IDIAP RESEARCH REPORT

TRACKING THE MULTI PERSON WANDERING VISUAL FOCUS OF ATTENTION

Kevin C. Smith [a] Sileye O. Ba [a]
Daniel Gatica-Perez [a] Jean-Marc odobez [a]

IDIAP–RR 05-80

AUGUST 2006

TO APPEAR IN
International Conference on Multimodal Interfaces (ICMI'06)

[a] IDIAP Research Institute

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

IDIAP Research Report 05-80

TRACKING THE MULTI PERSON WANDERING VISUAL FOCUS OF ATTENTION

Kevin C. Smith   Sileye O. Ba   Daniel Gatica-Perez   Jean-Marc odobez

AUGUST 2006

TO APPEAR IN
International Conference on Multimodal Interfaces (ICMI'06)

Abstract. Estimating the *wandering visual focus of attention* (WVFOA) for multiple people is an important problem with many applications in human behavior understanding. One such application, addressed in this paper, monitors the attention of passers-by to outdoor advertisements. To solve the WVFOA problem, we propose a multi-person tracking approach based on a hybrid Dynamic Bayesian Network that simultaneously infers the number of people in the scene, their body and head locations, and their head pose, in a joint state-space formulation that is amenable for person interaction modeling. The model exploits both global measurements and individual observations for the VFOA. For inference in the resulting high-dimensional state-space, we propose a trans-dimensional Markov Chain Monte Carlo (MCMC) sampling scheme, which not only handles a varying number of people, but also efficiently searches the state-space by allowing person-part state updates. Our model was rigorously evaluated for tracking and its ability to recognize when people look at an outdoor advertisement using a realistic data set.

1 Introduction

An advertising firm has been asked to produce an outdoor display ad campaign for use in shopping malls and bus stations. Internally, the firm has developed several competing designs, one of which must be chosen to present to the client. Is there some way to judge the best placement and content of outdoor advertisements? Currently, the outdoor advertising industry relies on recall surveys or traffic studies to measure the effectiveness of advertisements [19, 20]. However, these approaches are often too impractical or expensive to be commercially viable, and a tool that automatically measures the effectiveness of outdoor printed advertisements, such as television's Nielsen ratings system (which estimates television programs viewing based on a selected set of people's self reports) does not exist. A Nielsen-like system for outdoor display advertisements must *determine the number of people who have actually viewed the ad as a percentage of the total number of people exposed to it*. In this application, the tasks are to *automatically detect and track a varying number of people exposed to the advertisement, and estimate their visual focus of attention* (VFOA) to determine whether they looked at the ad. We have coined the term *Wandering VFOA* to describe this type of problem. It is also relevant for other areas including human-computer interaction, robot-human interaction, and surveillance.

The advertising literature contains a significant amount of work on determining VFOA from eye gaze [3, 12]. However, people in such studies are typically subject to constrained conditions (e.g. they must place their chin on a chin-rest and remain stationary as advertisements are placed in front of them), which renders these approaches useless for measuring public reaction in a real-life outdoor setting. On the other hand, while non-intrusive computer vision algorithms could determine eye gaze using high resolution head images (e.g. [15]), a wide field-of-view is required to detect FOA in an outdoor advertisement scenario where people are free to enter, leave, and move about an outdoor space freely.

In this paper, we present a probabilistic framework for estimating WVFOA for multiple people. Our paper contains three key contributions. First, we propose a principled solution to the problem via a mixed-state Dynamic Bayesian Network that jointly represents the number of people in the scene, their body and head locations, their interactions, and their WVFOA, in a true multi-person state-space formulation. Secondly, we present a method to do inference in the proposed model by trans-dimensional Markov Chain Monte Carlo (MCMC) sampling techniques. Finally, we apply our framework to an outdoor advertisement application to gather useful statistics such as the number of viewers, duration of viewing, and the total number of people exposed to the advertisement. This application, to our knowledge, has not been addressed previously. We rigorously evaluate our approach using realistic data and a detailed set of objective performance measures.

The remainder of the paper is organized as follows. Related work is discussed in Section 2. We present our model in Section 3. We describe how to model WVFOA in Section 4. We objectively evaluate our model on a video data set depicting people passing an outdoor advertisement in Section 5 and provides concluding remarks in Section 6.

2 Related Work

To our knowledge, our work is the first attempt to tackle the problem of wandering visual focus of attention for multiple people. However, some related problems have been studied. The 2002 workshop on Performance and Evaluation of Tracking Systems (PETS) defined a number of estimation tasks on data depicting people passing in front of a shop window, including 1) the number of people in the scene, 2) the number of people in front of the window, and 3) the number of people looking at the window [11]. Other research has studied detection and tracking of shopping groups in a store, and estimation of transaction time [6]. However, in these works, attempts were made to estimate VFOA from body motion *only*. Body motion alone does not contain enough information to accurately determine VFOA. Although there is little related work on the specific problem we address, a large body of research has been conducted on the *separate* issues of multi-person tracking, head pose tracking, and VFOA estimation.

Solving the multi-person tracking problem is a well studied topic, and many researchers have adopted a rigorous Bayesian joint state-space formulation to the problem using particle filtering (PF) techniques [7, 9, 14]. However, sampling on a joint state-space quickly becomes inefficient as the space dimension increases when people are added. Recent work has concentrated on using MCMC sampling to track multiple people more efficiently [9, 14, 24]. The model in [9] tracked a fixed number of interacting people using MCMC. In [14] this model was extended to handle varying numbers of people via reversible-jump MCMC. In this paper, we significantly extend the model of [14] by handling a more complex state-space which requires the non-trivial design of new jumps and proposal distributions (see Section 3.4).

There are two general approaches to solving the head pose tracking problem. The first one independently solves the head tracking and pose estimation problems: a head is first localized and then processed for pose estimation [1, 22]. Speed is the main advantage of this approach, as head pose needs to be estimated from a single location. However, as head pose estimation is very sensitive to head localization [1], the second approach, which jointly tracks a head and estimates its pose, can overall improve performance [16].

