Learning Image Statistics for Bayesian Tracking

Hedvig Sidenbladh
CVAP/NADA
Royal Institute of Technology (KTH)
SE–100 44 Stockholm, Sweden
hedvig@nada.kth.se

Michael J. Black
Dept. of Computer Science, Box 1910
Brown University
Providence, RI 02912, USA
black@cs.brown.edu

Abstract

This paper describes a framework for learning probabilistic models of objects and scenes and for exploiting these models for tracking complex, deformable, or articulated objects in image sequences. We focus on the probabilistic tracking of people and learn models of how they appear and move in images. In particular, we learn the likelihood of observing various spatial and temporal filter responses corresponding to edges, ridges, and motion differences given a model of the person. Similarly, we learn probability distributions over filter responses for general scenes that define a likelihood of observing the filter responses for arbitrary backgrounds. We then derive a probabilistic model for tracking that exploits the ratio between the likelihood that image pixels corresponding to the foreground (person) were generated by an actual person or by some unknown background. The paper extends previous work on learning image statistics and combines it with Bayesian tracking using particle filtering. By combining multiple image cues, and by using learned likelihood models, we demonstrate improved robustness and accuracy when tracking complex objects such as people in monocular image sequences with cluttered scenes and a moving camera.

1 Introduction

This paper extends recent work on learning the statistics of natural images and applies the results to the problem of tracking people in image sequences. We learn probabilistic models of how people appear in images and show how this information can be combined with probabilistic models that capture the statistics of natural scenes [10, 11, 14, 20, 26]. In particular, we learn models that characterize the probability of observing various image filter responses given, for example, we are looking at a human arm at a particular orientation. Filter responses corresponding to edges, ridges, and motion for the different limbs of the body and for generic scenes are considered. We show how these learned models can be combined in a Bayesian framework for tracking complex objects such as people. We employ a particle filtering method [7, 8] and demonstrate its behavior with examples of people tracking in monocular image sequences containing clutter, camera motion and self-occlusion.

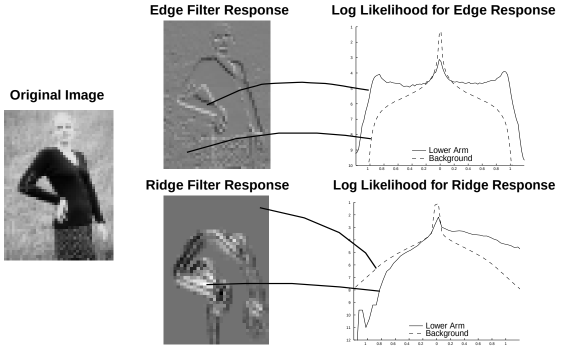

Figure 1: Learning the appearance of people and scenes. Distributions over edge and ridge filter response are learned from examples of human limbs and general scenes.

Reliable tracking requires a general model of human appearance in images that captures the range of variability in appearance and that is somewhat invariant to changes in clothing, lighting, and background. Motivated by [10], probability distributions for various filter responses are constructed. The approach is illustrated in Figure 1. Given a database of images containing people we manually determine the “ground truth” corresponding to limb boundaries and limb axes for the torso, head, upper and lower arms and upper and lower legs. Discrete probability distributions corresponding to edge and ridge filter responses on the marked boundaries and axes respectively are learned from the data.

We also collect ground truth data of limb motions between two frames and learn the distribution of temporal image differences between corresponding pixels. In the same spirit, we could learn probabilistic models of skin color [25] or other texture cues.

