definition of object appearance (births) and disappearances (deaths) from the scene through the definition of a set of reversible move types (see Section 3.4). In this work, we extend the model in [32] to handle a more complex object model and a larger state-space, necessitating the design of new move types and proposal distributions, a new observation model, and interperson and intraperson interactions.

2.2 Head-Pose Tracking

Head-pose tracking is the process of locating a person's head and estimating its orientation in space. Existing methods can be categorized in two ways: feature-based versus appearance-based approaches and parallel versus 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 for estimating the head pose [10], [13], [37]. A feature-based approach employing stereo vision was proposed in [42]. The major drawback of the feature-based approach is that it requires high-resolution head images, which is impractical in many situations. Occlusions and other ambiguities present difficult challenges to this approach as well.

In the appearance-based approach, instead of concentrating on specific facial features, 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 have proposed using neural networks [28], [19], principal component analysis [8], and multidimensional Gaussian distributions [41] as modeling tools.

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 some tracking algorithm, and features are extracted from the tracking results to perform pose estimation. This methodology has been used by several authors [37], [28], [19], [43], [41], [7]. 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 for improving 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 approach's improved accuracy through joint tracking and pose estimation. Benefits of this method can be seen in [42] and [3]. In this work, we adopt an appearance-based parallel approach to head-pose tracking, where we jointly track the bodies and the heads and estimate the poses of the heads of multiple people within a single framework.

2.3 Visual Focus of Attention

Estimating VFOA is of interest to several domains, as a person's VFOA is often strongly correlated with his behavior or activity. Strictly speaking, a person's VFOA is determined by his eye gaze. However, measuring the VFOA by using eye gaze is often difficult or impossible, as it can require either the movement of the subject to be constrained or high-resolution images of the eyes, which may not be practical [34], [22].

In [38], Stiefelhagen et al. made the important observation that VFOA can be reasonably derived by head pose in many cases. We rely on this assumption to simultaneously estimate the VFOA for multiple people without restricting their motion. Others have followed this work, such as Danninger et al. [9] (where VFOA is estimated using head pose in an office setting), Stiefelhagen [36] (where VFOA for multiple people and multiple targets is estimated through head pose), and Katzenmaier et al. [16] (where the head pose is used for determining the addressee in a human-human-robot interaction). Note that in these related works, the VFOA is modeled for a fixed number of seated people by using an unsupervised learning process.

2.4 Other Related Work

Although we believe that this work is the first attempt to estimate VFOA-W, there exist several previous works in a similar vein. The Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2002) defined a number of estimation tasks on videos 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 authors attempted to accomplish these tasks through various means, including [21], [25]. However, among these works, there were no attempts to use head pose or eye gaze to detect when people were looking at the window: All estimations were done using only body location, assuming that a person pausing in front of the window is looking at it. A preliminary version of this paper appeared in [30].

3 Joint Multiperson and Head-Pose Tracking

In a Bayesian approach to multiperson tracking, the goal is to estimate the posterior distribution for a target state $\mathbf{X}_t$, taking into account a sequence of observations $\mathbf{Z}_{1:t} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_t)$, $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$. The state or joint multiperson configuration 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

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

where the dynamic model $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ governs the temporal evolution of $\mathbf{X}_t$, 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 estimation of the state $\mathbf{X}_t$. Here, C is a normalization constant.

In practice, the estimation of the filtering distribution in (1) is often intractable. However, it can be approximated by applying the Monte Carlo method, where the target distribution (1) is represented by a set of N samples $\{\mathbf{X}_t^{(n)}, n = 1, \ldots, N\}$, where $\mathbf{X}_t^{(n)}$ denotes the $n\text{th}$ sample. In this work, we use RJMCMC, where a set of uniformly weighted samples form a so-called Markov chain. Given the sample set approximation of the posterior at time $t-1$, $p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) \approx \sum_n \delta(\mathbf{X}_{t-1} - \mathbf{X}_{t-1}^{(n)})$, the Monte Carlo approximation of (1) is written as follows:

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

In the following sections, we describe the joint multiperson and head tracking model, the dynamic model, the observation model, and how RJMCMC sampling is used for inference.

3.1 State-Space Definition for a Varying Number of People

The state at time t describes the joint configuration of people in the scene. Because the amount of people in the scene may vary, we define a state model designed to accommodate changes in dimension [32]. The joint state vector $\mathbf{X}_t$ is defined by $\mathbf{X}_t = \{\mathbf{X}_{i,t} | i \in \mathcal{I}_t\}$, where $\mathbf{X}_{i,t}$ is the state vector for person i, and $\mathcal{I}_t$ is the set of all person indices at time t. The total number of people that are 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 $\mathbf{X}_t = \emptyset$ (the empty set).

Each person is represented by two components: body $\mathbf{X}_{i,t}^b$ and head $\mathbf{X}_{i,t}^h$, $\mathbf{X}_{i,t} = (\mathbf{X}_{i,t}^b, \mathbf{X}_{i,t}^h)$, as seen in Fig. 2. The body component is represented by a bounding box, whose state vector contains four parameters $\mathbf{X}^b = (x^b, y^b, s^b, e^b)$ (we drop the i, t subindices to simplify notation). The point (x^b, y^b) is the continuous 2D location of the center of the bounding box, s^b is

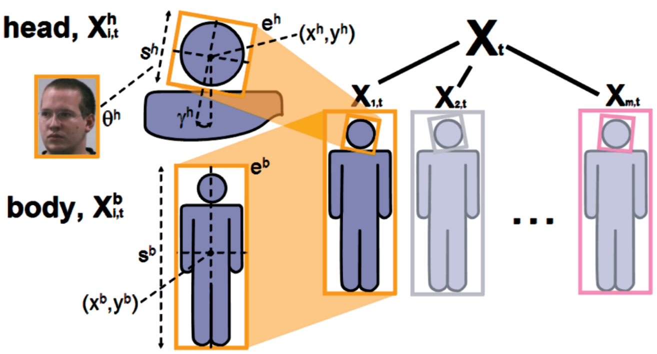


Fig. 2. **State model for varying numbers of people and their head-pose**. The joint multiperson state $\mathbf{X}_t$ consists of an arbitrary number of single-person states $\mathbf{X}_{i,t}$, each of which contains a body $\mathbf{X}^b_{i,t}$ and head $\mathbf{X}^h_{i,t}$ components. The body is modeled 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.

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 to its height.

The head component is represented by a bounding box, which may rotate in the image plane, along with an associated discrete exemplar used for representing the head-pose (see Fig. 4). 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. The discrete variable θ^h represents the head-pose exemplar, which models the out-of-plane head rotation. Note that the head pose is completely defined by the couple (γ^h, θ^h).

3.2 Dynamic and Interaction Model

The dynamic model governs the evolution of the state between time steps. It is responsible for predicting the motion of people (and their heads) and governing transitions between the head-pose exemplars. It is also responsible for modeling *interperson* interactions between the various people and *intraperson* interactions between the body and the head. We define the dynamic model for a variable number of objects as

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

where $p_V(\mathbf{X}_t|\mathbf{X}_{t-1})$ is the multiobject transition model, and $p_0(\mathbf{X}_t)$ is an interaction term. The multiperson transition model is defined more specifically as

$$p_V(\mathbf{X}_t|\mathbf{X}_{t-1}) = \begin{cases} \prod_{i \in \mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}) & \text{if} \quad \mathcal{I}_t \neq \emptyset, \\ k & \text{if} \quad \mathcal{I}_t = \emptyset, \end{cases} \tag{4}$$

where k is a constant. The single-person transition model is given by

$$p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}) = \begin{cases} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) & \text{if} \quad i \text{ previously existed, } i \in \mathcal{I}_{1:t-1}, \\ p(\mathbf{X}_{i,t}) & \text{if} \quad i \text{ is a previously unused index, } i \notin \mathcal{I}_{1:t-1}, \end{cases} \tag{5}$$

where $p(\mathbf{X}_{i,t})$ is a mixture that selects parameters from either a previously dead tracked object or a new proposal (see Section 3.5 for birth move). The first term $p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1})$ is given by

$$p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) = p\left(\mathbf{X}^b_{i,t}|\mathbf{X}^b_{i,t-1}\right) p\left(L^h_{i,t}|L^h_{i,t-1}\right) p\left(\theta^h_{i,t}|\theta^h_{i,t-1}\right), \tag{6}$$

where the dynamics of the body state $\mathbf{X}^b_i$ and the head spatial state component L^h_i are modeled as second-order autoregressive (AR) processes. This model applies for dead and live objects, as it is necessary for the positions of dead objects to be propagated for a certain duration in order to allow them to possibly be reborn. The head-pose exemplars θ^h_i are modeled by a discrete first-order AR process represented by a transition probability table.

The interaction model $p_0(\mathbf{X}_t)$ handles two types of interactions: interperson p_{0_1} and intraperson p_{0_2} : $p_0(\mathbf{X}_t) = p_{0_1}(\mathbf{X}_t) p_{0_2}(\mathbf{X}_t)$. For modeling interperson interactions, we follow the method proposed in [17], in which the interperson interaction model $p_{0_1}(\mathbf{X}_t)$ serves the purpose of restraining multiple trackers from fitting the same person by penalizing overlap. It accomplishes this by exploiting a pairwise Markov Random Field (MRF) whose graph nodes are defined by the people present at each time step. The links in the graph are defined 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 multiperson dynamic model based on the locations of a person's neighbors. This constraint is defined by a nonnegative penalty function $g = \frac{2\rho(X_i,X_j)\nu(X_i,Xj)}{\rho(X_i,X_j)+\nu(X_i,Xj)}$, which penalizes configurations that contain overlapping pairs of people, where S^{X_i} is the spatial support of $X_{i,t}$, $\rho(X_i, X_j) = \frac{S^{X_i} \cap S^{X_j}}{S^{X_i}}$ is the recall, and $\nu(X_i, Xj) = \frac{S^{X_i} \cap S^{X_j}}{S^{X_j}}$ is the precision, so that $g = 0$ for no overlap, and increased overlap increases the penalization function g.

We also introduce intraperson interactions to the overall motion model. The intraperson interaction model is meant to constrain the head model with respect to the body model so that they are configured in a physically plausible way (for example, the head is not detached from the body). The intraperson interaction model $p_{0_2}(\mathbf{X}_t)$ is defined as $p_{0_2}(\mathbf{X}_t) = \prod_{k \in \mathcal{I}_t} p(L^h_{k,t}|\mathbf{X}^b_{k,t})$, where $p(L^h_{k,t}|\mathbf{X}^b_{k,t}) \propto exp(-\lambda\, d^2 (L^h_{k,t}, \mathbf{X}^b_{k,t}))$, and the distance function $d(\cdot)$ is equal to zero when the head center is within a predefined region relative to the body (that is, the area defined by the top third of the body bounding box) and is equal to the euclidean distance between the head and nearest edge of the predefined region otherwise. This term penalizes head configurations that fall outside an acceptable range of the body, increasing as the distance between the head and body increases. With these terms defined, the Monte Carlo approximation in (2) can now be expressed as

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \approx C^{-1} p(\mathbf{Z}_t|\mathbf{X}_t) p_0(\mathbf{X}_t) \sum_n p_V\left(\mathbf{X}_t|\mathbf{X}^{(n)}_{t-1}\right), \tag{7}$$

$$= C^{-1} p(\mathbf{Z}_t|\mathbf{X}_t) \prod_{ij \in \mathcal{C}} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \prod_{k \in \mathcal{I}_t} p\left(L^h_{k,t}|\mathbf{X}^b_{k,t}\right) \sum_n p_V\left(\mathbf{X}_t|\mathbf{X}^{(n)}_{t-1}\right). \tag{8}$$

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 features. The body model consists of *binary* and *color* features, which are global in that they are defined pixelwise over the entire image. The binary features $\mathbf{Z}_t^{bin}$ make use of a foreground segmented image, whereas the color features $\mathbf{Z}_t^{col}$ exploit histograms in hue-saturation (HS) space. The head model is local in that its features $\mathbf{Z}^h$ are gathered independently for each person from an area around the head. They are responsible for the localization of the head and estimation of the head pose, and they 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 the notation. Assuming conditional independence of body and head observations, the overall likelihood is given by

$$p(\mathbf{Z}|\mathbf{X}) \triangleq p(\mathbf{Z}^{col}|\mathbf{Z}^{bin}, \mathbf{X})p(\mathbf{Z}^{bin}|\mathbf{X})p(\mathbf{Z}^{h}|\mathbf{X}). \quad (9)$$

The first two terms constitute the body model, and the third term represents the head model.