Previous work on automatic eye gaze detection, which defines VFOA includes [15], where the VFOA of a driver is determined from eye gaze as the driver's pupils are tracked from a high-resolution monocular video. Because the nature of WVFOA restrict us to lower resolutions, we follow previous works which have shown that VFOA can be reasonably approximated by head pose [18]. However, most existing work has been limited to situations with restricted head motion. In [18], the task was to estimate VFOA of a single person sitting in a meeting room from his head pose. In other situations with less restricted motion, modeling VFOA is more complex. Seminar room environments are such an example. In a recent work, head pose tracking was extended to tracking the head pose of a single person (the lecturer) on low resolution image using multi-view camera setup [21]. As an alternative to the above techniques, face detectors, such as described in [8], that are able to estimate face locations in images together with head pose could be used. However, such systems cannot be applied to solve the multi-people WVFOA estimation problem because they don't keep track of people's identities. To our knowledge, only the work by Otsuka *et al* (2005) deals with multiple people for VFOA estimation, where the number of people is known and fixed, and the problem of tracking is ignored as head pose tracking is obtained with a sensor.

Our approach presents a principled Bayesian solution for a problem which has not yet been addressed in literature, namely tracking the WVFOA for a varying number of interacting people using visual tracking techniques. The task involves the joint estimation of the number of people in a scene, the body and head locations, and head pose for each of them. This is a difficult problem as the size of the state-space (which consists of head and body location parameters and head pose parameters) can be quite large and changes dimensions as people enter and exit the scene.

3 Our Approach

In a Bayesian approach, tracking can be seen as the estimation of the filtering distribution of a state $\mathbf{X}_t$ given a sequence of observations $\mathbf{Z}_t$, $p(\mathbf{X}_t|\mathbf{Z}_{1:t})$. In our model, the state is a joint multi-person configuration and the observations consist of information extracted from a monocular image sequence $\mathbf{Z}_{1:t} = (\mathbf{Z}_1, ..., \mathbf{Z}_t)$. The filtering distribution is recursively computed by

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

where $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is a dynamic model governing the predictive temporal evolution of the state, $p(\mathbf{Z}_t|\mathbf{X}_t)$ is the observation likelihood (measuring how the predictions fit the observations), and C is a normalization constant.

Under the assumption that the posterior $p(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})$ can be approximated by a set of unweighted particles $\{\mathbf{X}_t^{(n)}$,
$n = 1, ..., N\}$ (where $\mathbf{X}_t^{(n)}$ denotes the n-th sample) the Monte Carlo approximation of Eq. 1 becomes

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

The filtering distribution of Eq. 2 can be inferred using MCMC sampling as outlined in Section 3.4.

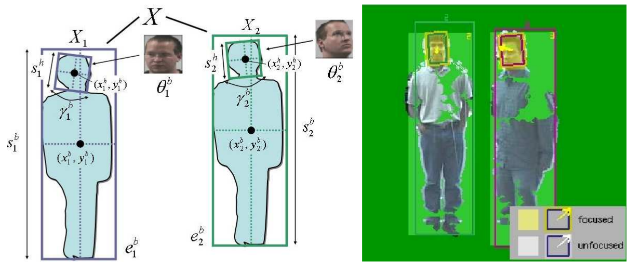

Figure 1: Left: The multi-person state for two people is defined by bounding boxes for the body and head, their related parameters, and the head pose; Right: foreground pixels are segmented using background subtraction.

3.1 State Model for Varying Numbers of People

The dimension of the state vector must be able to vary along with the number of people in the scene. The state at time t can contain from zero to an arbitrary number of people, and is defined by $\mathbf{X}_t = \{X_{i,t} | i \in \mathcal{I}_t\}$, where $\mathcal{I}_t$ is the set of person indexes, $m_t = |\mathcal{I}_t|$ denotes the number of people and $|\cdot|$ indicates set cardinality. The special zero-person case is denoted by $\mathbf{X}_t = \emptyset$.

The state of a single person contains a body and a head component, and is denoted by $\mathbf{X}_{i,t} = (\mathbf{X}^b_{i,t}, \mathbf{X}^h_{i,t})$. The body state vector is $\mathbf{X}^b = (x^b, y^b, s^b, e^b)$ where x^b, y^b is the 2D location of the body in the image, s^b is the height scale factor, and e^b is the eccentricity defined by the ratio of the width over the height. The head state vector is similarly defined as $\mathbf{X}^h = (L^h, \theta^h)$ where $L^h = (x^h, y^h, s^h, e^h, \gamma^h)$ denotes the 2D spatial configuration of the head, including the in-plane rotation γ^h, while θ^h is a discrete variable representing the head pose exemplar accounting for out-of-plane rotation head appearance changes (see Figure 1).

3.2 Multi-Person Dynamics and Interaction

Our dynamic model for a variable number of people is

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) \propto \prod_{i \in \mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) p_0(\mathbf{X}_t) \tag{3}$$

$$\stackrel{def}{=} p_V(\mathbf{X}_t|\mathbf{X}_{t-1}) p_0(\mathbf{X}_t), \tag{4}$$

where p_V is the predictive distribution and
$p_0(\mathbf{X}_t) = p_{0_1}(\mathbf{X}_t) p_{0_2}(\mathbf{X}_t)$ is a prior on the multi-person state configuration including interactions between different people (p_{0_1}) and between a body and its head (p_{0_2}). Following [9, 14, 24], we define p_V as

$$p_V(\mathbf{X}_t|\mathbf{X}_{t-1}) = \prod_{i \in \mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1}) \tag{5}$$

when $\mathbf{X}_{t-1} \neq \emptyset$ and constant otherwise. Additionally, we define $p(\mathbf{X}_{i,t}|\mathbf{X}_{t-1})$ as either the single-person dynamics
$p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1})$ if person i existed in the previous frame, or as a distribution $p_{init}(\mathbf{X}_{i,t})$ over potential initial person birth positions otherwise. The single person dynamic is given by

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

where the dynamics of the body state $\mathbf{X}^b_i$, the head spatial state component L^h_i, and the head-pose exemplars θ^h_i are modeled as 2nd order auto-regressive (AR) processes (a discrete version is exploited for θ^h).