The above distributions characterize the appearance of the “foreground” object. For reliable people tracking we must learn the prior distribution of filter responses in general scenes. We show that the likelihood of observing the filter responses for an image is proportional to the ratio between the likelihood that the foreground image pixels are explained by the foreground object and the likelihood that they are explained by some general background (cf. [17]):

$$p(\text{all cues} \mid \text{fgrnd}, \text{bgrnd}) = \mathrm{C}\,\frac{\mathrm{p}(\text{fgrnd cues} \mid \text{fgrnd})}{\mathrm{p}(\text{fgrnd cues} \mid \text{bgrnd})} \;.$$

This ratio is highest when the foreground (person) model projects to an image region that is unlikely to have been generated by some general scene but is well explained by the statistics of people. This ratio also implies that there is no advantage to the foreground model explaining data that is equally well explained as background. It is important to note that the "background model" here is completely general and, unlike the common background subtraction techniques, is not tied to a specific, known, scene.

Using these ideas, we extend previous work on person tracking by combining multiple image cues, by using learned probabilistic models of object appearance, and by taking into account a probabilistic model of general scenes in the above likelihood ratio. Experimental results suggest that a combination of cues provides a rich likelihood model that results in more reliable and computationally efficient tracking than can be achieved with individual cues. We present the results for the 3D tracking of human limbs in monocular image sequences in the presence of clutter, unknown backgrounds, self occlusion, and camera motion.

2 Related Work

Recent work on learning the low-order spatial statistics of natural scenes shows promise for problems in segmentation, graphics, and image compression [6, 10, 11, 14, 20, 26]. Here we extend this analysis in a number of directions. First, most previous work has considered the statistics of filter responses corresponding to first derivatives of the image function. Here we also examine filters corresponding to "ridges" [13] and show that, like edge filters, the distribution of responses is invariant across scale for general scenes. In addition to ridges and edges, motion is an important cue for tracking people. Previous tracking approaches have made simplifying assumptions about the noise in temporal image derivatives. The typical brightness constancy assumption assumes that this noise is Gaussian [21] while the actual distributions learned here for hand-registered training data show that it is actually highly non-Gaussian.

Our modeling of the image statistics of people versus backgrounds is similar in spirit to the work of Konishi et al. [10]. Given images where humans have manually marked what they think of as "edges," Konishi et al. learn a distribution p_{on} corresponding to the probability of a filter response for these edge locations. In our case, we model the filter responses at the boundary of a limb *regardless* of whether an actual edge is visible in the scene or not. An edge may or may not be visible at a limb boundary depending on the clothing and contrast between the limb and the background. Thus we can think of the p_{on} distribution of Konishi et al. [10] as a *generic* feature distribution while here we learn an *object-specific* distribution for people.

Konishi et al. [10] also computed the distribution p_{off} corresponding to the filter responses away from edges and used the log of the likelihood ratio between p_{on} and p_{off} for edge detection. We add additional background models for the statistics of ridges and temporal differences and exploit the ratio between the probability of foreground (person) filter responses and background responses for tracking. Finally, the absolute contrast between foreground and background is less important for detecting people than the orientation of the features (edges or ridges) and hence we perform local contrast normalization prior to filtering.

We exploit the above work on learned image statistics to track people in cluttered scenes with a moving camera. Recent Bayesian probabilistic formulations of the tracking problem [1, 4, 8, 16, 18] use particle filtering methods [7, 8] as we do here. Cham and Rehg [1] use a fixed template to represent the appearance of each limb. Deutscher et al. [4] assume large foreground-background contrast and multiple cameras. In our previous work [18] we used image motion as the cue for Bayesian tracking of 3D human models in monocular sequences. The approach used a robust likelihood model for temporal image differences. While this approach could deal with more complex imaging environments than [1, 4], like all optical flow tracking methods, it was prone to "drift".

Particle filtering methods represent a complex posterior probability distribution with a discrete set of samples. Each sample corresponds to a possible set of model parameters, ϕ, or poses of the body in our case. For each pose we can predict where in the image we expect to see limbs and then check whether the image filter responses support the hypothesis. This is in contrast to tracking approaches that first extract edges and then match the model to them [4, 8].