3.3.1 Body Model

An issue arises when defining an observation likelihood for a variable number of objects. Fairly comparing the likelihoods, which is a task essential to the filtering process, is more complicated when the number of objects may vary. For a fixed number of objects, the comparison of two observation likelihoods can be relatively straightforward. Given an observation likelihood for a single object, the joint multiobject observation likelihood can be defined as the product of the individual object likelihoods [17], [18], [44]. For a static number of objects, the observation likelihoods are directly comparable because the number of objects, and thus the number of factors in the likelihood, is fixed. Fairly comparing two likelihoods defined in this manner when the number of objects may vary is problematic, as the number of factors in the likelihood terms that we wish to compare may be different. This can eventually lead to observation likelihoods of different magnitude orders reflecting a variation in the number of factors rather than an actual difference in the overall likelihood level.

To address this issue, we propose a global body observation model, which allows for a direct comparison of observations containing different numbers of objects. Our model detects, tracks, and maintains consistent identities of people, adding and removing them from the scene when necessary. It is comprised of a binary feature and a color feature.

Body Binary Feature. We introduced the binary feature in a previous work [32], which relies on an adaptive foreground segmentation technique described in [35]. 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 multiperson configuration and foreground segmentation, the binary feature computes the distance between the observed overlap (between the spatial support of the multiperson 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 percent foreground and 35 percent background." The overlap is measured for F and B in terms of precision and recall: $\nu^F = \frac{S^{\mathbf{X}} \cap F}{S^{\mathbf{X}}}$, $\rho^F = \frac{S^{\mathbf{X}} \cap F}{F}$, $\nu^B = \frac{S^{\mathbf{X}} \cap B}{S^{\mathbf{X}}}$, 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 multiperson 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 nonzero person counts $m \neq 0$ is a single Gaussian distribution in precision-recall space (ν^F, ρ^F). The binary background term $p(\mathbf{Z}^{bin,B}|\mathbf{X})$, on the other hand, is defined as a set of GMMs learned for each possible person count $m \in \mathcal{M}$. For example, if the multiperson state hypothesizes that two people are present in the scene, the binary background likelihood term is the GMM density of the observed ν^B and ρ^B values learned for $m = 2$. For details on the learning procedure, see Section 5.

In Fig. 3, an example of the binary observation model trained to recognize $\mathcal{M} = \{1, 2, 3\}$ objects is shown. Learning of the GMM parameters was done using the Expectation-Maximization (EM) algorithm on 948 labeled images from the data set described in Section 5.2. As shown in Figs. 3a, 3b, and 3c, two *ground truth* people appear in the scene. The binary feature also encourages the tracker to propose hypotheses with good spatial fitting in a similar manner. For example, a poorly placed object might only cover a small fraction of the foreground blob corresponding to a person appearing in the image. In this case, the foreground ν and ρ measurements will not match the learned values well, as the learning has been done using tightly fitting example data.

Body Color Feature. The color feature is responsible for maintaining the identities of people over time and 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 color foreground likelihood compares an adaptive four-dimensional (4D) spatial-color model histogram HC with a 4D spatial-color observed histogram $H(X_t)$. The observation likelihood measures the similarity of the 4D histograms by $p(\mathbf{Z}^{col,F}|\mathbf{Z}^{bin,F}, \mathbf{X}) \propto \exp(-\lambda_F d_F^2(HC, H(X_t)))$, where $d_F(HC, H(X_t))$ is the Bhattacharya distance [6] between the histograms. The 4D histograms $H(i, bp, h, s)$ are collected as follows. The first dimension corresponds to the object i, and the remaining dimensions correspond to an object color model proposed by Pérez et al. [26]. In this object color model, the histogram is defined over three body parts bp corresponding to the head, torso, and legs. For each body part region, a 2D $\mathrm{HS} + \mathrm{V}$ histogram is computed using the HS-Value elements from the corresponding location in the training image. The $\mathrm{HS} + \mathrm{V}$ histogram is constructed by

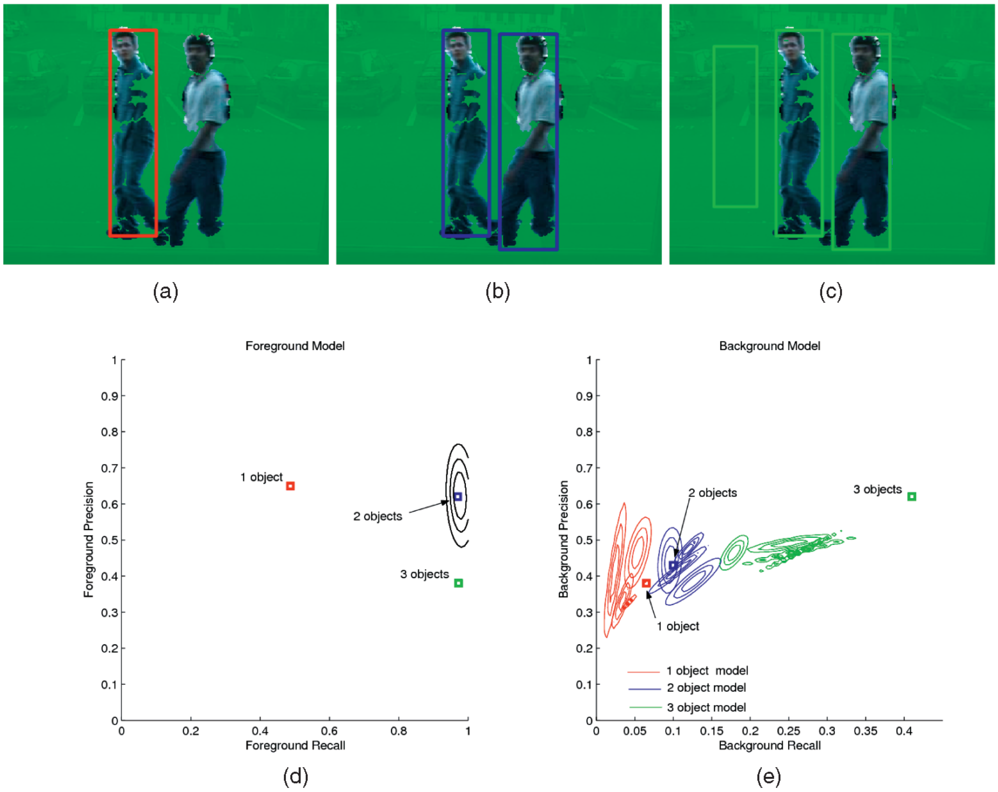


Fig. 3. **The binary observation model determines the number of objects and localizes the objects**. In (a), (b), and (c), two *ground truth* people appear in the scene. The binary foreground model consists of $K_{bf} = 1$ Gaussian, the black contour in (d). The background model consists of three GMMs of $K_{bb} = 4$ mixture components, each in (e) ($m = 1$: red contour, $m = 2$: blue contour, and $m = 3$: green contour). The square data points in (d) and (e) represent measured precision/recall observations from the hypotheses in (a), (b), and (c). The red square indicates the (ν, ρ) values for the hypothesis containing only one object in (a), the blue square indicates the two-object hypothesis in (b), and the green square indicates the three-object hypothesis in (c). Clearly, the two-object hypothesis, which agrees with the ground truth, fits the model better than the others. The binary observation model will associate the highest likelihood to the hypothesis matching the actual number of objects $(m = 2)$ (a) One object (shown in red). (b) Two objects in (shown in blue). (c) Three objects in (shown in green). (d) Binary foreground model. (e) Binary background model.

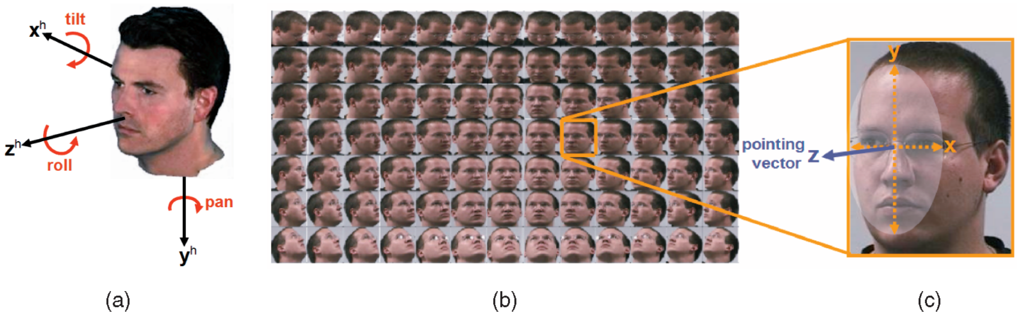


Fig. 4. **The head-pose model**. (a) The head pose represented by the angles resulting from the Euler decomposition of the head rotation with respect to the head frame, known as pan, tilt, and roll. (b) Set of discrete poses θ^h used for representing out-of-plane rotation exemplars from the Prima-Pointing database. (c) Pointing vector z^h (note z^h only depends on the pan and tilt angles when using the representation on (a)).

populating an $B_H \times B_S$ HS histogram (where $B_H = 8$ and $B_S = 8$ are the number of H and S bins) using only the pixels with H and S greater than 0.15. The $+$V portion of the HS $+$ V histogram contains a $B_V \times 1$ $(B_V = 8)$ Value histogram comprised of the pixels with HS lower than or equal to 0.15.[1]

The 4D adaptive color model HC is selected from a set of competing adaptive color models every frame. When an object first appears, pixel values extracted from the initial frame are used for initializing each competing color model. At the end of each subsequent frame, the point estimate solution for the objects' locations is used for extracting a 4D multi-person color histogram, which is compared to each model. The nearest matching competing model receives a vote and is updated with the extracted data by a running mean. When computing the foreground color likelihood in the following frame, the model with the most votes is used.

The background color likelihood helps reject configurations containing untracked people by penalizing unexpected colors. The background model is a static 2D HS color histogram, learned from empty training images. The background color likelihood is defined as $p(\mathbf{Z}_t^{col,B} | \mathbf{Z}_t^{bin,B}, \mathbf{X}_t) \propto$

1. This extra one-dimensional (1D) V histogram is appended as one extra row in the HS histogram, resulting in a "2D" HS $+$ V histogram.

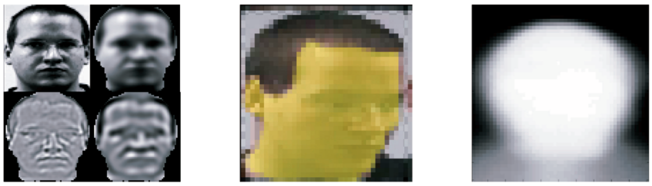


Fig. 5. **Head-pose observation features**. (a) Texture is used for estimating the head pose by applying three filters to (upper left) the original image. These filters include (upper right) a coarse-scale Gaussian filter, (lower left) a fine scale Gabor filter, and (lower right) a coarse scale Gabor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

$e^{\lambda_B d_B^2}$, where λ_B and d_B^2 are defined as in the foreground case but using the background images to compute the histogram.

