Automatic Detection and Tracking of Human Motion with a View-Based Representation

Ronan Fablet and Michael J. Black

Department of Computer Science
Brown University, Box 1910
Providence, RI02912, USA
{rfablet,black}@cs.brown.edu

Abstract. This paper proposes a solution for the automatic detection and tracking of human motion in image sequences. Due to the complexity of the human body and its motion, automatic detection of 3D human motion remains an open, and important, problem. Existing approaches for automatic detection and tracking focus on 2D cues and typically exploit object appearance (color distribution, shape) or knowledge of a static background. In contrast, we exploit 2D optical flow information which provides rich descriptive cues, while being independent of object and background appearance. To represent the optical flow patterns of people from arbitrary viewpoints, we develop a novel representation of human motion using low-dimensional spatio-temporal models that are learned using motion capture data of human subjects. In addition to human motion (the foreground) we probabilistically model the motion of generic scenes (the background); these statistical models are defined as Gibbsian fields specified from the first-order derivatives of motion observations. Detection and tracking are posed in a principled Bayesian framework which involves the computation of a posterior probability distribution over the model parameters (i.e., the location and the type of the human motion) given a sequence of optical flow observations. Particle filtering is used to represent and predict this non-Gaussian posterior distribution over time. The model parameters of samples from this distribution are related to the pose parameters of a 3D articulated model (e.g. the approximate joint angles and movement direction). Thus the approach proves suitable for initializing more complex probabilistic models of human motion. As shown by experiments on real image sequences, our method is able to detect and track people under different viewpoints with complex backgrounds.

Keywords: Visual motion, motion detection and tracking, human motion analysis, probabilistic models, particle filtering, optical flow.

1 Introduction

The extraction and the tracking of humans in image sequences is a key issue for a variety of application fields, such as, video-surveillance, animation, human-computer interface, and video indexing. The focus of a great deal of research has been the detection and tracking of simple models of humans by exploiting knowledge of skin color or static backgrounds [10,15,22]. Progress has also been made on the problem of accurate 3D

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2350, pp. 476–491, 2002.

tracking of high-dimensional articulated body models given a known initial starting pose [9,11,25]. A significant open issue that limits the applicability of these 3D models is the problem of automatic initialization. Simpler, lower-dimensional, models are needed that can be automatically initialized and provide information about the 3D body parameters. Towards that end, we propose a novel 2D view-based model of human motion based on optical flow that has a number of benefits. First, optical flow provides some insensitivity to variations in illumination, clothing, and background structure. Second, the dimensionality of the model is sufficiently low to permit automatic detection and tracking of people in video sequences. Third, the parameters of the model can be related to the pose of a 3D articulated model and are hence suitable for initializing more complex models. Finally, we develop a probabilistic formulation that permits our motion estimates to be exploited by higher level tracking methods.

The key idea behind our view-based representation is summarized in Figure 1. Motion capture data of actors performing various motions is used to generate many idealized training flow fields from various viewpoints. For each viewpoint, singular value decomposition (SVD) is used to reduce the dimensionality of the training flow fields to give a low-dimensional linear model. Training motions are projected onto this linear basis and temporal models of the linear coefficients for different activities are learned. It is worth noting that there is some psychophysical evidence for the existence of view-based representations of biological motions such as human walking [6,27].

Given this model, the automatic detection and tracking of human motion in image sequences is formulated using a principled Bayesian framework. In addition to the view-based human motion model, we learn a model for the optical flow of general scenes which is used to distinguish human motions from general background motions. Both foreground (person) and background statistical models are defined as Gibbsian fields [12,32] specified from the first-order statistics of motion measurements. Hence, we can exactly evaluate the likelihood of given motion observations w.r.t. learned probabilistic motion models. Therefore, the detection and tracking of human motion can be stated as Bayesian estimation, which involves the evaluation of the posterior distribution of model parameters w.r.t. a sequence of motion observations. For tracking, the prediction

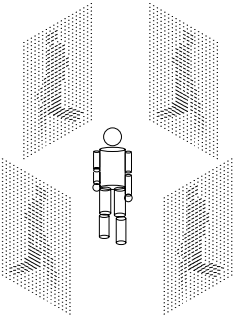

Fig. 1. The motion of a 3D articulated model is projected to derive the 2D image motion (optical flow) of a person from a variety of views. Natural 3D human motions are acquired with a commercial motion-capture system.

in time of this posterior distribution is derived from a prior distribution on the temporal dynamics of model parameters. Since we exploit general non-parametric probabilistic models, the posterior distribution is non-Gaussian and has no straightforward analytic form. Thus, we represent it explicitly using a discrete set of samples in a particle filtering framework [13,16].

2 Problem Statement and Related Work

In this paper, we focus, on the automatic detection and tracking of human motion in image sequences, without the complete recovery of the 3D body motion. While recent advances have been obtained for the tracking of 3D human motion using 2D image cues from monocular image sequences [14,25,30] or multi-view image sequences [4,9,11], these techniques require manual initialization (see [20] for a more complete review). Despite these successes, the complete recovery of 3D body motion is not always necessary and the detection and tracking of 2D human motion is sufficient for numerous applications. Furthermore, this 2D stage can also be regarded as a primary step towards the automatic initialization of more complex 3D schemes.

View-based models for object recognition are not new but here we apply these ideas to biological motion recognition [6,27]. We see these models as existing within a hierarchy from low-level image measurements to 3D motion models. Bregler [5] proposed a similar probabilistic hierarchy of models but the approach lacked powerful mid-level representations of human motion such as those proposed here and hence attempted to interpret at a high level, very low-level motion measurements. There have been other proposed intermediate representations such as the "cardboard" person model [19] and the scaled prismatic model [7] but these proved too high dimensional for automatic initialization.