Reliable tracking requires multiple spatial and temporal image cues. While many systems combine cues such as motion, color, or stereo for person detection and tracking [3, 24], the formulation and combination of these cues is often ad hoc. Additionally, the appearance of people changes in complex ways and previous approaches have used highly simplified noise models. In contrast, the learned models here account for the variation observed in training data. These edge, ridge, and motion models are then combined in a Bayesian framework.

Similar in spirit is the tracking work of Sullivan et al. [22, 23] who model the distributions of filter responses for a general background and a particular foreground where the foreground is represented by a generalized template. Given these, they determine if an image patch is background, foreground, or on the boundary by matching the distribution of filter responses in the patch with a learned mixture model of background and foreground filter responses. Our work differs in several ways: We model the ratio between the likelihoods for model foreground points being foreground and background, rather than evaluating the likelihood for model background and foreground in evenly distributed locations in the image. We use several different filter responses, and we use steerable filters [5] instead of isotropic ones. Furthermore, our objects (human limbs) are, in the general case, too varied in appearance to be modeled by generalized templates.

3 Learning Human and Scene Models

A human is modeled as a 3-dimensional articulated assembly of truncated cones. The model parameters, $\boldsymbol{\phi}$, consist of the relative angles between the limbs (cones) and their angular velocities along with the global position and rotation of the assembly and its translational and angular velocity [18]. In general we may also have background parameters $\boldsymbol{\beta}$ that describe, for example, the affine motion of the background. For this paper we treat the background image structure and motion as unknown and leave the explicit estimation of the background parameters for future work.

The model parameters, $\boldsymbol{\phi}$ determine $\{\mathbf{x}_f\}$, the set of image locations corresponding to the foreground (person). Let the set of background pixels be $\{\mathbf{x}_b\} = \{\mathbf{x}\} - \{\mathbf{x}_f\}$, where $\{\mathbf{x}\}$ is the set of all pixels.[1] Let $p(\mathbf{f} \mid \boldsymbol{\phi})$ be the likelihood of observing filter responses $\mathbf{f}$ given the parameters, $\boldsymbol{\phi}$, of the foreground object. Given appropriately sampled sets $\{\mathbf{x}\}$, $\{\mathbf{x}_b\}$, and $\{\mathbf{x}_f\}$ we treat the filter responses at all pixels as independent and write the likelihood as

$$p(\mathbf{f} \mid \boldsymbol{\phi}) = \prod_{\mathbf{x}\in\{\mathbf{x}_b\}} p_{\text{off}}(\mathbf{f}(\mathbf{x})) \prod_{\mathbf{x}\in\{\mathbf{x}_f\}} p_{\text{on}}(\mathbf{f}(\mathbf{x},\boldsymbol{\phi})) = \frac{\prod_{\mathbf{x}\in\{\mathbf{x}\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))}{\prod_{\mathbf{x}\in\{\mathbf{x}_f\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))} \prod_{\mathbf{x}\in\{\mathbf{x}_f\}} p_{\text{on}}(\mathbf{f}(\mathbf{x},\boldsymbol{\phi}))$$

since $\{\mathbf{x}_b\} = \{\mathbf{x}\} - \{\mathbf{x}_f\}$. p_{off} represents the probability of observing the filter response $\mathbf{f}(\mathbf{x})$ given that pixel $\mathbf{x}$ is in the background, while p_{on} represents the probability of observing $\mathbf{f}(\mathbf{x},\boldsymbol{\phi})$ given that $\mathbf{x}$ is in the foreground and the model parameters are $\boldsymbol{\phi}$.

Note $\prod_{\mathbf{x}\in\{\mathbf{x}\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))$ is independent of $\boldsymbol{\phi}$. We call this constant term κ_1 and simplify the likelihood as

$$p(\mathbf{f} \mid \boldsymbol{\phi}) = \kappa_1 \prod_{\mathbf{x}\in\{\mathbf{x}_f\}} \frac{p_{\text{on}}(\mathbf{f}(\mathbf{x},\boldsymbol{\phi}))}{p_{\text{off}}(\mathbf{f}(\mathbf{x}))} \,. \qquad (1)$$

This is the normalized ratio of the probability that the foreground pixels are explained by the person model versus that they are explained by a generic background model.