As in [9], the interaction model $p_{0_1}(\mathbf{X}_t)$ prevents two trackers from fitting the same person. This is achieved by exploiting a pairwise Markov Random Field (MRF) whose graph nodes are defined at each time-step by the people, and the links by the set $\mathcal{C}$ of pairs of proximate people. By defining an appropriate potential function

$\phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \propto \exp(-g(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}))$, the interaction model $p_{0_1}(\mathbf{X}_t) = \prod_{ij \in \mathcal{C}} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t})$ enforces constraints in the dynamic model of people based on the locations of the person's neighbors. The interaction potential between two people is defined by a penalty function, g, which is based on the overlap of the people (it is zero when they do not overlap, and increases as the area of overlap increases).

Unlike previous work [9, 14, 24], we propose to exploit a prior model on individual configurations, defined as $p_{0_2}(\mathbf{X}_t) = \prod_{k \in \mathcal{I}_t} p(L^h_{k,t}|\mathbf{X}^b_{k,t})$. This term ensures that the head and body spatial states are physically plausible, constraining the head location w.r.t. the current body configuration.

With these terms defined, the Monte Carlo approximation of the filtering distribution (Eq. 2) is re-expressed as

$$p(\mathbf{X}_t|\mathbf{Z}_{1:t}) \approx C^{-1} p(\mathbf{Z}_t|\mathbf{X}_t) \prod_{ij \in \mathcal{C}} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \times \prod_{k \in \mathcal{I}_t} p(L^h_{k,t}|\mathbf{X}^b_{k,t}) \sum_n p_V(\mathbf{X}_t|\mathbf{X}^{(n)}_{t-1}). \quad (7)$$

3.3 Observation Model

The observation model combines five features to estimate the likelihood of a proposed configuration. The first two are global body features. They consist of *binary* and *color* measurements, and are defined pixel-wise over the entire image. The binary measurements ($\mathbf{Z}^{bin}_t$) make use of a background-subtracted image, while color measurements ($\mathbf{Z}^{col}_t$) exploit histograms in Hue-Saturation (HS) color space. The remaining three features are head features, and consist of texture $\mathbf{Z}^{tex}_t$, skin $\mathbf{Z}^{sk}_t$, and silhouette $\mathbf{Z}^{sil}_t$ observations gathered independently for each person and contribute to the localization and estimation of the head pose. For the remainder of this section, the time index (t) has been omitted to simplify notation. Assuming conditional independence of observations, the overall likelihood is then given by

$$p(\mathbf{Z}|\mathbf{X}) = p(\mathbf{Z}^{col}|\mathbf{Z}^{bin}, \mathbf{X}) p(\mathbf{Z}^{bin}|\mathbf{X}) \left[\prod_{i \in \mathcal{I}} p(\mathbf{Z}^h_i|\mathbf{X}_i) \right]^{\frac{1}{m}},$$

with the individual head likelihood given by

$$p(\mathbf{Z}^h_i|\mathbf{X}_i) = p(\mathbf{Z}^{tex}_i|\mathbf{X}_i) p(\mathbf{Z}^{sk}_i|\mathbf{X}_i) p(\mathbf{Z}^{sil}_i|\mathbf{X}_i) \quad (8)$$

The normalization factor $\frac{1}{m}$ is used to make the head likelihood values comparable for different number of people. All likelihood models are detailed in the next subsections.

3.3.1 Body Model

Binary. Following [14] and using the adaptive background subtraction technique described in [17], each image is segmented into foreground ($\mathbf{Z}^{bin,F}$) and background ($\mathbf{Z}^{bin,B}$) pixels-wise observations (see Figure 1). Qualitatively, for a given multi-person configuration and foreground segmented image, the binary feature computes the distance between the observed overlap (between the area of the multi-object configuration $S^{\mathbf{X}}$ and the segmented image) and a learned value. The overlap is measured for foreground and background in terms of precision ν and recall ρ : $\nu^F = \frac{S^{\mathbf{X}} \cap F}{S^{\mathbf{X}}}$, $\rho^F = \frac{S^{\mathbf{X}} \cap F}{F}$, $\nu^B = \frac{S^{\mathbf{X}} \cap B}{S^{\mathbf{X}}}$, and $\rho^B = \frac{S^{\mathbf{X}} \cap B}{B}$ where F and B are the sets of foreground and background segmented pixels, respectively [14]. Incorrect locations or numbers of people will not match the learned values well, and will result in lower likelihood values. The likelihood is defined for the foreground and background as

$$p(\mathbf{Z}^{bin}|\mathbf{X}) = p(\mathbf{Z}^{bin,F}|\mathbf{X}) p(\mathbf{Z}^{bin,B}|\mathbf{X}). \quad (9)$$

The binary foreground likelihood term, $p(\mathbf{Z}^{bin,F}|\mathbf{X})$, is defined similarly for all non-zero person counts $m \neq 0$ as a single Gaussian distribution set in precision-recall space (ν^F,ρ^F). The binary background likelihood term, on the other hand, is defined as a set of Gaussian Mixture Models (GMMs) learned for each possible person count ($m \in \mathcal{M}$). If the state hypothesizes that two objects are present in the scene, for example, the binary background likelihood term is the GMM density of the the observed ν^B and ρ^B values from the GMM learned for $m = 2$.