Current approaches for the detection and the tracking of people in images and videos mainly rely on human appearance analysis and modeling. For instance, pedestrian detection has been achieved using low-resolution wavelet images of people [22] or body shape [10]. In [18], a Bayesian framework is also developed for object localization based on probabilistic modeling of object shape. The most successful of recent tracking methods exploit statistical models of object appearance (color or grey-level histograms [8], mixture models of color distributions [17], background subtraction [15], or edge-based models of object shape [16,29]).

Among all the methods developed for object detection and tracking, Bayesian approaches appear the most attractive, since they provide a principled probabilistic framework to combine multiple cues and to introduce *a priori* knowledge or constraints related to the class of objects to detect and track in the scene. For these statistical schemes, the key point is to provide appropriate statistical characterization of the entities of interest (foreground) and of the background. Recent work on Bayesian tracking has focused on this problem of foreground/background modeling [17,24,26,28].

In this paper, we also consider such a Bayesian approach. Unlike previous work, our main focus is on the definition of appropriate probabilistic models of dynamic information for human motion. As previously mentioned, whereas motion cues provide generic and rich information independent of object appearance, they are rarely exploited for

the detection and tracking of predefined types of objects. Motion information is indeed mainly exploited in motion detection schemes [21,23], when no *a priori* information is available about the class of entities to extract. This is due to the lack of generic probabilistic models of object motion, which could be used as alternatives or complements to statistical modeling of object appearance. However, in the context of human motion analysis, recent studies [2,31] targeted at motion estimation and activity recognition have stressed that human motion examples share specific characteristics, which make the definition and the identification of generic models of human motion feasible.

We further exploit and extend these previous approaches to handle the automatic detection and tracking of human motion in image sequences. Similarly to [2,31], we rely on learned bases of human motion. However, instead of considering only one motion basis set as in [2,31], we use a set of these motion bases. Consequently, our probabilistic modeling can be viewed as a mixture of human motion models. Moreover, unlike these previous approaches, our main concern is to design well-founded probabilistic models of human motion. Instead of assuming particular noise distributions such as Gaussian or some more robust distribution [2], our models are defined as Gibbsian fields [12,32] specified from the first-order statistics of motion measurements, and are directly learned from training examples. These probabilistic models are then used as the foreground motion models in our approach. In the same fashion we construct a statistical background model that accounts for generic motion situations (cf. [17,24,26]). Both models are exploited in the Bayesian framework for detection and tracking. These motion models could be combined with more traditional probabilistic models of appearance in this Bayesian framework. It would be straightforward to extend this work to detect and track other kinds of objects for other applications.

3 Human Motion Modeling

To detect and track human motion in image sequences, we rely on generative models, which are computed from training examples for different view angles using PCA (Principal Component Analysis). What is critical is that these models be sufficiently low-dimensional so as to permit efficient search and sufficiently expressive so as to be useful for initializing more complex models. Given these linear human motion bases, we build probabilistic likelihood models from the statistical analysis of the reconstruction error and of the distribution of the projection onto basis vectors.

In this section, we first present the training stage used to learn human motion bases. Then, the different features of the probabilistic human motion models are introduced.

3.1 Learning Human Motion Bases

The learning of motion bases from training examples has already been successfully exploited for parameterized motion estimation and activity recognition [2,31]. Similarly, we learn bases for full-body human motion from synthetic training examples generated from motion capture data. Here we focus on walking motions but the approach can be extended to more general human motion.

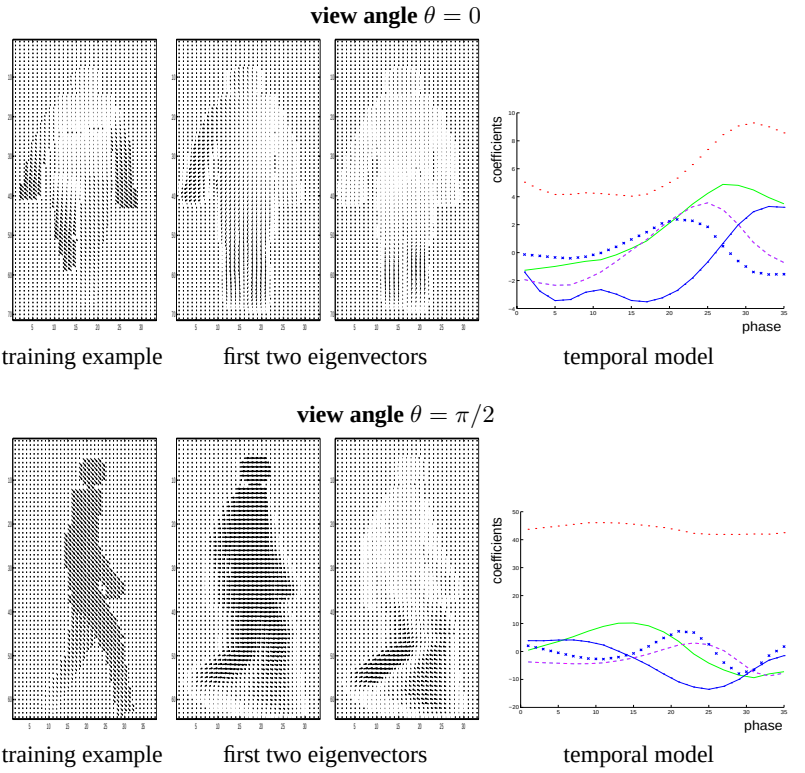

Fig. 2. Learning human motion models

Our training set consists of multiple walking sequences from four professional dancers (two men and two women). Given the 3D position of the body and its motion at any time instant we can predict the actual 2D flow field that this motion would generate from any viewing direction. To our knowledge, complex motion models such as those described here have never been used before because of the lack of reasonable training data. Here however, with 3D "ground truth" we can generate high-quality training flow fields from any desired viewpoint. Fig. 2 displays two example flow fields from the training set. To cope with the changes in optical flow as a function of viewing angle, we adopt a view-based model and separately learn motion bases for different viewing directions. Specifically, we generate a training set of 480 flow fields for each of twelve view angles $\{0, \pi/6, ..., 11\pi/6\}$. For each view, we perform PCA and keep as motion bases the first fifteen eigenvectors accounting on average for 0.95 of the total variance. Fig.2 shows the two first basis vectors for the view angles 0 and $\pi/2$. For a given human motion model $\mathcal{M}$, $\theta(\mathcal{M})$ denotes the associated view angle, $N_B(\mathcal{M})$ the number of basis vectors and $B(\mathcal{M}) = \{B_k(\mathcal{M}), k \in \{1, ..., N_B(\mathcal{M})\}\}$ the eigenvectors for the human motion basis.