The filter responses, $\mathbf{f}$, are computed from a set of filters that are chosen to capture the spatial and temporal appearance of people and natural scenes. In particular, the filter responses include edge responses f_e, ridge responses f_r and the motion responses f_m, so that $\mathbf{f} = [f_e, f_r, f_m]$.

Responses are computed at several different image scales. For this purpose, a Gaussian image pyramid is constructed. The lowest level, 0, is the original image, while pyramid level σ is obtained from $\sigma - 1$ by convolving with a Gaussian filter of standard deviation 1 and subsampling.

We assume that the responses from the different filters are independent for a given pixel location $\mathbf{x}$. Furthermore, the responses for edge and motion at different image levels σ are considered independent.[2] The response for ridges is only observed at one scale, depending on the size of the limb (this is discussed further in section 3.2).

Thus, the likelihood can be written as

$$p(\mathbf{f} \mid \boldsymbol{\phi}) = p(f_e \mid \boldsymbol{\phi})\, p(f_r \mid \boldsymbol{\phi})\, p(f_m \mid \boldsymbol{\phi}) = \kappa_1 \prod_{\sigma=0}^{s} \left(\prod_{\mathbf{x}\in\{\mathbf{x}_e\}} \frac{p^e_{\text{on}}(f_e(\mathbf{x},\boldsymbol{\phi},\sigma))}{p^e_{\text{off}}(f_e(\mathbf{x},\sigma))} \cdot \prod_{\mathbf{x}\in\{\mathbf{x}_m\}} \frac{p^m_{\text{on}}(f_m(\mathbf{x},\boldsymbol{\phi},\sigma))}{p^m_{\text{off}}(f_m(\mathbf{x},\sigma))} \right) \cdot \prod_{\mathbf{x}\in\{\mathbf{x}_r\}} \frac{p^r_{\text{on}}(f_r(\mathbf{x},\boldsymbol{\phi},\sigma(\boldsymbol{\phi}))}{p^r_{\text{off}}(f_r(\mathbf{x},\sigma(\boldsymbol{\phi})))} \qquad (2)$$

where $s = 3$ corresponds to four levels in the pyramid, the edge point set $\{\mathbf{x}_e\} \subseteq \{\mathbf{x}_f\}$ contains sampled pixel locations on the model edges (i.e. on the borders of the limbs), and the motion and ridge point sets $\{\mathbf{x}_m\}$ and $\{\mathbf{x}_r\}$ are equal to $\{\mathbf{x}_f\}$.[3]

The individual likelihood distributions p^z_{on} and p^z_{off} (where $z = e, r, m$) are non-Gaussian and are learned from training data. This training set consists of approximately 150 images and short sequences of people, in which the outline of torso, head, upper and lower arms and legs are marked manually. Examples of marked training images are given in Figure 2. The marked edges serve as ground truth for the learning of edge responses on and off actual limb edges. The area spanned by the two edges is computed from the marked edges an is used for learning of ridge responses on the limbs. The area spanned by the two edges is also warped between consecutive frames in sequences. The distribution of temporal differences between the warped image pairs is then learned.

[1] The spatial and temporal statistics of neighboring pixels are unlikely to be independent [23]. We therefore approximate the set $\{\mathbf{x}_f\}$ with a randomly sampled subset to approximate pixel independence. The number of samples in the foreground is always the same and covers the visible parts of the human model.

[2] This is a very crude assumption as edge responses are highly correlated across scale. Further work needs to be done to model these correlations.

[3] The point sets $\{\mathbf{x}_m\}$ and $\{\mathbf{x}_r\}$ need not be equal to $\{\mathbf{x}_f\}$. For example, it could be beneficial to exclude points near the edges from these sets. Note that the cardinality of these sets defines an implicit weighting of the likelihood terms of each cue.