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

$$p(\mathbf{Z}^h|\mathbf{X}) = \left[\prod_{i\in\mathcal{I}} p(\mathbf{Z}_i^{tex}|\mathbf{X}_i)p(\mathbf{Z}_i^{sk}|\mathbf{X}_i)p(\mathbf{Z}_i^{sil}|\mathbf{X}_i)\right]^{\frac{1}{m}}. \qquad (10)$$

The overall head likelihood is composed of the geometric mean of the individual head likelihood terms. The geometric mean provides a pragmatic solution to the problem of comparing likelihoods with a variable number of factors (corresponding to varying numbers of people). However, note that it is not justifiable in a probabilistic sense.

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 by using foreground segmentation. The texture and skin color features, which have appeared in previous works, including our own [3], [41], use appearance-dependent observations to determine the head pose of the subject.

Head-Pose Texture Feature. The head-pose texture feature reports how well the texture of an extracted image patch matches the texture of the discrete head-pose hypothesized by the tracker. Texture is represented using responses from three filters: A coarse-scale Gaussian filter, a fine Gabor filter, and a coarse Gabor filter, as seen in Fig. 5.

Texture models were learned for each discrete head pose θ^h. Training was done using several 64×64 images for each head pose taken from the Prima Pointing Database. Histogram equalization was applied to the training images to reduce variation in lighting, the filters were applied on a subsampled grid to reduce computation, and the filter responses 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 $\sigma_\theta = (\sigma_j^\theta)$, $j = 1, \ldots, N_{tex}$, of the corresponding training feature vectors were computed and used for defining the person texture likelihood model from (10) as

$$p(\mathbf{Z}_i^{tex}|\mathbf{X}_i) = \frac{1}{Z_\theta}\exp -\lambda_\theta^{tex} d_\theta\left(\mathbf{Z}_i^{tex}, e^{\theta_i}\right), \qquad (11)$$

where θ_i is the head pose associated with person i, and d_θ is the normalized truncated Mahalanobis distance defined as

$$d_\theta(u,v) = \frac{1}{N_{tex}}\sum_{j=1}^{N_{tex}} \max\left(\left(\frac{u_j - v_j}{\sigma_j^\theta}\right)^2,\ T_{tex}^2\right), \qquad (12)$$

where $T_{tex} = 3$ is a threshold set to make the distance more robust to outlier components. The normalization constant Z_θ and the parameter λ_θ^{tex} are learned from the training data using a procedure proposed in [40].

Head-Pose Skin Feature. The texture feature is a powerful tool for modeling the head pose but is 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 nonskin pixel (0). An example of a skin color mask can be seen in Fig. 5. The skin color binary models were learned from skin color masks extracted from the same training images used in the texture model by using a Gaussian skin color distribution modeled in normalized red-green (RG) space [2].

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 by using a temporally adaptive person-dependent skin color distribution model, which is updated with a maximum a posteriori (MAP) adaptation to the current person by using skin color pixels in the estimated head location). The skin color likelihood of a measurement $\mathbf{Z}_i^{sk}$ belonging to the head of person i is defined as

$$p(\mathbf{Z}_i^{sk}|\mathbf{X}_i) \propto \exp -\lambda_{sk}\|\mathbf{Z}_i^{sk} - M^{\theta_i}\|_1, \qquad (13)$$

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

Head-Pose Silhouette Feature. In addition to the pose-dependent head model, we propose to use a head silhouette likelihood model to aid in localizing the head by taking advantage of foreground segmentation information. A head silhouette model H^{sil} (see Fig. 5) is constructed by averaging head silhouette patches extracted from binary foreground segmentation images resized to 64×64 (see Section 5.2; 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 silhouette likelihood term is defined as

$$p(\mathbf{Z}_i^{sil}|\mathbf{X}_i) \propto \exp -\lambda_{sil}\|\mathbf{Z}_i^{sil} - H^{sil}\|_1, \qquad (14)$$

where λ_{sil} is an parameter tuned on training data.

In practice, we found that introducing this term (not defined in our previous work [3] or in others, like [41]) greatly improved the head localization in the combined body-head optimization process. Further details on the head-pose model can be found in [2].

3.4 Componentwise Reversible-Jump MCMC

Having defined the components of (2) (state-space, dynamic model, and observation model), we now define an RJMCMC sampling scheme to efficiently generate a Markov Chain representing the posterior distribution in (8).

As the state vector for a single person is 10-dimensional, the multiperson state-space can quickly become very large when allowing for an arbitrary number of people. Traditional Sequential Importance Resampling (SIR) PFs are known to be inefficient in such high-dimensional spaces [1]. The classic Metropolis-Hastings (MH)-based MCMC PF is more efficient [17] but does not allow for the dimensionality of the state-space to vary (the number of people must remain static). To solve this problem, we have defined a type of the RJMCMC sampling scheme [11] based on a method that we proposed previously [32], which includes a set of reversible move types (or *jumps*), which can change the dimension of the state-space (note that a different RJMCMC model was originally used for tracking in [44]).

The RJMCMC algorithm starts in an arbitrary configuration $\mathbf{X}^0$ sampled from the Markov chain belonging to the previous time step $t-1$. The first step is to select a *move type* v from the set of *reversible moves* Υ by sampling from a prior distribution on the move types $v \sim p(v)$. The next step is to choose a target object i^* (or two objects i^* and k^* in the case of a *swap* move) and apply the selected move type to form a proposal configuration $\mathbf{X}^*$. The proposal is evaluated in an *acceptance test*, and based on this test, either the previous state $\mathbf{X}^{(n-1)}$ or the proposed state $\mathbf{X}^*$ is accepted and added to the Markov chain for time t.

A reversible move defines a transition from the current state $\mathbf{X}$ and a proposed state $\mathbf{X}^*$ via a deterministic function h_v, and, when necessary, a generated random auxiliary variable $\mathbf{U}$ [11]. This transition can involve changing the dimension between $\mathbf{X}$ and $\mathbf{X}^*$. The transition function h_v is a *diffeomorphism* or an invertible function that maps one space to another. There is flexibility in defining the transition h_v, so long as it meets the following criteria: 1) it is a *bijection*, that is, if h_v defines a one-to-one correspondence between sets, 2) its derivative is invertible, that is, it has a nonzero Jacobian determinant, and 3) it has a corresponding *reverse move* h_v^R, which can be applied to recover the original state of the system. The reverse move must also meet the first two criteria. For move types that do not involve a dimension change, the reverse move is often the move type itself, in which case it is possible to recover the original multiobject configuration by reapplying the same move. Move types that involve a change in dimension usually cannot revert to the previous state and are defined in *reversible move pairs*, where one move is the reverse of the other.

Following [1], the general expression for the acceptance ratio for a transition defined by h_v from the current state $\mathbf{X}_t$ to a proposed state $\mathbf{X}_t^*$ (allowing for jumps in dimension) is given as follows:

$$\alpha(\mathbf{X}_t, \mathbf{X}_t^*) = \min \left\{1, \frac{p(\mathbf{X}_t^*|\mathbf{Z}_{1:t})}{p(\mathbf{X}_t|\mathbf{Z}_{1:t})} \times \frac{p(v^R)}{p(v)} \times \frac{q_v^R(\mathbf{X}_t, \mathbf{U}|\mathbf{X}_t^*, \mathbf{U}^*)}{q_v(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U})} \times \left|\frac{\partial h_v(\mathbf{X}_t, \mathbf{U})}{\partial(\mathbf{X}_t, \mathbf{U})}\right|\right\}, \tag{15}$$

where $\mathbf{U}$ is an auxiliary dimension-matching variable, and $\mathbf{U}^*$ is its reverse move counterpart. $p(\mathbf{X}_t^*|\mathbf{Z}_{1:t})$ is the target distribution evaluated at the proposed configuration $\mathbf{X}_t^*$, $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$ is the target distribution evaluated at the current configuration $\mathbf{X}_t$, $p(v)$ is the probability of choosing move type v, $p(v^R)$ is the probability of choosing the reverse move type v^R, $q_v(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U})$ is the proposal for a move from $(\mathbf{X}_t, \mathbf{U}) \rightarrow (\mathbf{X}_t^*, \mathbf{U}^*)$, $q_v^R(\mathbf{X}_t, \mathbf{U}|\mathbf{X}_t^*, \mathbf{U}^*)$ is the proposal distribution for the reverse move from $(\mathbf{X}_t^*, \mathbf{U}^*) \rightarrow (\mathbf{X}_t, \mathbf{U})$, and $\frac{\partial h_v(\mathbf{X}_t,\mathbf{U})}{\partial(\mathbf{X}_t,\mathbf{U})}$ is the Jacobian determinant of the diffeomorphism from $(\mathbf{X}_t, \mathbf{U}) \rightarrow (\mathbf{X}_t^*, \mathbf{U}^*)$. The Jacobian determinant is the matrix of all first-order partial derivatives of a vector-valued function, which reduces to one for our selected moves (see [29] for further details).

Instead of updating the whole of an object $\mathbf{X}_i$ in a single move, as in [44] and [32], we propose to split $\mathbf{X}_i$ into components of differing dimension $\{\mathbf{X}_i^b, L_i, \theta_i\}$ for some move types and update these components one by one to increase the efficiency of the sampling process. Haario et al. [12] showed that such Markov Chain Monte Carlo (MCMC) methods (which define proposal distributions that split the dimension of the state-space) are often more efficient and less sensitive to increasing dimension than those proposing moves over the full dimension for high-dimensional spaces [12]. In previous works using RJMCMC ([32] and [18]), a single update move was defined, in which *all* the parameters of a person were updated simultaneously. This was sufficient for simple object models, but we found it to be inefficient for our complex model representing the body, head, and head pose.

3.5 Reversible-Move-Type Definitions

In this work, we define a set of six reversible move types $\Upsilon = \{birth, death, swap, body\ update, head\ update, pose\ update\}$. The traditional update move is split into three component moves for efficiency. The split was made such that the set of parameters modified for each of the update move types only affects a few terms in the observation likelihood: *body update* modifies the location and size of the body ($\mathbf{X}_i^b$), *head update* modifies the location and size of the head (L_i), and *pose update* updates the head pose (θ_i).