These learned motion bases constrain the spatial configuration of the detected and tracked human motion. Additionally we model the temporal characteristics of human

motion. In this work, we use the method described in [1]. However, other kinds of temporal models could be employed (e.g. Hidden Markov Models (HMM), auto-regressive models or other time series models). The exploited temporal model is specified by a sequence $\tau(\mathcal{M}) = [a_1(\mathcal{M}), ..., a_{\phi_{max}}(\mathcal{M})]$, where $a_\phi(\mathcal{M})$ is a vector of linear coefficients at phase ϕ of the cyclic walking gait [2]. Given the motion basis $B(\mathcal{M})$, we learn the temporal model $\tau(\mathcal{M})$ using the method described in [1]. For each model $\mathcal{M}$, we compute the associate trajectories of the motion coefficients of the projection of the associated training examples onto basis $B(\mathcal{M})$ and the mean trajectories form the temporal models. In Fig. 2, the temporal models for $\theta = 0$ and $\theta = \pi/2$ are displayed.[1]

Hence, given a model $\mathcal{M}$, a phase ϕ and a magnitude γ, we can generate an optical flow field $w(\mathcal{M}, \phi, \gamma)$ corresponding to the human motion:

$$w(\mathcal{M}, \phi, \gamma) = \gamma \sum_{k=1}^{N_B(\mathcal{M})} a_{\phi,k}(\mathcal{M}) B_k(\mathcal{M}). \tag{1}$$

3.2 Statistical Modeling of Human Motion

Now, let $w(\mathcal{W})$ be an observed flow field in a window $\mathcal{W}$. Our generative model states that $w(\mathcal{W})$ equals $w(\mathcal{M}, \phi, \gamma)$ plus noise for some setting of the model parameters $\mathcal{M}$, $\mathcal{W}$, the phase ϕ and magnitude γ. Rather than assume an arbitrary noise model (e.g. Gaussian), here we learn it from training data.

We note that we can reduce the dimensionality of the model further by computing the optimal magnitude term γ' that minimizes the reconstruction error

$$E(w(\mathcal{W}), \phi, \mathcal{M}\}) = w(\mathcal{W}) - \gamma' \sum_{k=1}^{N_B(\mathcal{M})} a_{\phi,k}(\mathcal{M}) B_k(\mathcal{M}) \tag{2}$$

where γ' is given by

$$\gamma' = \left[\sum_{k=1}^{N_B(\mathcal{M})} \alpha_k a_{\phi,k}(\mathcal{M}) \right] / \left[\sum_{k=1}^{N_B(\mathcal{M})} \alpha_k^2 \right] \tag{3}$$

where $\{\alpha_k\}$ are the coefficients of the projection of $w(\mathcal{W})$ onto the bases $B(\mathcal{M})$.

The likelihood $P_{HM}(w(\mathcal{W})|\phi, \gamma, \mathcal{M})$ of the flow field $w(\mathcal{W})$ given the model parameters $(\phi, \gamma, \mathcal{M})$ is then specified from the reconstruction error $E(w(\mathcal{W}), \phi, \mathcal{M})$ and magnitude γ' as follows:

$$P_{HM}(w(\mathcal{W})|\phi, \gamma, \mathcal{M}) = P\left(E(w(\mathcal{W}), \phi, \mathcal{M}), \gamma'|\phi, \gamma, \mathcal{M}\right). \tag{4}$$

We can rewrite this as:

$$P_{HM}(w(\mathcal{W})|\phi, \gamma, \mathcal{M}) = P\left(E(w(\mathcal{W}), \phi, \mathcal{M})|\gamma', \phi, \gamma, \mathcal{M}\right) P\left(\gamma'|\phi, \gamma, \mathcal{M}\right). \tag{5}$$

[1] Observe that we have chosen to separately compute the spatial and temporal models for each view. This simplifies the learning and estimation problems compared to building a single spatio-temporal model for each view.

Since the magnitude of the reconstruction error obviously depends on the magnitude of the human motion, the likelihood $P\left(E(w(\mathcal{W}),\phi,\mathcal{M})|\gamma',\phi,\gamma,\mathcal{M}\right)$ is evaluated from the normalized reconstruction error $\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$ defined by:

$$\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}) = E(w(\mathcal{W}),\phi,\mathcal{M})/[\gamma'\|a_\phi(\mathcal{M})\|], \tag{6}$$

as $\gamma'\|a_\phi(\mathcal{M})\|$ is the magnitude of the human motion. Thus, further simplifying conditional dependencies, the likelihood $P_{HM}(w(\mathcal{W})|\phi,\gamma,\mathcal{M})$ is defined as the product of two terms as follows:

$$P_{HM}(w(\mathcal{W})|\phi,\gamma,\mathcal{M}) = P\left(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M}\right) P\left(\gamma'|\phi,\gamma,\mathcal{M}\right). \tag{7}$$

The first term, $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$, represents the likelihood distribution and will be learned from training examples. The second term, $P(\gamma'|\phi,\gamma,\mathcal{M})$, is exploited to specify the minimum motion magnitude of the motion to be detected and tracked and to smooth the temporal evolution of the magnitude γ of the tracked area.