Body Color. To maintain personal identities, we employ a HS color feature defined using color observations computed over foreground ($\mathbf{Z}^{col,F}$) and background ($\mathbf{Z}^{col,B}$) pixels. Assuming conditional independence between foreground and background, the color likelihood is written

$$\begin{aligned} p(\mathbf{Z}^{col}|\mathbf{Z}^{bin},\mathbf{X}) = \; & p(\mathbf{Z}^{col,F}|\mathbf{Z}^{bin,F},\mathbf{X}) \times \\ & p(\mathbf{Z}^{col,B}|\mathbf{Z}^{bin,B},\mathbf{X}). \end{aligned} \tag{10}$$

The first term determines how well the color of each measured person matches online learned models, and the second term determines how well the background matches a background model learned off-line. The foreground color likelihood makes use of a 4D histogram defined over a person index, spatial segment, and HS color space, built from an adaptive foreground color model composed of 2D HS color histograms for each person, spatially segmented for the head, torso, and legs. A similar 4D histogram is computed from the color foreground observations. The likelihood is defined using the Bhattacharya distance d_F between the learned and observed histograms $p(\mathbf{Z}^{col,F}|\mathbf{Z}^{bin,F},\mathbf{X}) \propto e^{\lambda_F d_F^2}$, where λ_F is a hyper-parameter [2]. Finally, the background color likelihood helps reject configurations with untracked people and is computed using the background pixels not appearing in $S^{\mathbf{X}}$.

3.3.2 Head Model

The head feature relies on head-pose dependent observation models defined over texture and skin measurements, as previously proposed in [16, 22]. In addition, we propose a novel term: a silhouette head feature defined using the background subtraction, which proved to be of great assistance for head localization in practice.

The head pose can be represented by the pan α^h, tilt β^h, and roll γ^h angles of the Euler decomposition of the head rotation w.r.t. the camera frame. However, as γ^h models in-plane rotation, out-of-plane head appearance changes only depend on the pan and tilt angles. To model these appearance changes we have constructed head pose models for each of the 93 discrete head poses $\theta^h \in \Theta = \{\theta^h_j = (\alpha^h_j, \beta^h_j), j = 1, ..., 93\}$ of the Prima-Pointing Database [5].

Head Pose Texture Model. Head pose texture is represented by the output of three filters: a Gaussian filter at coarse scale and two isotropic Gabor filters at two different scales. Training head patch images, resized to the same reference size (64×64), were preprocessed by histogram equalization to reduce light variation effects. The filter outputs at the locations of a subsampled grid are then concatenated into a single feature vector. Then, for each head pose θ ($\theta = \theta^h$ here, for simplicity), the mean $e^\theta = (e^\theta_j)$ and diagonal covariance matrix $\sigma_\theta = (\sigma^\theta_j)$ of the corresponding training feature vectors are computed and used to define the person texture likelihood model in Eq.8 as

$$p(\mathbf{Z}^{tex}_i|\mathbf{X}_i) = \prod_j \frac{1}{\sigma^{\theta_i}_j} \max(\exp -\frac{1}{2}\left(\frac{\mathbf{Z}^{tex}_{i,j} - e^{\theta_i}_j}{\sigma^{\theta_i}_j}\right)^2, T_{tex}), \tag{11}$$

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

Head Pose Skin Model. To make our head models more robust to background clutter we define a skin binary mask denoted by M^θ for each pose, θ. The masks M^θ are learned from skin masks extracted from the training images corresponding to pose θ by classifying pixels as skin or non-skin, using a Gaussian skin-color distribution modeled in the normalized RG space. The skin color likelihood of a measurement $\mathbf{Z}^{sk}_i$ belonging to the head of person i is defined as

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

where $||.||_1$ denotes the L_1 norm and λ_{sk} is a hyper parameter learned on training data. The measurement $\mathbf{Z}^{sk}_i$ is extracted from the location of person i by detecting skin pixels using a temporally adapted skin color distribution model.

Silhouette. In addition to the pose dependent head model, we propose to add a head silhouette likelihood model to take advantage of background subtraction information. The silhouette model, H^{sil} (see Figure 2), is constructed by averaging head silhouette patches extracted from binary foreground images resulting from background subtraction in the training set. The likelihood of a measured silhouette patch is then defined as:

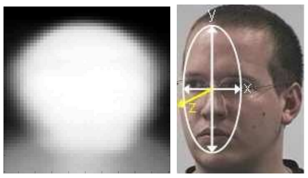

Figure 2: *Head Modeling*. Left: The head silhouette model. Right: the head pointing vector z^h.

$$p(\mathbf{Z}_i^{sil}|\mathbf{X}_i) \propto \exp -\lambda_{sil}||\mathbf{Z}_i^{sil} - H^{sil}||_1, \tag{13}$$

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

3.4 Inference with Trans-Dimensional MCMC

The state vector for a single person in our model is ten-dimensional. Inference on a state-space this large is taxing for traditional particle filters. When allowing for an arbitrary number of people, it becomes clear that an alternative solution is necessary. To solve the inference issue in such high dimensional state-space, we have adopted the Reversible-Jump MCMC (RJMCMC) sampling scheme proposed by several authors [14, 24] to efficiently sample over the posterior distribution. RJMCMC sampling has shown superior performance to a standard Sequential Importance Resampling PF for high dimensional spaces. However, unlike previous work [14, 24], where update moves were applied to the entire state of a single person, we propose to generalize the MCMC approach to update individual components of the state of a single person.