1. **Birth**. This adds a new object $\mathbf{X}_{i^*}^*$ with index i^* to the multiobject configuration $\mathbf{X}_t$ while keeping all other objects fixed, forming a proposed state $\mathbf{X}_t^*$. This move implies a dimension change from $m\Gamma \rightarrow m\Gamma + \Gamma$, where Γ denotes the dimension of a single object within the multiobject configuration. The birth move proposes the new multiobject configuration $\mathbf{X}_t^*$, generated from the birth proposal distribution $\mathbf{X}_t^* \sim q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U})$, by applying the transition function h_b and sampling a dimension-matching auxiliary variable $\mathbf{U}$, $\mathbf{U} \sim q(\mathbf{U})$. The birth move transition is given by $\mathbf{X}_t^* = h_b(\mathbf{X}_t, \mathbf{U})$, where the specific objects are defined as

$$\mathbf{X}_{i,t}^* = \begin{cases} \mathbf{X}_{i,t}, & i \neq i^*, \\ \mathbf{U}, & i = i^*. \end{cases} \tag{16}$$

The auxiliary variable $\mathbf{U}$ is responsible for dimension matching in the transition $(\mathbf{X}_t, \mathbf{U}) \rightarrow (\mathbf{X}_t^*)$ (that is, $\mathbf{U}$ acts as a placeholder for the missing dimension in $\mathbf{X}_t$). The proposal for the birth move $q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U})$ is given by

$$q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U}) = \sum_{i \in \mathcal{D}_t \cup \{i^+\}} q_b(i) q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U}, i), \tag{17}$$

where $q_b(i)$ selects the object to be added, i^+ is the next available unused object index, and $\mathcal{D}_t$ is the set of currently dead objects. The target object index sampled from $q_b(i)$ is denoted as i^*, making the proposed set of objects indices a union of the current set $\mathcal{I}_t$ and the target object index i^*, $\mathcal{I}_t^* = \mathcal{I}_t \cup \{i^*\}$. The object-specific proposal distribution for a birth move is given by

$$q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U}, i) = \begin{cases} \frac{1}{C(\mathbf{X}_t)} \frac{1}{N} \sum_{n=1}^{N} p\Big(\mathbf{X}_{i^*,t}^*|\mathbf{X}_{t-1}^{(n)}\Big) \prod_{j \in \mathcal{I}_t} p\Big(\mathbf{X}_{j,t}|\mathbf{X}_{t-1}^{(n)}\Big) & \\ \quad \delta(\mathbf{X}_{j,t}^* - \mathbf{X}_{j,t}) & \text{if } i = i^*, \\ \quad 0 & \text{otherwise}, \end{cases} \tag{18}$$

where in the case of $i = i^*$, the proposal can be rewritten as

$$q_b(\mathbf{X}_t^*|\mathbf{X}_t, \mathbf{U}, i^*) = \frac{1}{C(\mathbf{X}_t)} \left(\frac{1}{N} \sum_{n=1}^{N} \omega_n p\Big(\mathbf{X}_{i^*,t}^*|\mathbf{X}_{t-1}^{(n)}\Big) \right) \prod_{j \in \mathcal{I}_t} \delta\Big(\mathbf{X}_{j,t}^* - \mathbf{X}_{j,t}\Big), \tag{19}$$

where

$$\omega_n = \prod\nolimits_{j \in \mathcal{I}_t} p\Big(\mathbf{X}_{j,t}|\mathbf{X}_{t-1}^{(n)}\Big), \quad C(\mathbf{X}_t) = \frac{1}{N} \sum\nolimits_{n=1}^{N} \omega_n = \frac{1}{N} \sum\nolimits_{n=1}^{N} p_V\Big(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)}\Big). \tag{20}$$

When $i^* = i^+$, a previously unused object index is chosen, and $p(\mathbf{X}_{i^*,t}^*|\mathbf{X}_{t-1}^{(n)})$ reduces to $p(\mathbf{X}_{i^*,t}^*)$ (5). In this case, initial-sized parameters of a new object are sampled from learned Gaussian distributions. Location parameters are selected using cluster sampling for efficiency (a hierarchical process in which the image is broken into smaller regions, a region is randomly selected based on the probability of selecting its contents, and a point is sampled from the selected region) on a smoothed foreground segmented image. If a previously dead object is chosen to be reborn ($i^* \neq i^+$), the new object parameters are taken from the dead object. Initial head and pose parameters are chosen to maximize the head likelihood in both cases. Refer to [29] for further details. After simplification, it can be shown that α_b reduces to

$$\alpha_b = \min \left(1, \frac{p(\mathbf{Z}_t|\mathbf{X}_t^*)}{p(\mathbf{Z}_t|\mathbf{X}_t)} \times \frac{\prod_{j \in \mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}^*, \mathbf{X}_{j,t}^*)}{1} \times \frac{p(\upsilon = d)}{p(\upsilon = b)} \times \frac{q_d(i^*)}{q_b(i^*)} \right). \tag{21}$$

2. **Death**. This is the reverse of a birth move $h_b^R = h_d$. The death move is designed so that it may revert the state back to the initial configuration after a birth, or $(\mathbf{X}_t, \mathbf{U}) = h_d(h_b(\mathbf{X}_t, \mathbf{U}))$. The death move removes an existing object $\mathbf{X}_{i^*,t}$ with index i^* from the state $\mathbf{X}_t$, keeping all other objects fixed. This move implies a dimension change from $m\Gamma \rightarrow m\Gamma - \Gamma$. It proposes a new state $\mathbf{X}^*$ and an auxiliary variable $\mathbf{U}^*$, generated from the death proposal distribution $(\mathbf{X}_t^*, \mathbf{U}^*) \sim q_d(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t)$ by applying the transition function h_{death}. The transition is given by $(\mathbf{X}_t^*, \mathbf{U}^*) = h_d(\mathbf{X}_t)$, where the specific objects are defined as

$$\mathbf{X}_{i,t}^* = \mathbf{X}_{i,t}, \;\; i \neq i^*, \qquad \mathbf{U}^* = \mathbf{X}_{i,t}, \;\; i = i^*. \tag{22}$$

The proposal for the death move $q_d(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t)$ is given by

$$q_d(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t) = \sum_{i \in \mathcal{I}_t} q_d(i) \; q_d(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t, i), \tag{23}$$

where $q_d(i)$ selects the object index i^* to be removed and placed in the set of dead objects $\mathcal{D}_t$, and the object-specific proposal distribution is

$$q_d(\mathbf{X}_t^*, \mathbf{U}^*|\mathbf{X}_t, i) = \begin{cases} \prod_{j \in \mathcal{I}_t, j \neq i^*} \delta(\mathbf{X}_{j,t}^* - \mathbf{X}_{j,t}) & \text{if } i = i^*, \\ 0 & \text{otherwise } (i \neq i^*). \end{cases} \tag{24}$$

In practice, the death move selects an object according to $q_d(i)$ (which is uniform over the set of existing objects in our model) and removes that object from the state-space. Refer to [29] for further details. After simplification, α_d is expressed as

$$\alpha_d = \min \left(1, \frac{p(\mathbf{Z}_t|\mathbf{X}_t^*)}{p(\mathbf{Z}_t|\mathbf{X}_t)} \times \frac{1}{\prod_{j \in \mathcal{C}_i^*} \phi(\mathbf{X}_{i^*,t}, \mathbf{X}_{j,t})} \times \frac{p(\upsilon = b)}{p(\upsilon = d)} \times \frac{q_b(i^*)}{q_d(i^*)} \right). \tag{25}$$

3. **Swap**. This exchanges the parameters of a pair of objects indices i^* and k^*, allowing the tracker to recover from events in which the identity of two people become confused (for example in occlusion). The transition is given by $\mathbf{X}_t^* = h_s(\mathbf{X}_t)$, where specific objects are defined as follows:

$$\mathbf{X}_{i,t}^* = \begin{cases} \mathbf{X}_{i,t}, & i \neq i^*, i \neq k^*, \\ \mathbf{X}_{k^*,t}, & i = i^*, \\ \mathbf{X}_{i^*,t}, & i = k^*. \end{cases} \tag{26}$$

The proposal for the swap move $q_s(\mathbf{X}_t^*|\mathbf{X}_t)$ is defined as

$$q_s(\mathbf{X}_t^*|\mathbf{X}_t) \triangleq \sum_{i,k \in \mathcal{I}_t} q_s(i,k) \; q_s(\mathbf{X}_t^*|\mathbf{X}_t, i, k), \tag{27}$$

where the target object indices i^* and k^* are randomly sampled from $q_s(i,k)$. The object-specific proposal distribution exchanges the state values and histories (past state values) of objects i^* and k^*. It can be shown [29] that the expression for the α_s reduces to

$$\alpha_s = \min\left(1, \frac{p(\mathbf{X}_t^*|\mathbf{Z}_{1:t})}{p(\mathbf{X}_t|\mathbf{Z}_{1:t})}\right). \tag{28}$$

4. **Body update**. This modifies the body parameters of a current object $\mathbf{X}_{i,t}^{*b}$, with index $i = i^*$ keeping the head of person i^* and all other people fixed. The update move transition is given by $(\mathbf{X}_t^*, \mathbf{U}^*) = h_{body}(\mathbf{X}_t, \mathbf{U})$, where the specific objects are defined as

$$\left(\mathbf{X}^{*b}_{i,t}, \mathbf{X}^{*h}_{i,t}\right) = \begin{cases} \left(\mathbf{X}^{b}_{i,t}, \mathbf{X}^{h}_{i,t}\right) & i \neq i^*, \\ \left(\mathbf{U}, \mathbf{X}^{h}_{i,t}\right) & i = i^*, \end{cases} \qquad \mathbf{U}^* = \mathbf{X}^{b}_{i^*,t}. \tag{29}$$

The body update move proposal is defined as

$$q_{body}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}) = \sum_{i \in \mathcal{I}_t} q_{body}(i)\ q_{body}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i). \tag{30}$$

The object-specific proposal distribution is defined as

$$\begin{aligned} &q_{body}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i) = \\ &\frac{1}{N}\sum_n p\left(\mathbf{X}^{*b}_{i^*,t}|\mathbf{X}^{b,(n)}_{t-1}\right) p\left(\overline{\mathbf{X}^{*b}_{i^*,t}}|\mathbf{X}^{b,(n)}_{t-1}\right) \delta\left(\overline{\mathbf{X}^{*b}_{i,t}} - \overline{\mathbf{X}^{b}_{i^*,t}}\right) \\ &\prod_{j \neq i^*} p\left(\mathbf{X}_{j,t}|\mathbf{X}^{(n)}_{t-1}\right) \delta\left(\mathbf{X}^*_{j,t} - \mathbf{X}_{j,t}\right), \end{aligned} \tag{31}$$

where $\overline{\mathbf{X}^{*b}_{i^*,t}}$ denotes all state parameters except $\mathbf{X}^{*b}_{i^*,t}$, and $\mathbf{X}^{*b}_{i,t}$ denotes the proposed body configuration for target i^*. This implies randomly selecting a person i^* and sampling a new body configuration for this person from $p(\mathbf{X}^{*b}_{i^*,t}|\mathbf{X}^{b,(n^*)}_{t-1})$ by using an appropriate sample n^* from $t-1$, leaving the other parameters unchanged. Thus, α_{body} can then be shown to reduce to [29]

$$\alpha_{body} = \min\left(1, \frac{p(\mathbf{Z}^b_t|\mathbf{X}^{*b}_{i^*,t}) \prod_{l \in \mathcal{C}_{i^*}} \phi(\mathbf{X}^*_{i^*,t}, \mathbf{X}^*_{l,t})}{p(\mathbf{Z}^b_t|\mathbf{X}^{b}_{i^*,t}) \prod_{l \in \mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}, \mathbf{X}_{l,t})}\right). \tag{32}$$

5. **Head update**. This modifies the head parameters of a current object $L^{*h}_{i^*,t}$ with index i^*. The transition is given by $(\mathbf{X}^*_t, \mathbf{U}^*) = h_{head}(\mathbf{X}_t, \mathbf{U})$, where the specific objects are defined as

$$\left(\mathbf{X}^{*b}_{i}, L^*{}_i, \theta^*_i\right) = \begin{cases} (\mathbf{X}^{b}_{i,t}, L_{i,t}, \theta_{i,t}) & i \neq i^*, \\ (\mathbf{X}^{b}_{i,t}, \mathbf{U}, \theta_{i,t}) & i = i^*, \end{cases} \qquad \mathbf{U}^* = L_{i,t}. \tag{33}$$