3.3 Likelihood Distribution of the Reconstruction Error

The definition of the likelihood distribution $P\left(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M}\right)$ is based on the first-order statistics of $\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$. Let Λ denote the quantization space of these flow field differences and $\Gamma(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})) = \{\Gamma(\lambda,\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}))\}_{\lambda\in\Lambda}$ the histogram of $\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$ quantized over Λ. The computation of the likelihood $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$ must be independent of the size of the window $\mathcal{W}$ in order to compare the likelihoods of the projection error over a set of windows with different sizes. This leads us to consider the normalized histogram $\overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}))$ as the characteristic statistics of $\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$.

Based on the Maximum Entropy criterion (ME) [32], $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$ is expressed using the following Gibbsian formulation:

$$P\left(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M}\right) \propto \exp\left[\Psi_\mathcal{M} \bullet \overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}))\right], \tag{8}$$

where $\Psi_\mathcal{M} = \{\Psi_\mathcal{M}(\lambda)\}_{\lambda\in\Lambda}$ are the Gibbsian potentials which explicitly specify the distribution $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$. $\Psi_\mathcal{M} \bullet \overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}))$ is the dot product between model potentials $\Psi_\mathcal{M}$ and normalized histogram $\overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$ defined by:

$$\Psi_\mathcal{M} \bullet \overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})) = \sum_{\lambda\in\Lambda} \Psi_\mathcal{M}(\lambda)\overline{\Gamma}(\lambda,\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})). \tag{9}$$

Since we will compare values of the likelihood $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$ for different windows $\mathcal{W}$ and models $\mathcal{M}$, the normalization constant $Z_\mathcal{M}$ defined by:

$$Z_\mathcal{M} = \sum_{(w(\mathcal{W}),\phi)} \exp\left[\Psi_\mathcal{M} \bullet \overline{\Gamma}(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M}))\right], \tag{10}$$

has to be explicitly known and computable. Let us stress that this issue was not handled in [2,31] since only one motion basis was considered. It can be rewritten as:

$$Z_\mathcal{M} = \left(\sum_{\lambda\in\Lambda} exp\left[\frac{\Psi_\mathcal{M}(\lambda)}{|\mathcal{W}|}\right]\right)^{|\mathcal{W}|}. \tag{11}$$

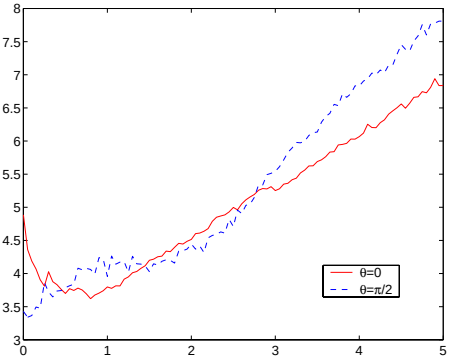

Fig. 3. Potentials $\Psi_{\mathcal{M}}$ specifying the likelihood of the reconstruction error $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$. We give the plots of $\{-\Psi_{\mathcal{M}}(\lambda)\}_{\lambda\in\Lambda}$ for the view angles $\theta = 0$ and $\theta = \pi/2$.

Thus, the exact expression of the likelihood $P(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M})$ is

$$P\left(\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})|\mathcal{M}\right) = \left(\sum_{\lambda\in\Lambda}\exp\left[\frac{\Psi_{\mathcal{M}}(\lambda)}{|\mathcal{W}|}\right]\right)^{|\mathcal{W}|}\exp\left[\Psi_{\mathcal{M}}\bullet\overline{\Gamma}(\widetilde{E}(w,\phi,\mathcal{M}))\right]. \tag{12}$$

In this expression, only the first term depends on $|\mathcal{W}|$ whereas it is by definition independent of the observed motion error $\widetilde{E}(w(\mathcal{W}),\phi,\mathcal{M})$. Therefore, to make the comparison of the likelihoods according to different window sizes feasible, we will compute the expression (12) for a reference window size $|\mathcal{W}|_{ref}$. In practice, we use the window size of the training examples.

We learn the potentials $\Psi_{\mathcal{M}}$ for a model $\mathcal{M}$ from the training examples used to compute the motion basis $B(\mathcal{M})$. More precisely, given the normalized histogram $\overline{\Gamma}(\mathcal{M})$ of the reconstruction error for this training set, the potentials $\Psi_{\mathcal{M}}$ estimated w.r.t. the Maximum Likelihood (ML) criterion are given by:

$$\Psi_{\mathcal{M}}(\lambda) = \log\left(\overline{\Gamma}(\lambda,\mathcal{M})/\sum_{\lambda'\in\Lambda}\overline{\Gamma}(\lambda',\mathcal{M})\right). \tag{13}$$

Fig. 3 displays the plot of the potentials $\Psi_{\mathcal{M}}$ for the view angles $\theta = 0$ and $\theta = \pi/2$. These two distributions are non-Gaussian. Besides, it is also worth mentioning that the main peak does not necessarily occur in 0. Thus, there can be a weak bias in the reconstruction from the learned motion basis.

3.4 Prior Distribution of Magnitude

To further constrain the detection and tracking of human motion, we exploit the fact that we aim at identifying moving entities with a motion magnitude greater than a given motion detection level μ. In addition, the magnitude γ' is more likely to evolve smoothly over time. Therefore, $P(\gamma'|\phi,\gamma,\mathcal{M})$ is written as:

$$P(\gamma'|\phi,\gamma,\mathcal{M}) \propto \delta_\mu(\|w(\mathcal{M},\phi,\gamma')\|)\mathcal{N}(\gamma'-\gamma,\sigma^2_{mag}), \tag{14}$$

where $\|w(\mathcal{M},\phi,\gamma')\|$ is the norm of the reconstructed flow $w(\mathcal{M},\phi,\gamma')$ given by relation (1). $\delta_\mu(.)$ is a smooth step function centered in μ and $\mathcal{N}(.,\sigma^2_{mag})$ a normal distribution with variance σ^2_{mag}.