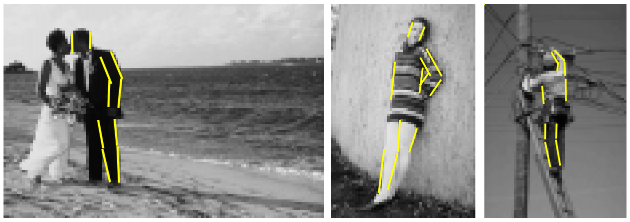

Figure 2: Example images from the training set with limb edges manually marked.

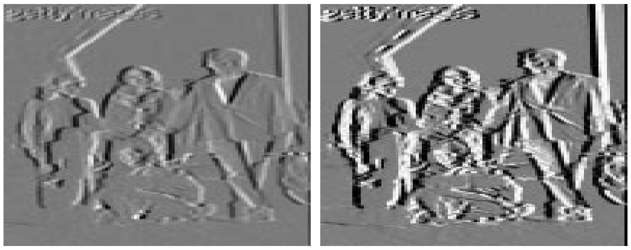

Figure 3: Left: Horizontal gradient image. Right: Contrast-normalized horizontal gradient image, used for learning.

3.1 Edge Cue

The edge response f_e is a function of $[f_x, f_y]$, the first derivatives of the image brightness function in the horizontal and vertical directions. Edges are modeled in terms of these filter responses at the four finest levels $\sigma = 0, 1, 2, 3$ in the image pyramid.

More specifically, the image response for an edge of orientation θ at pyramid level σ is formulated as the image gradient perpendicular to the edge orientation:

$$f_e(\mathbf{x}, \theta, \sigma) = \sin\theta \, f_x(\mathbf{x}, \sigma) - \cos\theta \, f_y(\mathbf{x}, \sigma) \qquad (3)$$

where $f_x(\mathbf{x}, \sigma)$ and $f_y(\mathbf{x}, \sigma)$ are the image gradients at pyramid level σ and image position $\mathbf{x}$.

For our purposes, the most interesting property of an edge or a ridge is not its absolute contrast, but rather the scale and orientation of the feature. Therefore, before computing image derivatives at a location $\mathbf{x} = [x, y]$, we perform local contrast normalization using a hyperbolic tangent nonlinearity [19]. Regions with high contrast will be normalized to a contrast of 1, while areas of low contrast are normalized to contrast of 0 (Figure 3). The resulting filter responses then depend more on orientation than on contrast.

Our experiments indicate that the edge response is independent of scale [19]. We therefore build a scale-independent empirical edge likelihood distribution using filter responses from multiple scales.

Foreground. For each of the images in the training set (Figure 2), the edge orientation θ_l of each limb l is computed from the manually marked edges. For all pyramid levels, locations $\mathbf{x}$ are sampled on the marked edge, with $\theta = \theta_l$. For each limb, l, a separate histogram is constructed, of steered edge responses $f_e(\mathbf{x}, \theta_l, \sigma)$ [5], for the sampled foreground edge locations. The normalized histogram for each limb l represents $p^e_{\mathrm{on}}(f_e \mid l)$, the probability function of edge response f_e conditioned on limb l, given that the limb projects to an actual limb. Given an observed response $f_e(\mathbf{x}, \theta_l, \sigma)$, the likelihood of observing this response in the foreground (on limb l) is $p^e_{\mathrm{on}}(f_e(\mathbf{x}, \theta_l, \sigma) \mid l)$.

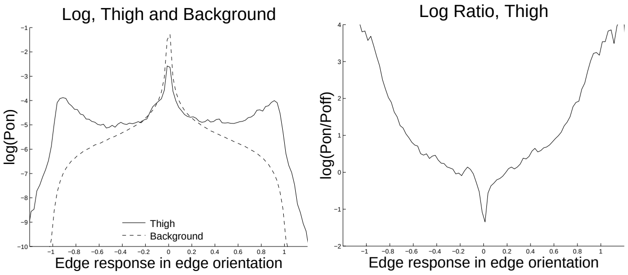

Figure 4: Edge filter responses. Left: Log likelihood for background, $p^e_{\mathrm{off}}(f_e)$, and thigh, $p^e_{\mathrm{on}}(f_e \mid thigh)$. Right: Log ratio between distributions, $\log(\frac{p^e_{\mathrm{on}}(f_e \mid thigh)}{p^e_{\mathrm{off}}(f_e)})$.