The head update move proposal is defined as

$$\begin{aligned} &q_{head}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}) = \\ &\sum_{i \in \mathcal{I}_t} q_{head}(i)\ q_{head}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i), \end{aligned} \tag{34}$$

where the object-specific proposal distribution is defined as

$$\begin{aligned} &q_{head}(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i) = \\ &\frac{1}{N}\sum_n p\left(L^*_{i^*,t}|\mathbf{X}^{(n)}_{t-1}\right) p\left(\overline{L^*_{i^*,t}}|\mathbf{X}^{(n)}_{t-1}\right) \delta\left(\overline{L^*_{i^*,t}} - \overline{L_{i^*,t}}\right) \\ &\prod_{j \neq i^*} p\left(\mathbf{X}_{j,t}|\mathbf{X}^{(n)}_{t-1}\right) \delta(\mathbf{X}^*_{j,t} - \mathbf{X}_{j,t}), \end{aligned} \tag{35}$$

where $\overline{L^*_{i^*,t}}$ denotes all state parameters except $L^*_{i^*,t}$. This implies selecting a person i^* and sampling a new head configuration for this person from $p(\mathbf{X}^{*h}_{i^*,t}|\mathbf{X}^{h,(n^*)}_{t-1})$ by using an appropriate sample n^* from the $t-1$ time, leaving the other parameters unchanged. α_{head} can then be shown to reduce to [29]

$$\alpha_{head} = \min\left(1, \frac{p(\mathbf{Z}^h_t|L^*{}_{i^*,t})}{p(\mathbf{Z}^h_t|L_{i^*,t})} \times \frac{p(L^*{}_{i^*,t}|\mathbf{X}^{*b}_{i^*,t})}{p(L_{i^*,t}|\mathbf{X}^{b}_{i^*,t})}\right), \tag{36}$$

6. **Pose update**. This modifies the pose parameter θ_{t,i^*} of a person with index i^*. Like the previous update moves, it is self reversible and does not change the dimension of the state. The move transition is given by $(\mathbf{X}^*, \mathbf{U}^*) = h_\theta(\mathbf{X}, \mathbf{U})$, where

$$\left(\mathbf{X}^{*b}_{i}, L^*{}_i, \theta^*_i\right) = \begin{cases} (\mathbf{X}^{b}_{i}, L_{i}, \theta_{i}) & i \neq i^*, \\ (\mathbf{X}^{b}_{i}, L_{i}, \mathbf{U}) & i = i^*, \end{cases} \qquad \mathbf{U}^* = \theta_{i^*}. \tag{37}$$

The head-pose update move proposal is defined as

$$q_\theta(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}) = \sum_{i \in \mathcal{I}_t} q_\theta(i)\ q_\theta(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i), \tag{38}$$

where the object-specific proposal distribution is defined as

$$\begin{aligned} &q_\theta(\mathbf{X}^*_t, \mathbf{U}^*|\mathbf{X}_t, \mathbf{U}, i) = \\ &\frac{1}{N}\sum_n p\left(\theta^*_{i^*,t}|\theta^{(n)}_{t-1}\right) p\left(\overline{\theta^*}_{i^*,t}|\overline{\theta^*}^{(n)}_{t-1}\right) \delta\left(\overline{\theta^*}_{i^*,t} - \overline{\theta}_{i^*,t}\right) \\ &\prod_{j \neq i^*} p\left(\mathbf{X}_{j,t}|\mathbf{X}^{(n)}_{t-1}\right) \delta(\mathbf{X}^*_{j,t} - \mathbf{X}_{j,t}), \end{aligned} \tag{39}$$

where $\theta^*_{i^*,t}$ denotes the proposed head-pose configuration for target i^*, and $\overline{\theta^*}_{i^*,t}$ denotes all state parameters except $\theta^*_{i^*,t}$. This implies selecting a person index i^* and sampling a new head-pose for this person from $p(\theta^*_{i^*,t}|\theta^{(n^*)}_{t-1})$ by using an appropriate sample n^* from the previous time step, leaving the other parameters unchanged. α_θ can then be shown [29] to reduce to

$$\alpha_\theta = \min\left(1, \frac{p(\mathbf{Z}^h_t|\mathbf{X}^{*h}_{i^*,t})}{p(\mathbf{Z}^h_t|\mathbf{X}^{h}_{i^*,t})}\right). \tag{40}$$

3.6 Inferring a Solution

The first N_b samples added to the Markov Chain are part of the *burn-in* period, which allows the Markov Chain to reach the target density. The chain, after this point, approximates the filtering distribution, which represents a belief distribution of the current state of the objects, given the observations. It does not, however, provide a single answer to the tracking problem. To find this, we compute a *point estimate solution*, which is a single state computed from the filtering distribution, which serves as the tracking output. To determine the set of objects in the scene, we compute the mode of the object configurations in the Markov Chain (each sample contains a set of object indices, and we select the set that is repeated most often, accounting for identity changes resulting from swap moves). Using these samples, we find the mean configuration of each of the body and head spatial configuration parameters $(\mathbf{X}^b_{i,t}, L^h_{i,t})$. For the out-of-plane head rotations represented by the discrete exemplar θ_i, we compute the mean of the corresponding Euler angles for pan and tilt. The detailed steps of our joint multiperson body-head tracking and VFOA-W estimation model are summarized in Fig. 6.

At each time step, t, the posterior distribution of (8) for the previous time step is represented by a set of N *unweighted* samples $p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) \approx \{\mathbf{X}_{t-1}^{(n)}\}_{n=1}^{N}$. The approximation of the current distribution $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$ is constructed according to steps 1 and 2, from which a *point estimate solution* for head and body parameters is determined in step 3. The values of these parameters are used in step 4 to determine if a person's attention is directed at the advertisement (*focused*) or not (*unfocused*).

1) Initialize the Markov Chain by choosing a sample from the $t-1$ Markov Chain with the mode configuration (m_{t-1}^{mode}). Apply the motion model to each object, $\prod_{i\in\mathcal{I}_t} p(\mathbf{X}_{t,i}|\mathbf{X}_{t-1,i}^{(n)})$, and accept as sample $n=0$.
2) RJMCMC Sampling. Draw $N+N_B$ samples according to the following schedule.
   - 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^*, k^* for swap) according to the target proposal $q_\upsilon(i)$ for chosen move type.
   - Sample New Configuration $\mathbf{X}^*{}_t$ from the move-specific proposal distribution q_{υ^*}. For move type υ, this implies:
     - *Birth* - add a new person i^* according to (17), $m_t^{(n)*} = m_t^{(n)} + 1$.
     - *Death* - remove an existing person i^* according to (23), $m_t^{(n)*} = m_t^{(n)} - 1$.
     - *Swap* - swap the parameters of two existing people i^*,k^* $\mathbf{X}_{i,t}^{(n)} \rightarrow \mathbf{X}_{k,t}^{(n)*}$, $\mathbf{X}_{k,t}^{(n)} \rightarrow \mathbf{X}_{i,t}^{(n)*}$.
     - *Body Update* - update the body parameters $X_{i,t}^{b,(n)*}$ of an existing person i^*.
     - *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 (21), (25), (28), (32), (36), or (40).
   - Accept/Reject. Accept the proposal $\mathbf{X}^*{}_t$ if $\alpha \geq 1$, otherwise accept with probability α. If accepted, add it to the Markov Chain $\mathbf{X}_t^{(n)} = \mathbf{X}^*{}_t^{(n)}$. If rejected, add the previous sample in the Markov Chain to the current position $\mathbf{X}_t^{(n)} = \mathbf{X}_t^{(n-1)}$.
3) Compute a Point Estimate Solution from the Markov Chain (as in Section 3.6):
   - to avoid bias in the Markov Chain, discard the first N_B *burn-in* samples. The sample set $\{\mathbf{X}_t^{(n)}\}_{n=N_B+1}^{N_B+N}$ represents an approximation of the filtering distribution.
   - form a sample set W from the mode configuration $\hat{X}_t$ as described in Section 3.6. Compute the *point estimate* body $\hat{X}_t^b$ and head $\hat{X}_t^h$ parameters from their mean value in W.
4) Determine the VFOA-W for each person in the scene according to Section 4.

Fig. 6. Algorithm for joint multiperson body and head tracking and VFOA-W estimation with RJMCMC.

4 Modeling the VFOA for a Varying Number of Wandering People

The VFOA-W task is to automatically detect and track a varying number of people able to move about freely and to estimate their VFOA. The VFOA-W problem is significantly more complex than the traditional VFOA problem because it allows for the number of people in the video to vary, and it allows for the people in the video to freely walk about the scene, whereas in previous works [36], the number of people appearing in a single video was fixed, and they were constrained to remain seated (for their VFOA to be estimated). The advertising application chosen as an introduction to VFOA-W represents a relatively simple instance of the problem, as we only attempt to measure VFOA for a single visual target, though it is straightforward to extend this model for multiple targets.

At each time t, a person's VFOA-W is defined as being in any of the two states f_t:

- *Focused* $f_t = 1$: looking at the advertisement.
- *Unfocused* $f_t = 0$: not looking at the advertisement.

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

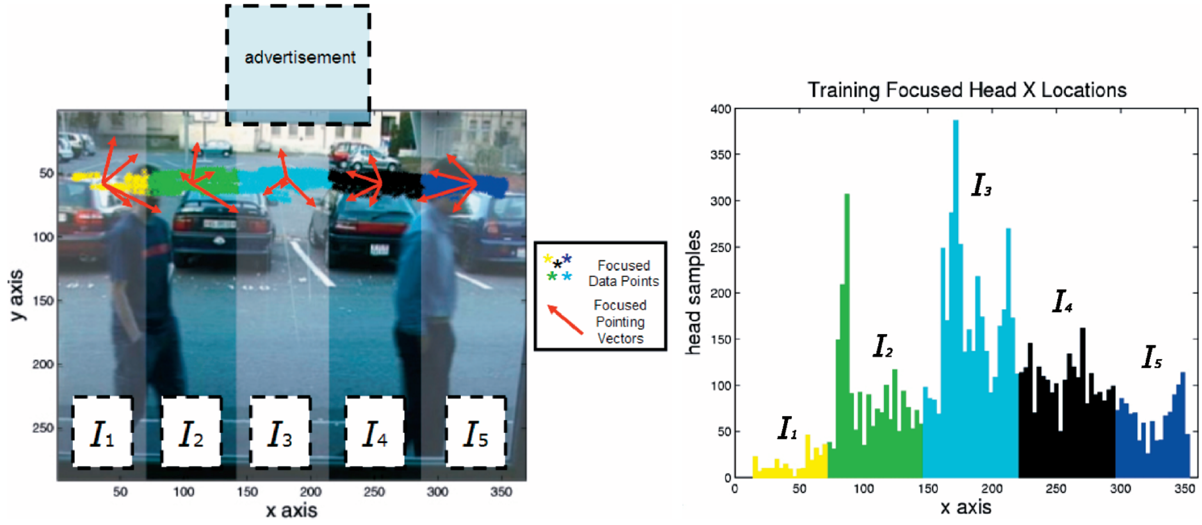
4.1 VFOA-W Modeling with a GMM

Estimating VFOA-W can be posed in a probabilistic framework as finding the focus state maximizing the a posteriori probability $\hat{f} = \arg\max_f p(f|z^h) \propto p(z^h|f)p(f)$, where z_t^h is the head pointing vector of the person parameterized by a pan and tilt angle (see Fig. 4). We assume the prior on the VFOA-W state $p(f)$ to be uniform; thus, it has no effect on the VFOA-W estimation. To model the probability of being in a focused state, we consider the horizontal head position x^h and head pointing vector (see Fig. 7). Because the target 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, $k=\{1,\dots,5\}$, and modeled the probability of a focused state as

$$\begin{aligned} p\big(z^h|f=1\big) &= \sum_{k=1}^{K} p\big(x^h \in I_k, z^h|f=1\big) \\ &= \sum_{k=1}^{K} p(x^h \in I_k)p(z^h|x^h \in I_k, f=1), \end{aligned} \tag{41}$$

where the first term $p(x^h \in I_k)$ models the probability of a person's head location belonging to region I_k, and the second term $p(z^h|x^h \in I_k, f=1)$ models the probability of focused head pose, given the region that the head belongs to. The inclusion of the head location in modeling the VFOA-W allowed us to solve an issue that was not previously addressed in [24], [34], [38]: Resolving the VFOA-W of a person whose focus state depends on their location.