Inferring a solution to the tracking problem in RJMCMC is accomplished by constructing an Markov Chain, the stationary distribution of which is equal to that of the filtering distribution defined in Eq. 7. The Markov Chain is defined over a variable dimensional space to accommodate the varying number of people, and is sampled according to the Metropolis-Hastings (MH) algorithm. Starting from an arbitrary configuration, Metropolis-Hastings repetitively samples a new configuration $\mathbf{X}^*$ from a proposal distribution $q(\mathbf{X}^*|\mathbf{X})$, and adds the proposed sample to the Markov Chain with probability

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)q(\mathbf{X}|\mathbf{X}^*)}{p(\mathbf{X})q(\mathbf{X}^*|\mathbf{X})}\right). \tag{14}$$

Otherwise, a sample constructed from the current configuration is added to the Markov Chain with probability $1-\alpha$. In practice, the new configuration is chosen by first selecting a *move type*, υ^* from a set of moves Υ with prior probability p_{υ^*}. The acceptance ratio α can be re-expressed through *dimension-matching* [4] as

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)p_\upsilon q_\upsilon(\mathbf{X})}{p(\mathbf{X})p_{\upsilon^*}q_{\upsilon^*}(\mathbf{X}^*)}\right), \tag{15}$$

where q_{υ^*} is a move-specific distribution and q_υ is its reverse-move counterpart.

We define six different move types in our model: *birth*, *death*, *swap*, *body update*, *head update*, and *pose update*. A move can either change the dimensionality of the state (as in birth or death moves) or keep it fixed (as in the case of swap and update moves). Once the move type has been determined, a proposal configuration $\mathbf{X}^*$ is sampled from a move-specific proposal distribution $q_{\upsilon^*}(\mathbf{X}^*)$, the likelihood of the proposed configuration is evaluated, the acceptance ratio is computed, and the proposed sample is either added to the Markov Chain (if it passes the acceptance test) or discarded (in which case, the previous configuration $\mathbf{X}$ is added to the Markov Chain).

For the first three move types
(1) Birth of a new person, implying a dimension change from m_t to $m_t + 1$,
(2) Death of an existing person, implying a dimension decrease, from m_t to $m_t - 1$, and
(3) Swap of the identifiers of two existing people, implying no change in dimension,
the details for computing the acceptance ratios and move-specific proposal distributions are described in [14].

However, in [14], a single update move was defined in which *all* the parameters of a randomly selected person were updated simultaneously. Instead, we propose to split the person state space and define several update moves (body update, head update, and pose update). This is done for two reasons. First, the state for a single person in our model is much more complex, and splitting the update moves allows us to separate the problem of finding a good configuration for an entire person into three smaller problems: finding a good configuration for the body, finding a good configuration for the head location, and finding a good configuration for the head pose (the body and head likelihoods are defined in such a way that they are conditionally independent). Secondly, splitting the update moves helps us to avoid the *likelihood balancing* problem, which can arise when one of the components of the likelihood dominates the others. This can result in well-tracked bodies, but poorly estimated head pose, for example. Following the dynamic decomposition for a person into body, head, and pose (Eq. 6), we propose to employ the following update moves (see [14] for the appropriate methodology to define the move proposal),

(4) Body update involves defining the proposal as $q_{\upsilon^*}(\mathbf{X}^*) = \sum_i \frac{1}{m_t} q_{u,b}(\mathbf{X}^*|i)$ with $q_{u,b}(\mathbf{X}^*|i) =$

$$\frac{1}{N}\sum_n p(\mathbf{X}_{i,t}^{b,*}|\mathbf{X}_{t-1}^{(n)})p(\overline{\mathbf{X}_{i,t}^{b,*}}|\mathbf{X}_{t-1}^{(n)})\delta(\overline{\mathbf{X}_{i,t}^{b,*}} - \overline{\mathbf{X}_{i,t}^{b}}),$$

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

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

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

$$\min\left(1, \frac{p(\mathbf{Z}_{i^*,t}^h|\mathbf{X}_{i^*,t}^{h,*})p(L_{i^*,t}^{h,*}|\mathbf{X}_{i^*,t}^{b,*})}{p(\mathbf{Z}_{i^*,t}^h|\mathbf{X}_{i,t}^{h})p(L_{i,t}^{h}|\mathbf{X}_{i,t}^{b})}\right). \tag{16}$$

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

$$\min\left(1, \frac{p(\mathbf{Z}_{i^*,t}^h|\mathbf{X}_{i^*,t}^{h,*})}{p(\mathbf{Z}_{i^*,t}^h|\mathbf{X}_{i,t}^{h})}\right). \tag{17}$$

4 WVFOA Modeling

For our application, the WVFOA of a visible person is defined as being in one of two states: *focused* (she/he is looking at the advertisement) or *unfocused* (she/he is not). As seen in Figure 3, passing people focus their attention on the advertisement from different locations with a variety of different head poses. To infer the WVFOA at each time step for each person in the scene, we rely on the head location and pose estimates provided by the MCMC filter, which track and maintain identity of people over time, even through occlusion. Simply applying a face detector to solve the WVFOA problem for multiple people will fail for several reasons: (1) the range of head poses is beyond that of a typical face detectors, (2) existing state-of-the-art face detectors such as that described in [8] have no mechanism to maintain identity between time steps.

The WVFOA is determined by extracting the pointing vector z^h from the pose estimate (see Fig. 2), which is characterized by the pan and tilt angles, as well as the horizontal head position x^h (see Figure 3). As the ranges of z^h corresponding to the *focused* state are directly dependent on the location of the head in the image, we modeled the likelihood of a *focused* state as

$$p(z^h) = \sum_{k=1}^{K} p(x^h \in I_k, z^h) = \sum_{k=1}^{K} p(x^h \in I_k)p(z^h|x^h \in I_k). \tag{18}$$

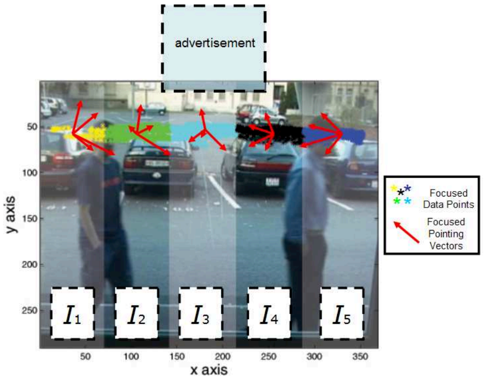

Figure 3: *WVFOA Modeling.* WVFOA is determined by head pose and horizontal position in the image. The horizontal axis is divided into 5 regions. Yellow, green, cyan, black, and blue data points represent *focused* training head locations in each region. At the center of each region red arrows represent 4 focused pointing vectors samples.