4 Generic Motion Modeling

In the Bayesian framework described in Section 5, the detection and the tracking of human motion exploits the ratio of the likelihood of the observation within a given window explained, on the one hand, by a human motion model (foreground model) and on the other hand by a generic motion model (background model). Since no ground truth exists for the flow fields of general scenes, we cannot directly derive this model using observed statistics of the flow field w. As an alternative, it is defined from the statistics of temporal image differences. Thus, it allows us to handle noise on a static background but also dynamic situations which do not correspond to human motion.

The probabilistic distribution attached to this model is specified using the first-order statistics of the difference of pairs of successive images. Given a window $\mathcal{W}$ and an image difference ΔI, we evaluate its normalized histogram $\overline{\Gamma}(\Delta I(\mathcal{W})) = \{\overline{\Gamma}(n,\Delta I(\mathcal{W}))\}_{n\in\{-N,\ldots,N\}}$ where N is the number of grey-levels in the images. Similarly to the statistical modeling of the reconstruction error in ubsection 3.3, the likelihood $P_{GM}(\Delta I(\mathcal{W}))$, that the image difference $\Delta I(\mathcal{W})$ within window $\mathcal{W}$ is a sample of the generic motion model, is expressed as:

$$P_{GM}(\Delta I(\mathcal{W})) \propto \exp\left[\Psi_{GM} \bullet \overline{\Gamma}(\Delta I\mathcal{W})\right], \tag{15}$$

$$\text{with } \Psi_{GM} \bullet \overline{\Gamma}(\Delta I(\mathcal{W})) = \sum_{n=-N}^{N} \Psi_{GM}(n)\overline{\Gamma}(n,\Delta I(\mathcal{W})). \tag{16}$$

We cope with the normalization issue in a similar way as in Subsection 3.3. To estimate the potentials Ψ_{GM}, we consider a set of image sequences acquired with a static camera, involving different kinds of moving objects (pedestrians, cars, trees) and backgrounds. The normalized histogram $\overline{\Gamma}_{GM}$ of the image differences is evaluated from this set of sequences and the estimation of the potentials Ψ_{GM} w.r.t. the ML criterion leads to:

$$\Psi_{GM}(n) = \log\left(\overline{\Gamma}_{GM}(n) / \sum_{n'=-N}^{N} \overline{\Gamma}_{GM}(n')\right). \tag{17}$$

Fig. 4 displays the plot of the estimated potentials Ψ_{GM} of the generic motion model. While this distribution is obviously non-Gaussian, it is similar in spirit to the robust function as used for robust motion estimation [3].

5 Bayesian Formulation

The detection and the tracking of human motion is stated as a Bayesian inference problem. More precisely, given a sequence of observations, i.e. a sequence of observed flow

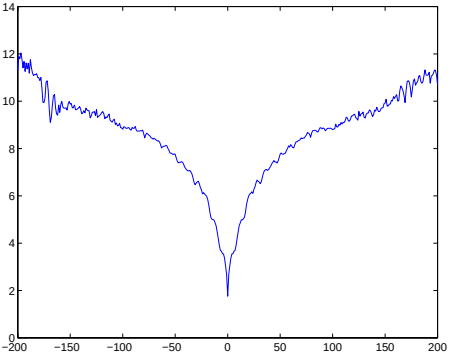

Fig. 4. Plot of the potentials Ψ_{GM} of the generic motion model.

fields and image differences, we aim at evaluating the posterior distribution of the model parameters which are in our case the location (i.e., the window $\mathcal{W}$) and type of human motion (i.e., model $\mathcal{M}$, phase ϕ and magnitude γ of the sought human motion sample).

In this Section, we detail this Bayesian formulation which exploits the statistical motion models, that we have previously defined, in a data-driven likelihood. We will define prior distribution over model parameters appropriate for detection and tracking. Then, we will briefly outline how we evaluate in practice the posterior distribution using particle filtering.

5.1 General Overview

Let us denote by $\overline{w}_t = \{w_0, w_1, ..., w_t\}$ and $\overline{\Delta I}_t = \{\Delta I_0, \Delta I_1, ..., \Delta I_t\}$ the sequences of flow fields and image differences up to time t. The flow fields $\{w_t\}$ are estimated using the robust technique described in [3].

The goal of detecting and tracking human motion at time t is regarded as the evaluation of the posterior distribution $P(\phi_t, \gamma_t, \mathcal{W}_t, \mathcal{M}_t | \overline{w}_t, \overline{\Delta I}_t)$. Below, we will denote by Θ_t the model parameters $[\phi_t, \gamma_t, \mathcal{W}_t, \mathcal{M}_t]$. Using Bayes rule and assuming that observations at time t are independent from observations at previous instants given model parameters Θ_t, we obtain:

$$P(\Theta_t | \overline{w}_t, \overline{\Delta I}_t) = k \underbrace{P(w_t, \Delta I_t | \Theta_t)}_{\text{data-driven likelihood}} \underbrace{P(\Theta_t | \overline{w}_{t-1}, \overline{\Delta I}_{t-1})}_{\text{prior at time } t-1}, \tag{18}$$

where k is a constant independent of Θ_t.

5.2 Data-Driven Likelihood

The data-driven distribution $P(w_t, \Delta I_t | \Theta_t)$ evaluates the likelihood that the observations at time t account for human motion model $\mathcal{M}_t$ within the window $\mathcal{W}_t$. Assuming the motion characteristics within the window $\mathcal{W}_t$ are independent on those of the background $\mathcal{R} \backslash \mathcal{W}_t$, where $\mathcal{R}$ is the image support, $P(w_t, \Delta I_t | \Theta_t)$ is explicitly given by:

$$P(w_t, \Delta I_t | \Theta_t) = k' P_{HM}(w_t(\mathcal{W}_t) | \phi_t, \gamma_t, \mathcal{M}_t) P_{GM}(\Delta I_t(\mathcal{R} \backslash \mathcal{W}_t) | \mathcal{W}_t), \tag{19}$$

where k' is a normalization factor. Exploiting the independence between $\mathcal{R}\backslash\mathcal{W}_t$ and $\mathcal{W}_t$, $P_{GM}(\Delta I_t(\mathcal{R}\backslash\mathcal{W}_t)|\mathcal{W}_t)$ can be rewritten as the ratio of $P_{GM}(\Delta I_t(\mathcal{R})|\mathcal{W}_t)$ and $P_{GM}(\Delta I_t(\mathcal{W}_t)|\mathcal{W}_t)$. Further simplifying conditional dependencies, we obtain:

$$P(w_t, \Delta I_t|\Theta_t) = k' P_{HM}(w_t(\mathcal{W}_t)|\phi_t, \gamma_t, \mathcal{M}_t) \frac{P_{GM}(\Delta I_t(\mathcal{R}))}{P_{GM}(\Delta I_t(\mathcal{W}_t))}, \tag{20}$$