The terms of the VFOA-W model in (41) are defined as follows: Each region is defined by its center and width, denoted by x_{I_k} and σ_{I_k}, respectively. The probability of a head location x^h belonging to region I_k is modeled by a



(a) (b)

Fig. 7. **VFOA-W modeling**. (a) VFOA-W is determined by *head pose* and horizontal *position* in the image. The horizontal axis is split into $K_{vfoa-w} = 5$ regions $(I_1, \ldots, I_5)$, and a VFOA-W 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. (b) Over 9,400 training points representing a person in a *focused* state (also seen in (a)) were split into five regions along the horizontal axis and used for training a model for each region.

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 Fig. 7. The probability of being unfocused is modeled as a uniform distribution $p(z^h|f = 0) = T_{vfoa-w}$.

The parameters of the VFOA-W model (Gaussian mean and covariance matrix) and the uniform distribution modeling the unfocused state distribution were learned from the training data described in Section 5. Though our VFOA-W model does not make use of the vertical head location, it is straightforward to generalize the model to do this. To reduce noisy VFOA-W estimations, a smoothing filter with a 10-frame window was applied to the GMM output.

4.2 VFOA-W Modeling with an HMM

The VFOA-W GMM does not take into account the temporal dependencies between the focus states. Such dependencies can be modeled using an HMM. If we denote a sequence of focus states by $f_{1:T}$ and a sequence of head pose observations as $z^h_{1:T}$, the joint posterior probability of the observation and the states can be written as

$$p(f_{1:T}, z^h_{1:T}) = p(f_0) \prod_{t=1}^{T} (z^h_t | f_t) p(f_t | f_{t-1}). \tag{42}$$

In this equation, the emission probabilities $p(z^h_t|f_t)$ are modeled as before (GMM for focused and uniform for unfocused). However, in the HMM case, a transition matrix is used for modeling the temporal VFOA-W state transition $p(f_t|f_{t-1})$, which is defined as a 0.8 probability to stay in the same state and 0.2 to change state. Given $z^h_{1:T}$, VFOA-W recognition is done by finding the optimal sequence maximizing $p(f_{1:T}|z^h_{1:T})$ using the Viterbi algorithm [27].

5 Training and Parameter Selection

5.1 Experimental Setup

To simulate the advertising application described in the Introduction, a fake advertisement was placed in an exposed window, with a camera set behind. Several actors were instructed to pass in front of the window and were allowed to look at the advertisement (or not) as they would naturally (actors were used due to privacy concerns for actual passers-by). A recording of a 10-min duration (360×288 resolution, 25 fps) was made, in which a maximum of three people appear in the scene simultaneously. The recorded data includes challenging events such as people occluding each other in sequence h and people entering/exiting the scene.

5.2 Training and Parameter Selection

The recorded video data was organized into a disjoint training and test set of equal size. The training set, consisting of nine sequences (for a total of 1,929 frames), was manually annotated for body location, head location, and focused/unfocused states.

Table 1 provides a list of many of the key parameters of our model. Parameters were either learned automatically from training data (*learned*), tuned by hand (*hand-tuned*), or selected without exhaustive tuning (*untuned*). The parameters for the foreground segmentation were hand tuned by observing results on the training set. The binary body model was trained using background subtraction and training set annotations. Using this information, GMMs were trained for the foreground and background models (K parameters were selected through cross validation). Head annotations were used for learning the parameters of the Gaussian skin color distribution in the head-pose skin feature. The silhouette mask was also trained using the head annotations by averaging the binary patches corresponding to head annotations. Parameters for the VFOA-W model, including T_{vfoa-w}, were optimized on the training data (bootstrapped to 9,400 training

TABLE 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}	4	learned	*observation model* body binary model number of Gaussians (background)
λ_F	20	hand-tuned	*observation model* body color foreground parameter
λ_{sil}	200	hand-tuned	*observation model* head silhouette parameter
Z_θ, λ_θ^{tex}	-	learned	*observation model* head texture parameters
T_{tex}	$exp(\frac{-9}{2})$	untuned	*observation model* head texture threshold
λ_{sk}	0.5	hand-tuned	*observation model* head skin color parameter
p_{birth}	0.05	untuned	*RJMCMC* prior probability of choosing a *birth* move
p_{death}	0.05	untuned	*RJMCMC* prior probability of choosing a *death* move
p_{swap}	0.05	untuned	*RJMCMC* prior probability of choosing a *swap* move
p_{body}	0.283	untuned	*RJMCMC* prior probability of choosing a *body update* move
p_{head}	0.283	untuned	*RJMCMC* prior probability of choosing a *head update* move
p_{pose}	0.283	untuned	*RJMCMC* prior probability of choosing a *pose update* move
N	300,600,800	hand-tuned	*RJMCMC* number of samples in chain for 1,2,3 simultaneous people, resp.
N_B	0.25*N	hand-tuned	*RJMCMC* number of *burn-in* samples
K_{vfoa-w}	5	untuned	*VFOA-W model* number of Gaussians
T_{vfoa-w}	*0.00095*	learned	*VFOA-W model* likelihood threshold

points; see Fig. 7) to achieve the highest VFOA-W event recognition performance (see Section 6 for details). The training set was also used for learning prior-sized models (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 Fig. 4). The texture parameters Z_θ and λ_θ^{tex} were learned according to the method described in [40].

6 Evaluation

In order to evaluate the performance of our application, the test set was annotated similarly to the training set. The test set consists of 10 sequences summarized in Table 2. Sequences a to d contain three people (appearing sequentially) passing in front of the window. Sequences e to i contain two people, whereas sequence j contains three people appearing simultaneously. We compared our results with the ground truth over 200 experiments on the 10 test sequences (as we use a stochastic process, we ran 20 runs/sequence). The length of the Markov Chain was chosen such that there was a sufficient number of samples for good-quality tracking (see Table 1). Experimental results are illustrated in Fig. 9 and fully shown in companion videos [14].

TABLE 2
Test Set Data Summary

sequence	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
length (s)	15	13	12	10	5	6	4	4	4	11
# people (simultaneous / total)	(1 / 3)				(2 / 2)					(3 / 3)
# looks at advertisement	2	3	0	3	2	2	2	1	2	4

6.1 Multiperson Body and Head Tracking Performance

To evaluate the tracking performance, we adopt a set of measures proposed in [31], with some minor changes to names and notation. These measures evaluate three tracking qualities: The ability to estimate the number and placement of people in the scene (*detection*), how tightly the estimated bounding boxes fit the ground truth (*spatial fit*), and the ability to persistently track a particular person over time (*tracking*). The overall results are given in Table 3, with illustrations for sequences b, e, h, and i in Fig. 9 and further details available at [14].

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 $\overline{FN}$. As indicated in Table 3, for a given person in a given frame, there is a 1.8 percent chance of our method producing a false-positive error and a 1.1 percent chance of producing a false-negative error. The *Counting Distance* $\overline{CD}$ measures how close the estimated number of

TABLE 3
Multiperson Tracking Results Averaged over the Entire Test Set

Tracking Quality Measured	Measure	Value
detection	False positive rate	$\overline{FP} = .0183 \pm .0031$
	False negative rate	$\overline{FN} = .0107 \pm .0038$
	Counting distance	$\overline{CD} = .0344 \pm .0078$
spatial fit	Body fit	$\overline{fit} = .8655 \pm .0075$
	Head fit	$\overline{fit} = .8484 \pm .0078$
tracking	Tracking purity	$\overline{P} = .9280 \pm .0171$

people is to the actual number (normalized per person per frame). A $\overline{CD}$ value of zero indicates a perfect match. As shown in Table 3, the $\overline{CD}$ is near zero, indicating good performance.

Spatial fitting between the ground truth region and the tracker output is measured for the body and the head by using the *f-measure* $F = \frac{2\nu\rho}{\nu+\rho}$, where ρ is recall, and ν is precision. A perfect fit is indicated by $F = 1$, no overlap by $F = 0$. Table 3 indicates that the spatial fitting for both the head and body were quite good, which is above 80 percent.

To evaluate the tracking performance, we rely on the *purity* measure, which estimates the degree of consistency with which the estimates and ground truths were properly identified ($\overline{P}$ near 1 indicates a well-maintained identity, and $\overline{P}$ near 0 indicates poor performance; see [31] for details). Table 3 shows that our model had good tracking quality (0.93), though it dropped to 0.81 in sequence h, where two people occlude one another as they cross paths.

6.2 Advertisement Application Performance

To evaluate the performance of the advertisement application, the results from our model were compared with ground truth annotations. Results appear in Fig. 8 (summarized in Table 4) and the companion videos [14]. For evaluation, we considered six criteria defined below and report results for the GMM and HMM models for each. To reduce errors caused by people partially appearing in the image, VFOA-W results are computed on a region of interest defined from eight frames after a person appears until eight frames before they exit the scene.

1. **The number of people exposed to the advertisement**. Over the entire test set, 25 people passed the advertisement, whereas our RJMCMC tracking model estimated that 25.15 people appeared on the average (over 20 runs, $std\ dev = 0.17$). In Fig. 8a, we can see that the number of people was correctly estimated for every sequence, except for a, c, and i.
2. **The number of people who looked at the advertisement**. Of the 25 people, 20 actually focused on the advertisement at some point. The GMM VFOA-W model estimated that 22.95 people looked (mean), whereas the HMM VFOA-W model estimated that 21.2 people looked (mean).
3. **The number of events where someone looked at the advertisement**. The VFOA-W recognition sequences were broken into continuous segments or events, where a look-event is a focused state for $t \geq 3$ frames. Twenty-one look-events actually occurred over the test set. The GMM model estimated 28.5 look-events (mean), whereas the HMM model estimated 21.45 look-events (mean). This result was determined through a standard symbol-matching technique.
4. **Time spent looking at the advertisement**. Over the entire test set, people spent 37.28 sec looking at the advertisement. The GMM model estimated that people looked at the ad for 38.59 seconds (mean), whereas the HMM estimated 37.89 seconds (mean).
5. and 6. **VFOA-W recognition rate estimation**. The VFOA-W recognition rate is computed with respect to frames and events (continuous segments of frames with a similar VFOA-W state). The frame-based recognition rate is computed directly as the number of frames in which the estimate and ground truth agree over the number of frames. The overall frame-based recognition rates are 83.90 percent (mean GMM) and 92.53 percent (mean HMM). The aforementioned *F-measure*, $F = \frac{2\rho\nu}{\rho+\nu}$, is used for computing the event-based recognition rate [16], where ρ is the event-based recall (the number of segments where the ground truth and estimate agree, normalized by the number of segments in the ground truth), and ν is the precision (the number of segments where the ground truth and estimate agree, normalized by the number of segments in the estimate). The overall event-based recognition rates are 90.37 percent (mean GMM) and 93.85 percent (mean HMM). Results for each sequence appear in Figs. 8e and 8f.

6.3 Varying the Number of Particles

To study the model's dependency on the number of samples, we conducted a series of experiments on sequence i, which is omitted for space reasons. In summary, $N = 600$ samples were required for good performance on an Intel Pentium IV 3.2 GHz processor (in Matlab, between < 1 and 5 seconds of processing time per frame). We refer the reader to [14] for details.