Figure 4: *Experimental Setup.* Left: The advertisement poster. Center: Inside the building, a camera is aimed at the window. Right: Outside, the advertisement in the window is noticeable

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

The two terms of the WVFOA model in Equation 18 are defined as followed. The image horizontal axis (x axis) is divided into K intervals, I_k, whose centers and width are denoted by x_{I_k} and σ_{I_k}, respectively. The probability of a location x^h to belong to interval I_k is modeled by a Gaussian distribution $p(x^h \in I_k) = \mathcal{N}(x^h, x_{I_k}, \sigma_{I_k})$. Then, in each interval I_k, the focused pointing vector distribution $p(z^h|I_k)$ is modeled with a Gaussian distribution.

The parameters of the WVFOA model (Gaussian mean and covariance matrix) are learned from the training data. Though our WVFOA model does not make use of the vertical head location, it is straightforward to generalize the models we propose by defining the set I_k to be head location areas in the image plane instead of x-axis intervals.

Finally, a person is determined to be to be *focused* when his/her likelihood $p(z^h)$ is greater than a threshold, T. WVFOA model parameters, including T, were set on the training data to achieve the highest WVFOA event recognition performance (see next Section).

5 Evaluation

As described in the introduction, we applied our model to a hypothetical Nielsen-like outdoor advertisement application. The task is to determine the number of people who actually look at an advertisement as a percentage of the total number of people exposed to it. Additionally, we estimate the time-of-attention (TOA), the number of times people looked, and provide a performance evaluation of tracking quality.

5.1 Data and Experimental Protocol

An experiment was set up as seen in Fig. 4. A fake advertisement was placed in an exposed window with a camera set behind. The camera view can be seen in Figures 5 and 6, with the bottom edge of the poster appearing at the top of the image above the heads of the subjects. A series of recordings were made of actors passing outdoors in front of the window over a period of 10 minutes (actors were used due to privacy concerns). The subjects were allowed to look at the advertisement on their own accord. The data consists of up to three simultaneous people in the scene, and includes several difficult tracking events such as people crossing paths and occluding each other. Though simulated, we believe the data to be a fair representation of a real-life scenario.

The data was organized into a training and test set of equal size. The test set consists of nine sequences, a through i, of approximately 10-second length each. Sequences a-c contain three people each (appearing one at a time) passing in front of the window. Sequences d-h contains two people appearing simultaneously. Sequence i contains three people appearing simultaneously.

The training set was manually annotated for body location, head location, and focused/unfocused state. Using this data, we learned the parameters for the background subtraction model, and the likelihood parameters for binary features, head silhouette features, and skin color distribution. The training set was also used to learn prior sizes for the body and head, and the WVFOA distribution.

An objective evaluation was performed on our model over 180 experiments (20 runs per sequence). The number of samples used was chosen such that there was a sufficient number for good quality tracking according to the number of people in the scene (300 for one person, 600 for two people, and 800 for three). A discussion on the effect of varying the number of samples is presented in section 5.4. The evaluation is separated into two parts: tracking performance and advertisement application performance, discussed in the next two sections.

5.2 Tracking Performance

In this work, we used a set of multi-person tracking measures recently proposed by Smith et al. in [13], and adopted its notation. These measures evaluate a trackers ability to estimate the number and placement of people in the scene (*configuration*), and to persistently track a particular person over time (*identification*) by comparing it with a hand-labeled ground truth (refer to [13] for details). The *F-measure*, which combines the overlap measures recall ρ and precision ν, ($F = \frac{2\nu\rho}{\nu+\rho}$), is used to evaluate the quality of tracking of both the body and the head, as F is only high when both ρ and ν are high ($F = 1$ indicates perfect tracking). An example of the evolution of the tracking errors and F over time can be seen in the lower two panes of Figure 5. Measures $\overline{FP}$ and $\overline{FN}$ give a rate of *False Positive* and *False Negative* errors per person over the course of its lifetime. $\overline{CD}$ (*Configuration Distance*) measures how close the estimated number of people is to the actual number per person per frame. $\overline{FIT}$ and $\overline{FIO}$ count the number of *Falsely-Identified Trackers* and *Falsely-Identified Objects* per person per frame. *Tracker Purity* $\overline{TP}$ and *Object Purity* $\overline{OP}$ estimate the degree of consistency with which the estimates and ground truths were properly identified.

Tracking results are presented in Table 1. From this Table, it is clear from the F measure for both the body and the head that our model performed with a high quality of tracking which was stable across the entire data set. On average, for a given person, our model generated a FP error on 1.8% of its lifetime. This is mostly due to short delays removing a tracker when a person leaves the scene. Similarly, FN errors occur on average on 1.0% of an person's lifetime due to delayed initializations and early deaths.

On average, an person was falsely identified (FIO) 1.6% of its lifetime and a tracker was falsely identified (FIT) 5.7% of its lifetime. These rates indicate that our model was able to maintain person identity through