Since $P_{GM}(\Delta I_t(\mathcal{R}))$ does not depend on model parameters Θ_t, this simplifies into the following expression:

$$P(w_t, \Delta I_t|\Theta_t) = k'' \frac{P_{HM}(w_t(\mathcal{W}_t)|\phi_t, \gamma_t, \mathcal{M}_t)}{P_{GM}(\Delta I_t(\mathcal{W}_t))}, \tag{21}$$

where k'' is a normalization factor.

Thus, the data-driven term $P(w_t, \Delta I_t|\phi_t, \gamma_t, \mathcal{W}_t, \mathcal{M}_t)$ is completely determined from the expression of the likelihoods $P_{HM}(w_t(\mathcal{W}_t)|\phi_t, \gamma_t, \mathcal{M}_t)$ and $P_{GM}(\Delta I_t(\mathcal{W}_t))$ given by relations (12), (14) and (15).

5.3 Prior Distribution on Model Parameters

The prior distribution $P(\Theta_t|\overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ describes the temporal dynamics of the model parameters Θ_t. For the detection task at time $t = 0$, this prior reduces to $P(\Theta_0)$. Since we have neither *a priori* knowledge about the location of the human motion in the image nor the human motion type, the initial prior distribution is chosen to be uniform.

For tracking purpose, the prior distribution $P(\Theta_t|\overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ is expressed as the marginalization of the joint distribution over all model parameters up to time t over all observations up to time t. Adopting a first-order Markov assumption on model parameters, this leads to the following integral formulation:

$$P(\Theta_t|\overline{w}_{t-1}, \overline{\Delta I}_{t-1}) = \int P(\Theta_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) \underbrace{P(\Theta_{t-1}|\overline{w}_{t-1}, \overline{\Delta I}_{t-1})}_{\text{posterior at time } t-1} d\Theta_{t-1}. \tag{22}$$

This integral involves the product of two terms: $P(\Theta_{t-1}|\overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ the posterior distribution at time $t-1$ and $P(\Theta_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ the prior distribution over model parameters describing their temporal dynamics. Assuming conditional independence of model parameters ϕ_t, γ_t and $\mathcal{M}_t$ w.r.t. to $[\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}]$, the latter term is rewritten as:

$$\begin{aligned} P(\Theta_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) = {} & P(\mathcal{M}_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) \\ & \times P(\phi_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) \\ & \times P(\gamma_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) \\ & \times P(\mathcal{W}_t|\gamma_t, \phi_t, \mathcal{M}_t, \Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}). \end{aligned} \tag{23}$$

$P(\mathcal{M}_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ defines the evolution of the human motion model assigned to the tracked window. It is directly related to the temporal evolution of the view angles

between the tracked entity and the camera. We can assume that $\mathcal{M}_t$ depends only on $\mathcal{M}_{t-1}$. We thus resort to the specification of the first-order Markov chain $P(\mathcal{M}_t|\mathcal{M}_{t-1})$. Assuming the view angle evolves smoothly over time, these transitions are defined by:

$$P(\mathcal{M}_t|\mathcal{M}_{t-1}) = \begin{cases} \alpha, & \text{if } \mathcal{M}_t = \mathcal{M}_{t-1} \\ \dfrac{1-\alpha}{2}, & \text{if } \theta(\mathcal{M}_t) = \theta(\mathcal{M}_{t-1}) \pm \pi/6[2\pi] \end{cases}. \tag{24}$$

Typically, we set in practice $\alpha = 0.7$.

Concerning phase ϕ_t, it can be assumed that it evolves smoothly along time and $P(\phi_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ is taken to be a wrapped Gaussian distribution centered in ϕ_{t-1} modulo the length of the walk cycle. For the magnitude γ_t, we exploit that we have estimated at time $t-1$ the magnitude which leads to the lowest reconstruction error. We then assign this value to γ_t.

The prior distribution $P(\mathcal{W}_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1})$ over the window position $\mathcal{W}_t$ is assumed to be Gaussian around the predicted window $\mathcal{W}_t^{pred}$.

$$P(\mathcal{W}_t|\phi_t, \gamma_t, \mathcal{M}_t, \Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1}) = \mathcal{N}(\mathcal{W}_t - \mathcal{W}_t^{pred}, \sigma_{pos}), \tag{25}$$

where $\mathcal{N}(., \sigma_{pos})$ is a Gaussian distribution with diagonal covariance σ_{pos}. The location of the predicted window $\mathcal{W}_t^{pred}$ is computed from the displacement of the center of the previous window $\mathcal{W}_{t\text{-}1}$ according to the reconstructed flow $w(\mathcal{M}_{t-1}, \gamma_t, \phi_{t-1})$.

5.4 Computation of the Posterior Distribution

The direct computation of the posterior distribution $P(\Theta_t|\overline{w}_t, \overline{\Delta I}_t)$ is not feasible, since no analytic form of this likelihood function over the whole model parameter space can be derived. However, for any values of the model parameters Θ_t, we can evaluate the likelihood of the observations formed by the flow field and the image difference at time t given these model parameter values. Therefore, we can approximate the posterior distribution $P(\Theta_t|\overline{w}_t, \overline{\Delta I}_t)$ by a set of samples using a particle filtering framework [13, 16].