7 Discussion and Limitations

Although our proposed model yielded convincing results on the preceding experiments, there exist some limitations to the models and data set. In this section, we discuss some of these limitations and how they might be addressed in future work.

7.1 Multiperson Tracking

Separability of classes in the binary background observation model limits the number of people that the model can track simultaneously. As the number of people increases, the learned background model loses its ability to discriminate between different numbers of objects (that is, the fewer the objects in the scene, the more confident our estimation). In independent experiments, the binary observation model was found to be robust for up to five simultaneous objects, though this limitation depends on the typical size of the objects with respect to the scene and the variability of object size. An

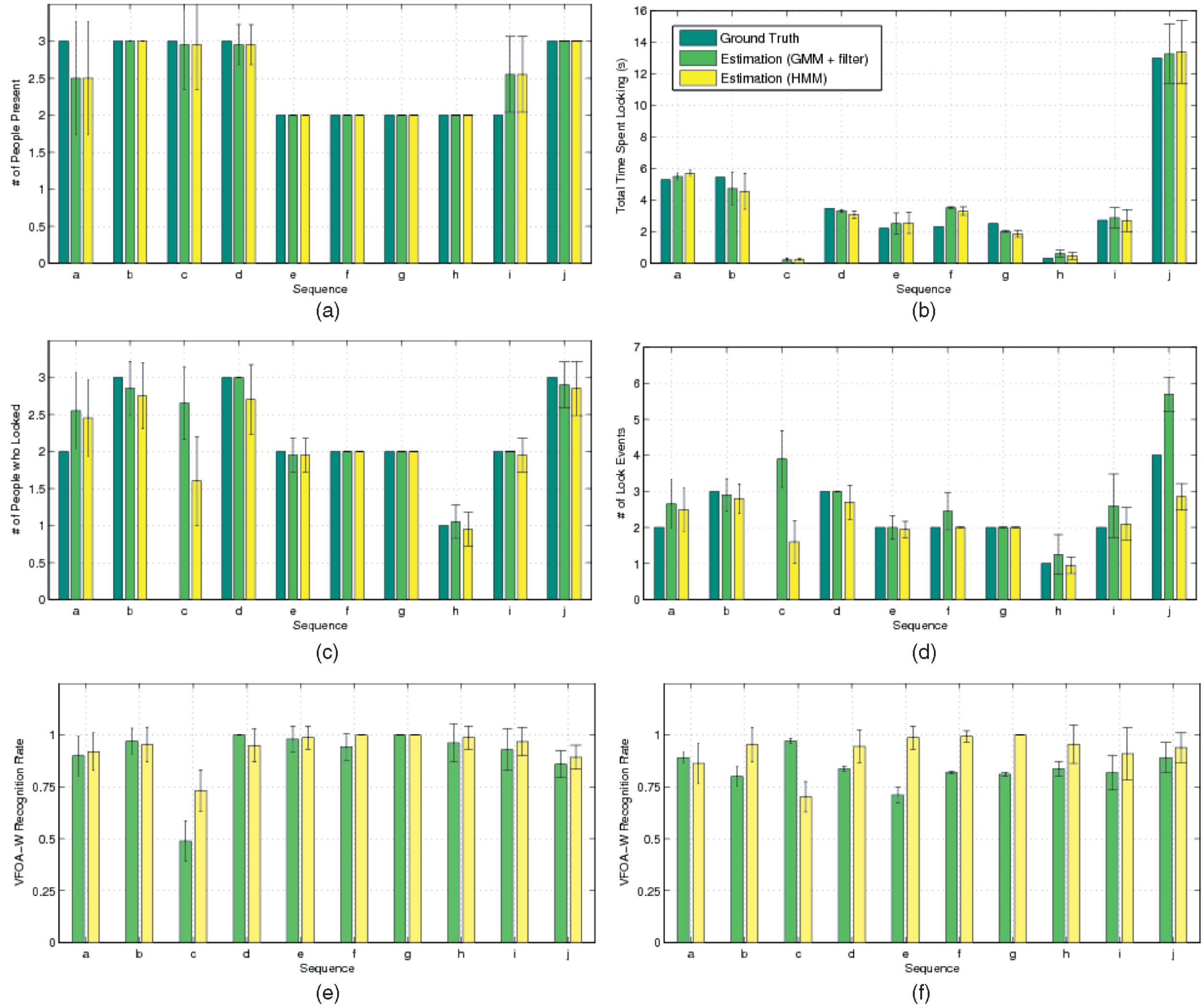


Fig. 8. **Advertisement application VFOA-W results**. (a) The results for estimating the number of people exposed to the advertisement for each sequence. (b) The results for estimating the amount of time people spent looking. (c) The results for estimating the number of people who looked. (d) The results for estimating the number of "look events." (e) The event-based recognition rate of *focused* and *unfocused* states. (f) The frame-based recognition rate. The ground truth appears in dark green, GMM results in light green, and HMM results in yellow.

alternative approach to the binary observation model proposed in [33] addresses this limitation.

Our observation model is also limited in its ability to handle occlusion. Though it performs well for full occlusion in our experiments (with a relatively small number of people), our approach would be less robust in situations where a monocular camera view is insufficient to resolve the occlusion due to the camera placement or multiple occlusions. This is a common problem to monocular tracking algorithms. A multiview approach such as that proposed by Fleuret et al. [5] may better address these types of situations, which can occur in realistic environments.

Finally, because it models relative size and overlap of the foreground and background, the binary observation model is not robust in situations where the typical size of a person varies dramatically (for example, if a person appears much smaller in the background than in the foreground).

7.2 Head Tracking and Pose Estimation

The head-pose estimation is principally limited by the performance of the texture and skin color models. The performance of these models is dependent on the resolution of the head in the image. Lower resolution leads to greater error in the head-pose estimation (and, thus, the VFOA-W estimation). In our experiments, the head was typically approximately 40×60 pixels. In [4], the head-pose model presented in this work was shown to yield good tracking results for head sizes of 20×30 pixels, though data from multiple cameras were used.

The performance of the texture and skin color models also depends on the placement of the camera relative to the head. Experiments in [3] show that our head-pose model performs better for near-frontal faces (12 degrees of mean error) than for faces near profile poses (18 degrees of mean error).

7.3 VFOA-W Modeling

The relatively simple x-axis positional model used for VFOA-W is sufficient to yield good results to estimate VFOA for moving people. A more complex scenario may require a more geometrically complex VFOA-W model, which takes into account the observed head pose and the locations of the advertisement, person, and camera.

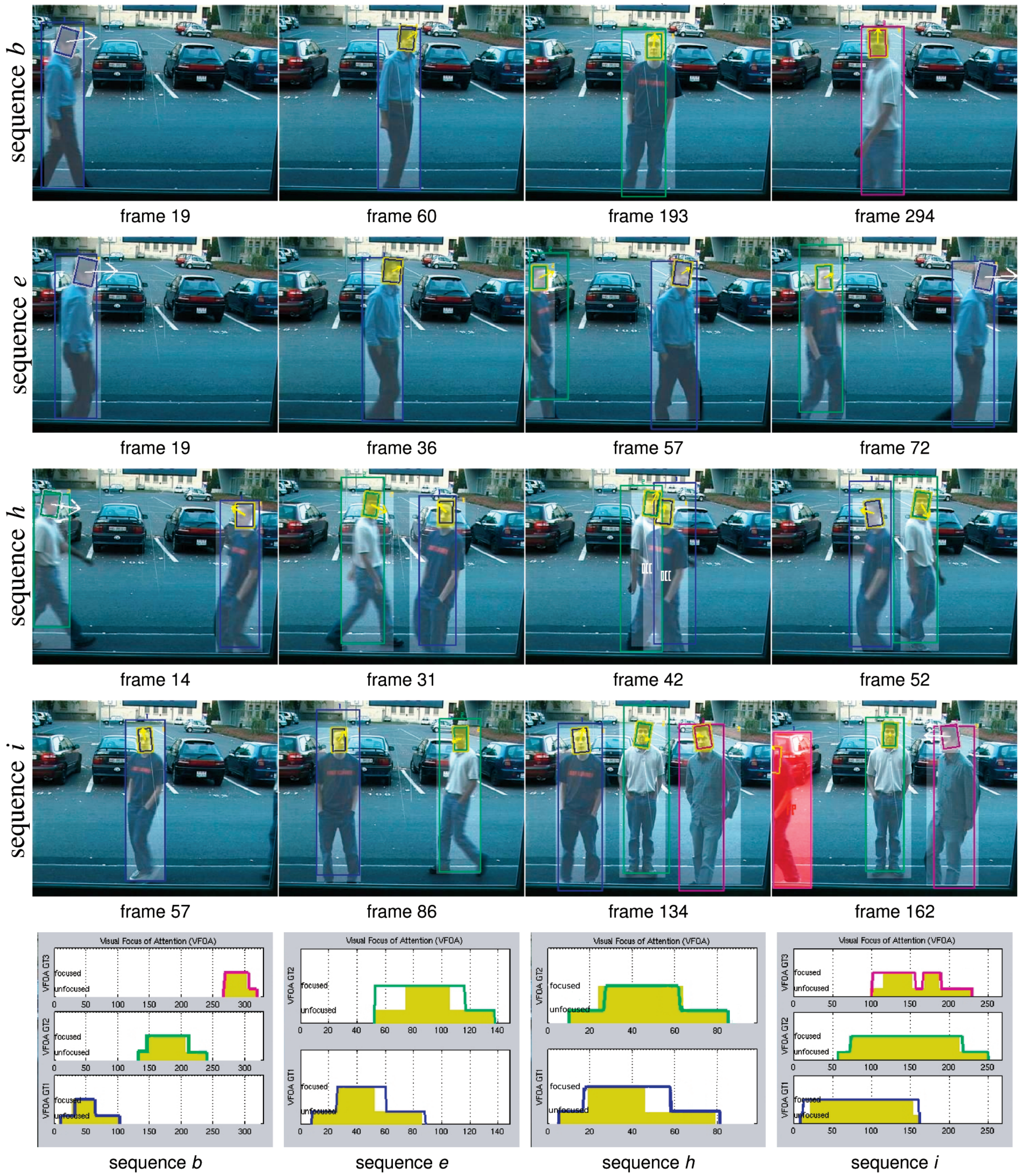


Fig. 9. **Tracking and VFOA-W results for sequences b, e, h, and i**. Tracking results appear as boxes around the body and head. A yellow pointing vector/head border indicates a *focused* state, whereas a white pointing vector/head border indicates an *unfocused* state. The ground truth appears as shaded boxes for the head and the body (the head area is shaded yellow when labeled as *focused* and is gray when labeled as *unfocused*). VFOA-W results for the GMM model appear at the bottom. The yellow bars represent the ground truth (raised indicates a *focused* state, lowered indicates *unfocused*, and no yellow bar indicates that the person is not present in the scene). GMM VFOA-W estimates appears as colored lines. VFOA-W performance was nearly perfect for b, with good event-based recognition in all sequences. Mild frame-based VFOA-W recognition errors occurred in e, h, and i. Frame 162 of sequence i shows an FP error generated as a tracker was placed where no ground truth was present.

7.4 Data Set

Although our data set was useful in demonstrating the ability of our VFOA-W algorithm to perform in a realistic situation, it contains some limitations. First, only four actors appeared throughout the data set. Second, the actors did not walk into the far background and, thus, their size did not vary appreciably. Third, the maximum number of actors appearing simultaneously did not exceed three, and the actors only crossed paths in one test sequence and one training sequence (causing an occlusion). Finally, though tested outdoors, the lighting conditions were relatively stable. The design of a future VFOA-W data set should take these issues into account.