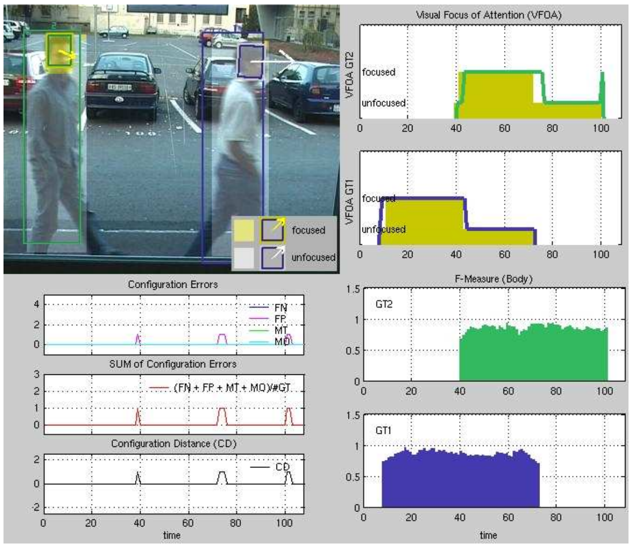

Figure 5: *Tracking Results*. Upper Left: A frame of test data with tracking results and ground truth overlayed. Tracking results appear as green and blue boxes around the head and body with a pointing vector projection, and the ground truth appears as shaded boxes. Both the ground truth and tracking results appear yellow when person is looking at the ad, and gray when unfocused. Upper Right: WVFOA results for both people over the duration of the test sequence. The ground truth appears as yellow bars (raised indicates a *focused* state, lowered when *unfocused*, and non-existent when the object does not appear in the scene), the tracking results appear as blue and green lines. Lower Left: The top plot contains a history of individual tracking errors, the middle plot contains a summation over all the errors, the bottom plot shows CD. Lower Right: F measures quality of tracking for each person. Video results are available at http://www.idiap.ch/$\sim$smith/.

occlusion, as illustrated in Fig. 6. The FIT's are mostly due to a person leaving the scene followed a person entering from the same place in a short period of time, which caused the model to believe it was the same person (in sequences a and b). FIT and FIO errors in sequence h are due to short misidentification caused by occlusion. $\overline{TP}$ and $\overline{OP}$ were both high in most cases. On average, a given person was correctly identified for 97% of his/her lifetime, a given tracker correctly identified its person for 89% of its lifetime. The problems with $\overline{TP}$ in sequences a and b were caused by the situation described previously.

5.3 Advertisement Application Performance

To evaluate the performance of the advertisement application, we compared the results from our model with a true gaze-based ground truth which was hand-labeled at each frame for the state of the WVFOA (either *focused* or *unfocused*). To evaluate our model's performance, we evaluated the following quantities (the results of which appear in Table 2): (1) the number of people present in the scene, (2) the number of people who looked (*focused*) at the advertisement, (3) the number of times someone looked (*focused*) at the advertisement (look-events defined by at least 3 consecutive focused frames), and (4) the frame-based and event-based recognition rates of *focus* on the advertisement.

Over the entire set, 22 people passed the advertisement, of which 20 *focused* on the advertisement. Our system, on average over all runs, estimated that 22 people passed (std $= .17$), of which 21 looked (std $= .09$).

A look-event is defined as *focused* state for a continuous period of at least 3 frames. The total number of look-events in the data set was 22, 21 of which our system recognized on average (std = .89). This result was determined through a symbol matching technique. However, our model estimated 37 total look-events on average (std $= 1.1$). This disparity can be attributed to problems in head pose estimation for heads partially (or fully) outside the image as people enter or leave the scene (in the upper right-hand pane of Fig. 5 this situation occurred for the green estimate near frame 100). Look-event estimation would improve if we did not consider

Table 1: Tracking Results. F body and F head indicate the tightness of the bounding boxes (1 is a perfect fit to the ground truth, 0 is no overlap with the ground truth). $\overline{FP}$ and $\overline{FN}$ are the rates of false positives and false negatives per person, per frame. $\overline{CD}$ measures the difference between the estimated number of people and the actual number of people, per person, per frame. For other measures, please refer to the text and [13].

seq	F	F	Configuration			Identification			
#	Body	Head	$\overline{FP}$	$\overline{FN}$	$\overline{CD}$	$\overline{FIT}$	$\overline{FIO}$	$\overline{TP}$	$\overline{OP}$
a	.88	.88	.021	.001	.022	.262	0	.67	1.0
b	.87	.86	.028	.006	.034	.139	0	.79	.99
c	.87	.85	.012	.005	.016	.026	0	.86	.99
d	.86	.84	.049	0	.049	0	0	.92	1.0
e	.88	.87	.037	.001	.038	0	0	.97	1.0
f	.86	.87	0	.048	.048	0	0	.96	1.0
g	.86	.89	.009	.017	.025	0	0	.99	.98
h	.82	.86	0	.016	.016	.09	.15	.86	.78
i	.88	.87	.008	.003	.011	0	0	.96	.99
avg	.86	.86	.0018	.011	.028	.057	.016	.88	.97

Table 2: Ad application results. *# people* indicates the number of people present in the scene. *# peop. looked* indicates the number of people who looked (*focused*) at the advertisement. *# look-events* refers to the total number of times someone looked at the advertisement. *WVFOA F* is the recognition rate of focused/unfocused events. GT refers to the ground truth and EST refers to the system estimation. REC refers to the # of recognized events.

seq	length (s)	# people		# peop. looked		# look events			WVFOA (F)	
		GT	EST	GT	EST	GT	EST	REC	event	frame
a	15	3	2.5	2	2.9	2	4.7	2.0	.56	.77
b	13	3	3.0	3	3.0	3	4.0	2.9	.83	.88
c	10	3	2.9	3	3.0	3	4.4	3.0	.86	.79
d	5	2	2.0	2	2.0	2	3.3	2.0	.79	.79
e	6	2	2.0	2	2.0	3	3.1	2.1	.78	.71
f	4	2	2.0	2	2.0	2	2.4	2.0	.93	.93
g	4	2	2.0	1	1.1	1	2.9	1.0	.75	.35
h	4	2	2.8	2	2.0	2	4.9	2.0	.65	.69
i	11	3	3.0	3	3.0	4	7.6	4	.72	.87

WVFOA when a person appears at the edge of the scene.