At time t, we first draw N_{part} particles $\{s_n\}$, each one being assigned model parameter values $\Theta_{t-1}^{s_n}$. We propagate this set of particles at time t using the temporal dynamics specified by the prior distributions $P(\Theta_t|\Theta_{t-1}, \overline{w}_{t-1}, \overline{\Delta I}_{t-1})$. This supplies us with a new set of particles $\{s'_n\}$, for which we compute the likelihoods $P(w_t, \Delta I_t|\Theta_t^{s'_n})$ using (21). When normalized to sum to one, these likelihoods (or weights) associated with each particle s'_n approximate the posterior distribution at time t.

At time $t = 0$, for detection purposes, we need to perform a global search over model parameters (i.e., position, motion type, phase and scale). We exploit an hierarchical strategy to approximate the posterior distribution by subsampling at different resolutions the space of the model parameters. This scheme provides a coarse location of the detected human motion, which will be refined by tracking.

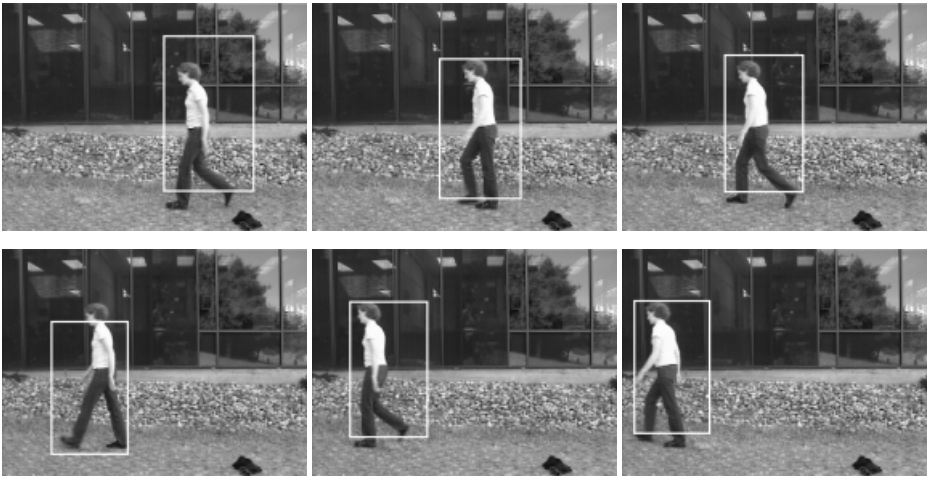

Fig. 5. Tracking a human walking in a straight line. We display the location of the expected window at frames 0, 10, 20, 30, 40 and 50.

6 Experiments

Parameter setting. We present preliminary results of detection and tracking of human motion in different real image sequences acquired with a static camera. In order to visualize the posterior distribution in the frame at time t, we display the expected location $< \mathcal{W}_t | \overline{w_t}, \overline{\Delta I_t} >$ of the detected and tracked window including human motion, which is approximated by the following sum over the set of particles $\{s_n\}_{n \in \{1,\ldots,N_{part}\}}$ at time t: $< \mathcal{W}_t | \overline{w_t}, \overline{\Delta I_t} >= \sum_{n=1}^{N_{part}} \pi^{s_n} \mathcal{W}_t^{s_n}$ where π_{s_n} is the normalized version of the likelihood $P(w_t, \Delta I_t | \Theta_t^{s_n})$.

In the subsequent experiments, we used the following parameter settings. As far the data-driven likelihood is concerned, the main parameter to set is the motion detection level μ. Since this parameter has a physical meaning in terms of average displacement within the expected window comprising the human motion, it is easy to set. We will use $\mu = 1.0$. Besides, the variance of the prediction for the magnitude γ is taken to be $\sigma_{mag}^2 = 1.0$. These parameters could be learned from training data.

For the prior distribution specifying the temporal dynamics, we set $\alpha = 0.7$ for the Markov chain characterizing the transitions between human motion models, and the covariance σ_{pos} has diagonal terms equaling 5.0 for the square root of the variance on the position of the center of the tracked window, and 1.0 for the variance in terms of window scaling.

Human walking in a straight line. The first processed example is an sequence of 60 frames involving a human walking in a straight line. We display in Fig. 5 the results for frames 0, 10, 20, 30, 40 and 50. Our method accurately recovers the window size and the location of the walking pedestrian with no manual initialization. As previously mentioned, the initialization provides a coarse estimate of the location of the human

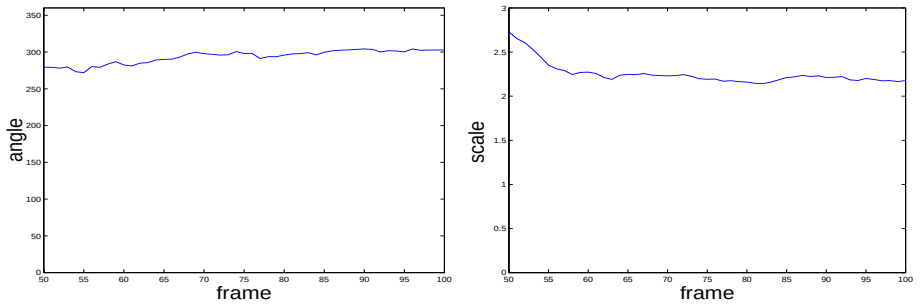

Fig. 6. Tracking a human walking in a straight line. Plot of the estimated values of scale and view angle parameters.

Fig. 7. Tracking a human walking in a straight line. We display the location of the expected window at frames 0, 10, 20, 30, 40 and 50.

motion, since the coarse-to-fine strategy is not iterated until the finest resolution. Also, as shown in Fig. 6, the expected value of the viewing angle stabilizes around $10\pi/6$, whereas one could expect to obtain $3\pi/2$. Even though there is a bias (corresponding to the quantization step of the view angles), it provides us with correct direction of the human motion. This bias might be due to differences in magnitude between the observed human motion and the training examples.