TABLE 4
VFOA-W Estimation Summary for GMM and HMM Models

	error rates (in % of the ground truth)				VFOA-W recognition rates	
	# people	time focused	# people looked	# look events	event-based	frame-based
hmm	3.40%	1.63%	6.00%	2.14%	93.95%	92.53%
gmm + filter	3.40%	3.51%	14.75%	35.45%	90.37%	83.90%

8 CONCLUSION

In this paper, we have introduced *VFOA-W* and presented a principled probabilistic approach to solving it. Our approach expands on state-of-the-art RJMCMC tracking models, with novel contributions to object modeling, likelihood modeling, and the sampling scheme. It is a general model that can be easily adapted to similar tasks. We applied our model to a real-world advertising application and provided a rigorous objective evaluation of its performance in this context. We compared two VFOA-W models (the GMM and HMM) and found the temporal dependencies of the HMM to yield superior performance. From these results, we have shown that our proposed model is able to track a varying number of moving people and determine their VFOA-W with good quality (exhibiting only a 6 percent error rate in determining the number of people who looked at the ad). For future work, it may be useful to investigate using a spatially dependent face/pose detector to help localize and model the head, which may outperform some of our head pose tracking features.

ACKNOWLEDGMENTS

This work was funded by the Swiss National Science Foundation (SNSF) through the National Center for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2), by the European Information Society Technologies (IST) Augmented Multiparty Interaction with Distant Access Project (AMIDA, FP6-033812, AMIDA-35), and by the European IST Content Analysis and Retrieval Technologies Applied to Knowledge Extraction from Massive Recordings (CARETAKER) Project. They also thank their colleagues at the IDIAP Research Institute, who contributed their time for recording the data set. This work was done while K. Smith was at the IDIAP Research Institute.

REFERENCES

[1] C. Andrieu, N. de Freitas, and M. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, no. 1, pp. 5-43, 2003.
[2] S. Ba, "Joint Head Tracking and Pose Estimation for Visual Focus of Attention Recognition," PhD dissertation, École Polytechnique Fédérale de Lausanne (EPFL), Feb. 2007.
[3] S. Ba and J. Odobez, "Evaluation of Multiple Cues Head-Pose Tracking Algorithms in Indoor Environments," *Proc. Int'l Conf. Multimedia and Expo*, July 2005.
[4] S. Ba and J.M. Odobez, "Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups," *Classification of Events, Activities and Relationships*, 2007.
[5] J. Berclaz, F. Fleuret, and P. Fua, "Robust People Tracking with Global Trajectory Optimization," *Proc. Computer Vision and Pattern Recognition*, 2006.
[6] A. Bhattacharyya, "On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99-109, 1943.
[7] L. Brown and Y. Tian, "A Study of Coarse Head-Pose Estimation," *Proc. Workshop Motion and Video Computing*, 2002.
[8] T. Cootes, G. Edwards, and J. Taylor, "Active Appearance Model," *Proc. Fifth European Conf. Computer Vision*, June 1998.
[9] M. Danninger, R. Vertegaal, D. Siewiorek, and A. Mamuji, "Using Social Geometry to Manage Interruptions and Co-Worker Attention in Office Environments," *Proc. Conf. Graphics Interface*, 2005.
[10] A. Gee and R. Cipolla, "Estimating Gaze from a Single View of a Face," *Proc. Int'l Conf. Pattern Recognition*, Oct. 1994.
[11] P. Green, "Reversible Jump MCMC Computation and Bayesian Model Determination," *Biometrika*, vol. 82, pp. 711-732, 1995.
[12] H. Haario, E. Saksman, and J. Tamminen, "Componentwise Adaptation for High-Dimensional MCMC," *Computational Statistics*, vol. 20, no. 2, pp. 265-274, 2005.
[13] A.T. Horprasert, Y. Yacoob, and L.S. Davis, "Computing 3D Head Orientation from a Monocular Image Sequence," *Proc. Int'l Soc. Optical Eng.*, 1996.
[14] *IEEE Digital Library*, http://www.computer.org/portal/site/csdl/, 2006.
[15] M. Isard and J. MacCormick, "Bramble: A Bayesian Multi-Blob Tracker," *Proc. Eighth Int'l Conf. Computer Vision*, July 2001.
[16] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech," *Proc. Sixth Int'l Conf. Multimodal Interfaces*, 2004.
[17] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-Based Particle Filter for Tracking Multiple Interacting Targets," *Proc. Eighth European Conf. Computer Vision*, May 2004.
[18] Z. Khan, T. Balch, and F. Dellaert, "MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1805-1819, 2005.
[19] V. Kruger, S. Bruns, and G. Sommer, "Efficient Head-Pose Estimation with Gabor Wavelet Networks," *Proc. 11th British Machine Vision Conf.*, Sept. 2000.
[20] J. MacCormick and A. Blake, "A Probabilistic Exclusion Principle for Tracking Multiple Objects," *Proc. Seventh Int'l Conf. Computer Vision*, Sept. 1999.
[21] L. Marcenaro, L. Marchesotti, and C. Regazzoni, "Tracking and Counting Multiple Interacting People in Indoor Scenes," *Proc. Third Performance Evaluation of Tracking and Surveillance*, June 2002.
[22] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior Recognition Based on Head-Pose and Gaze Direction Measurement," *Proc. Int'l Conf. Intelligent Robots and Systems*, 2002.
[23] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe, "A Boosted Particle Filter: Multi-Target Detection and Tracking," *Proc. Eighth European Conf. Computer Vision*, May 2004.
[24] 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," *Proc. Seventh Int'l Conf. Multimodal Interfaces*, Oct. 2005.
[25] A.E.C. Pece, "From Cluster Tracking to People Counting," *Proc. Third Performance Evaluation of Tracking and Surveillance Workshop*, June 2002.
[26] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-Based Probabilistic Tracking," *Proc. Seventh European Conf. Computer Vision*, May 2002.
[27] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Readings in Speech Recognition*, vol. 3, no. 53A, pp. 267-296, 1990.

[28] R. Rae and H. Ritter, "Recognition of Human Head Orientation Based on Artificial Neural Networks," *IEEE Trans. Neural Networks,* vol. 9, no. 2, pp. 257-265, 1998.
[29] K. Smith, "Bayesian Methods for Visual Multi-Object Tracking with Applications to Human Activity Recognition," PhD dissertation, École Polytechnique Fédérale de Lausanne (EPFL), 2007.
[30] K. Smith, S. Ba, D. Gatica-Perez, and J.M. Odobez, "Tracking the Multi-Person Wandering Visual Focus of Attention," *Proc. Seventh Int'l Conf. Multimodal Interfaces,* Nov. 2006.
[31] K. Smith, D. Gatica-Perez, S. Ba, and J.M. Odobez, "Evaluating Multi-Object Tracking," *Proc. CVPR Workshop Empirical Evaluation Methods in Computer Vision,* June 2005.
[32] K. Smith, D. Gatica-Perez, and J.M. Odobez, "Using Particles to Track Varying Numbers of Objects," *Proc. Computer Vision and Pattern Recognition,* June 2005.
[33] K. Smith, P. Quelhas, and D. Gatica-Perez, "Detecting Abandoned Luggage Items in a Public Space," *Proc. Ninth Workshop Performance Evaluation of Tracking and Surveillance,* June 2006.
[34] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining Driver Visual Attention with One Camera," *IEEE Trans. Intelligent Transportation Systems,* vol. 4, no. 4, pp. 205-218, 2004.
[35] C. Stauffer and E. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. Computer Vision and Pattern Recognition,* June 1999.
[36] R. Stiefelhagen, "Tracking Focus of Attention in Meetings," *Proc. Fourth IEEE Conf. Multimodal Interfaces,* 2002.
[37] R. Stiefelhagen, M. Finke, and A. Waibel, "A Model-Based Gaze Tracking System," *Proc. Int'l Joint Symp. Intelligence and Systems,* 1996.
[38] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From Gaze to Focus of Attention," *Visual Information and Information Systems,* pp. 761-768, 1999.
[39] H. Tao, H. Sawhney, and R. Kumar, "A Sampling Algorithm for Detection and Tracking Multiple Objects," *Proc. ICCV Workshop Vision Algorithms,* Sept. 1999.
[40] K. Toyama and A. Blake, "Probabilistic Tracking in a Metric Space," *Proc. Eighth IEEE Int'l Conf. Computer Vision,* 2001.
[41] Y. Wu and K. Toyama, "Wide Range Illumination Insensitive Head Orientation Estimation," *Proc. Automatic Face and Gesture Recognition,* Apr. 2001.
[42] R. Yang and Z. Zhang, "Model-Based Head-Pose Tracking with Stereo Vision," Technical Report MSR-TR-2001-102, Microsoft Research, 2001.
[43] L. Zhao, G. Pingali, and I. Carlbom, "Real-Time Head Orientation Estimation Using Neural Networks," *Proc. Int'l Conf. Image Processing,* Sept. 2002.
[44] T. Zhao and R. Nevatia, "Tracking Multiple Humans in Crowded Environment," *Proc. Computer Vision and Pattern Recognition,* June 2004.

Kevin Smith received the BS degree in mechanical engineering and the MS degree in electrical engineering from the University of Illinois, Urbana-Champaign, in 1998 and 2002, respectively, and the PhD degree in computer science from the École Polytechnique Fédérale de Lausanne (EPFL) in 2007 while working at the IDIAP Research Institute. He is currently a postdoctoral researcher with the Computer Vision Laboratory (CVLab), EPFL. His interests include statistical machine learning and problems in image processing, tracking, segmentation, and medical imaging. He is a member of the IEEE.

Sileye O. Ba received the MS degree in applied mathematics and signal processing from Dakar University in 2000, the MS degree in mathematics, computer vision, and machine learning from the École Normale Supérieure (ENS), Cachan, Paris, in 2002, and the PhD degree from the École Polytechnique Fédérale de Lausanne (EPFL) in 2007 while working as a research assistant at the IDIAP Research Institute. His PhD dissertation covered vision-based sequential Monte Carlo methods for head-pose tracking and visual focus of attention (VFOA) recognition from video sequences. He is currently a postdoctoral researcher at IDIAP, working on using audiovisual features for VFOA and addressees recognition of meeting participants.

Jean-Marc Odobez received the bachelor's degree from the École Nationale Supérieure de Télécommunications de Bretagn (ENSTBr) in 1990 and the PhD degree in signal processing and telecommunications from Rennes University, France, in 1994. He performed his PhD research in the Institut de Recherche en Informatique et Systèmes Aléatoires/Institut National de Recherche en Informatique et en Automatique (IRISA/INRIA), Rennes, on dynamic scene analysis (image stabilization, object detection and tracking, image sequence coding) using statistical models (robust estimation, 2D statistical labeling with Markov Random Field). He then spent one year as a postdoctoral fellow at the General Robotics, Automation, Sensing and Perception (GRASP) Laboratory, University of Pennsylvania, working on visually guided navigation problems. From 1996 to September 2001, he was an associate professor at the Université du Maine, France. He his currently a senior researcher at the IDIAP Research Institute, working on the development of statistical methods for multimodal dynamic scene analysis, including media analysis and human tracking and activity analysis. He is the author or coauthor of more than 60 reviewed papers. He is a member of the IEEE.

Daniel Gatica-Perez received the BS degree in electronic engineering from the University of Puebla, Mexico, in 1993, the MS degree in electrical engineering from the National University of Mexico in 1996, and the PhD degree in electrical engineering from the University of Washington, Seattle, in 2001. He joined the IDIAP Research Institute in 2002, where he is currently a senior researcher. He is currently an associate editor of the *IEEE Transactions on Multimedia*. His research interests include multimedia signal processing and information retrieval, computer vision, and statistical machine learning applied to these domains. He is a member of the IEEE.