To evaluate the overall quality of WVFOA estimation, we compute a recognition rate as an F measure (defined in Section 5.2) for event-based and frame-based WVFOA. To compute the event-based F, the ground truth and estimated WVFOA are segmented over the entire sequence into focused and unfocused events, symbol matching is performed, and F is computed on matched segments. The overall event-based F is 0.76 (std = .13). The frame-based F is computed by matching the estimated WVFOA for each frame to the ground truth. The overall frame-based F-measure is 0.76 (std $=$ .06). Poor frame-based F results in sequence g occurred because the subject only looked for a very short period of time (0.3s) as he entered the image during which time his head was partially outside of the image. However, our model still managed to detect this event with $F = .75$.

Finally, we also computed the time-of-attention (TOA) as the total amount of time that people spent looking at the advertisement. The total TOA over the entire data set is 37.2s. Our system estimated the TOA to be 44.5s on average (std $=$ .55s). Over-estimation can be attributed to false alarms as people enter and leave the scene.

5.4 Varying the Number of Samples

To study the model's dependency on the number of samples, we conducted experiments on sequence i (which contained three people in the scene) varying the number of samples $N = \{50, 100, 200, 600, 800, 1000\}$.

Figure 6: Tracking two people through occlusion.

Table 3: Varying the number of particles for sequence i.

# particles	# EST looks	WVFOA (F) event	frame	(F) body	head	$\overline{FP}$	$\overline{FN}$	$\overline{CD}$
50	12.8	.53	.80	.84	.85	.013	.006	.019
100	10.6	.59	.81	.85	.86	.010	.008	.018
200	9.9	.62	.80	.86	.86	.008	.009	.017
600	7.3	.73	.86	.87	.87	0	.004	.012
800	7.55	.72	.87	.88	.87	.008	.003	.011
1000	7.2	.72	.86	.88	.87	.008	.003	.011

Resulting measures are given in Table 3. For all N, the model correctly estimated the number of people and the number of people who looked. With less samples, the quality of tracking (as measured by F, $\overline{FP}$, $\overline{FN}$, and $\overline{CD}$) suffered. The head tracking and head pose estimation was noticeably shakier with lower numbers of samples, and the WVFOA estimation suffered as well. This is shown by the increasing error in the number of estimated looks, and by the lower frame-based and event-based F. The model stabilized around approximately $N = 600$. The computational complexity was roughly linear to the number of samples, with a cost ranging from < 1 second ($N = 50$) to ≈ 5 seconds ($N = 600$), non-optimized in Matlab.

6 Conclusion and Future Work

We have presented a principled, probabilistic approach for tracking WVFOA. We evaluated the performance of our model rigorously in the context of a real-world advertising application. The results of our evaluation demonstrates that the model we proposed is able to track a varying number of moving people with good quality and determine their WVFOA. We believe that our model can be used for other applications with similar tasks. For future work, we plan to investigate ways of reducing the occurrence of false look-events when heads appear partially as people enter/exit the scene. We also plan on investigating the usefulness of using a face detector as one of the features in our model. Other future work will include modeling multiple human-to-human interaction using WVFOA or explicitly modeling occlusion, two aspects for which our model could be extended in a principled way.

References

[1] L.M. Brown and Y-L. Tian. Comparative study of coarse head pose estimation. *Work. on Motion and Video Computing*, p125-130, 2002.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, vol 2, p142-149, 2000.

[3] L. Gerald. Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising*, vol 26(1), p61-73, 1997.

[4] P. Green Reversible Jump MCMC Computation and Bayesian Model Determination. *Biometrika*, vol 82, p711-732, 1995.

[5] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. *Work. on Vis. Obs. of Deictic Gestures*, 2004.

[6] I. Haritaoglu and M. Flickner. Detection and tracking of shopping groups in stores. In *CVPR*, p431-438, 2001.

[7] M. Isard and J. MacCormick. BRAMBLE: A Bayesian multi-blob tracker. In *ICCV*, vol 2, p34-41, 2001.

[8] M. J. Jones, and P. Viola Fast Multi-view Face Detection In *CVPR*, 2003

[9] Z. Khan, T. Balch, and F. Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *ECCV*, vol 3024, p279-290, 2004.

[10] K. Otsuka, Y. Takemae,J. Yamato, and H. Murase. A probabilistic inference of multi party-conversation structure based on Markov switching models of gaze patterns, head direction and utterance. In *ACM ICMI*, p191-198, 2005

[11] A. E. C. Pece. From cluster tracking to people counting. In *3rd PETS workshop*, p9-17, 2002.

[12] Rik G.M. Pieters, E. Rosbergen and M. Hartog. Visual attention to advertising: The impact of motivation and repetition. In *Conf. on Adv. in Cons. Research*, vol 23, p242-248, 1995.

[13] K. Smith, D. Gatica-Perez, S. Ba, J.-M. Odobez. Evaluating multi-object tracking. In *CVPR Work. on Empirical Evaluation Methods in Computer Vision*, vol 3, p36, 2005.

[14] K. Smith, D. Gatica-Perez, J.-M. Odobez. Using particles to track varying numbers of objects. In *CVPR*, vol 1, p962-969, 2005.

[15] P. Smith, M. Shah, N. Lobo. Determining driver visual attention with one camera. *Trans. Intel. Syst.*, vol 4(4), p205-218, 2003.

[16] S. Ba, J. Odobez. Evaluation of multiple cues head pose tracking algorithms in indoor environments. In *ICME*, 2005.

[17] C. Stauffer and E. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, vol 2, p246-252, 1999.

[18] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *Visual Information and Information Systems*, p761-768, 1999.

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

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

[21] M. Voit, K. Nickel, and R. Stiefelhagen. Estimating the Lecturer's Head Pose in Seminar Scenarios - A Multi-view Approach. In *MLMI Work.*, p230-240, 2005

[22] Y. Wu, K. Toyama. Wide range illumination insensitive head orientation estimation. In *AFGR* 2001.

[23] T. Yu and Y. Wu. Collaborative tracking of multiple targets. In *CVPR*, vol 1, p834-841, 2004.

[24] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR*, vol 2, p406-413, 2004.