Walking pedestrian in presence of background motion. The second processed example is a video of a street acquired from the top a building. Therefore, it does not exactly refer to the kind of motion situation learned from the training examples. In this sequence, the tree in the upper right corner is slightly moving in the wind. In spite of these difficulties, our approach recovers the location of the walking pedestrian and the expected view angle is estimated to be approximately $8\pi/6$, which gives a correct guess of the direction of the human motion.

Note further, that with the position, scale, viewing direction, and phase of the gate, that we could now predict the 3D configuration of the body (since we knew this during training). Thus our posterior distribution provides the basis for a probabilistic proposal distribution for more detailed tracking. In future work, we will use our 2D models for initialisation of 3D human motion tracking.

7 Conclusion

We have presented a Bayesian framework for the automatic detection and tracking of human motion in image sequences. It relies on the design of probabilistic generative models of human motion learned for training examples. We also define a statistical

model accounting for generic motion situations. These two motion models enable us to define the likelihood of observing a particular example flow field given the parameters of the human motion model. Then, the detection and the tracking of human motion involves the evaluation of the posterior distribution over the model parameters w.r.t. a sequence of motion observations. The computation of this posterior distribution exploits a model of the temporal dynamics of the model parameters and is achieved using a particle filtering framework. We have demonstrated the effectiveness of our methods for different real image sequences comprising human walking.

Future research directions will involve different issues. First of all, the probabilistic human motion models provide complementary tools to appearance modeling usually considered for the detection and tracking of people. The Bayesian framework exploited in our work could be easily extended to combine both appearance and motion models. Additionally, we could enrich the characterization of human motion by learning more complex temporal models of human motion using time series analysis tools such as HMMs or linear and non-linear auto-regressive models. With a more varied training set, we could learn more general models of 2D image motion. Finally, the proposed probabilistic human motion models could also be used to characterize and analyze other categories of dynamic events, not necessarily human related, such as dynamic phenomena occurring in meteorological image sequences.

Acknowledgments. This work was support in part by an INRIA postdoctoral grant, by the DARPA HumanID Project (ONR contract N000140110886) and by a gift from the Xerox Foundation.

We thank Michael Gleicher for providing the 3D motion capture database, Manolis Kamvysselis and Hedvig Sidenbladh for help processing the mocap data and Robert Altschuler for providing video data used to train the generic motion model. Michael J. Black thanks Allen Jepson for discussions on foreground and background models.

References

1. M. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parametrized models of image motion. *CVPR*, pp. 561–567, 1997.
2. M.J. Black. Explaining optical flow events with parametrized spatio-temporal tracking. *CVPR*, pp. 326–332, 1999.
3. M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):74–104, 1996.
4. C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, pp. 8–15, 1998.
5. C. Bregler. Learning and recognizing human dynamics in video sequences. *CVPR*, pp. 568–574, 1997.
6. I. Bülthoff, H.H. Bülthoff, and P. Sinha. A survey on the automatic indexing of video data. *Nature Neuroscience*, 1(3):254–257, 1998.
7. T-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, 1:239–245, 1999.
8. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. , *CVPR*, pp. 142–149, 2000.

9. J. Deutscher, A. Blake, and I. Reid. Articulated motion capture by annealing particle filtering. *CVPR*, pp. 126–133, 2000.
10. D. Gavrila. Pedestrian detection from a moving vehicle. *ECCV*, II, pp.37–49, 2000.
11. D. Gavrila and L. Davis. 3-D model-based tracking of human in action: a multi-view approach. *CVPR*, pp. 73–80, 1996.
12. S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *PAMI*, 6(6):721–741, 1984.
13. N. Gordon, D.J. Salmond, and A.F. Smith. A novel approach to nonlinear/non-Gaussian bayesian estimation. *IEEE Trans. on Radar, Sonar and Navigation*, 140(2):107–113, 1996.
14. Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. *NIPS*, 12, pp. 820–826, 2000.
15. I. Haritaoglu, D. Harwood and L. Davis. A real time system for detecting and tracking people *IVC*, 1999
16. M. Isard and A. Blake. Condensation: conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.
17. M. Isard and J. MacCormick. BraMBLE: a Bayesian multiple-blob tracker. *ICCV*, II, pp. 34–41, 2001.
18. M. Isard J. Sullivan, A. Blake and J. MacCormick. Object localization by bayesian correlation. *ICCV*, pp. 1068–1075, 1999.
19. S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *Int. Conf. on Automatic Face and Gesture Recog.*, pp. 38–44, 1996.
20. T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, to appear.
21. J.M. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. *Video Data Compression for Multimedia Computing*, chapter 8, pp. 295–311. H. H. Li, S. Sun, and H. Derin, eds, Kluwer, 1997.
22. C. Papageorgiou and T. Poggio. Trainable pedestrian detection. *ICCV*, pp. 1223–1228, 1999.
23. N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *PAMI*, 22(3):266–280, 2000.
24. J. Rittscher, J. Kato, S. Joga, A. Blake. A probabilistic background model for tracking. *ECCV*, 2:336-350, 2000.
25. H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *ECCV*, pp. 702–718, 2000.
26. H. Sidenbladh and M. J. Black. Learning image statistics for Bayesian tracking. *ICCV*, II, pp. 709–716, 2001.
27. P. Sinha, H.H. Bülthoff, and I. Bülthoff. View-based recognition of biological motion sequences. *Invest. Ophth. and Vis. Science*, 36(4):1920, 1995.
28. J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. *ECCV*, II pp. 307–323, 2000.
29. K. Toyama and A. Blake. Probabilistic tracking in a metric space. *ICCV*, 2:50–57, 2001.
30. S. Wachter and H. Nagel. Tracking of persons in monocular image sequences. *CVIU*, 74(3):174–192, 1999.
31. Y. Yacoob and M.J. Black. Parametrized modeling and recognition of activities. *CVIU*, 73(2):232–247, 1999.
32. S.C. Zhu, T. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *IJCV*, 27(2):107–126, 1998.