The log likelihood, $\log(p^e_{\mathrm{on}})$, for the thigh is shown in Figure 4 (similar distributions are obtained for the other limbs).

Background. The background edge distribution is learned from a large set of general images with and without people. A normalized histogram of responses $f_e(\mathbf{x}, \theta, \sigma)$ is created by sampling image locations $\mathbf{x}$ and orientations θ uniformly at all pyramid levels σ. This gives $p^e_{\mathrm{off}}(f_e)$, the probability distribution over edge responses, given that we look at locations and orientations that *do not correspond to* the edges of human limbs. According to this distribution, the likelihood of observing a certain edge response $f_e(\mathbf{x}, \theta, \sigma)$ in the background is $p^e_{\mathrm{off}}(f_e(\mathbf{x}, \theta, \sigma))$. Figure 4 shows the logarithm of this distribution.

The background is more likely than the limb edge to have low contrast and hence the background distribution has a higher peak near 0. Large filter responses are also less likely than for limbs, which means that the background distribution has very low values when the response approaches 1. Therefore, if a large filter response is observed, the corresponding image location will have higher probability of originating from a limb boundary, and the log ratio (Figure 4) between foreground and background likelihood will be larger than 0. If the response is low, the probability of the pixel belonging to the background is high, and the ratio will be smaller than 0. Note that the distributions here have different shapes than those learned by others [10, 11, 26] due, in part, to the effects of contrast normalization.

3.2 Ridge Cue

In the same spirit as with edge cues, we use the second derivatives of the image in the direction of the model ridge

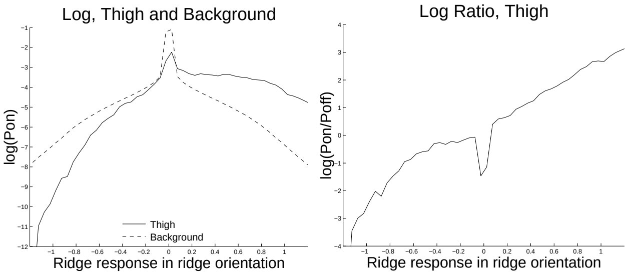

Figure 5: Ridge filter responses. Left: Log likelihoods for background, $p^r_{\text{off}}(f_r)$, and thigh, $p^r_{\text{on}}(f_r \mid thigh)$. Right: Log ratio between distributions, $\log(\frac{p^r_{\text{on}}(f_r \mid thigh)}{p^r_{\text{off}}(f_r)})$.

for comparing real-world and model ridges. The filter response used is f_r, a function of $[f_{xx}, f_{xy}, f_{yy}]$, the second derivatives of the brightness function in the horizontal and vertical directions.

A ridge is an elongated structure in the image. Following Lindeberg [13], we define ridge response as the second gradient perpendicular to the ridge ($|f_{\theta\theta}|$), minus the second gradient parallel to the ridge ($|f_{(\theta-\frac{\pi}{2})(\theta-\frac{\pi}{2})}|$). This will suppress other non-elongated maxima in the image ("blobs"). More specifically, the image response for a ridge of orientation θ, at pyramid level σ is formulated as

$$\begin{aligned} f_r(\mathbf{x}, \theta, \sigma) = {} & |\sin^2\theta\, f_{xx}(\mathbf{x}, \sigma) + \cos^2\theta\, f_{yy}(\mathbf{x}, \sigma) - \\ & 2\sin\theta\,\cos\theta\, f_{xy}(\mathbf{x}, \sigma)| - \\ & |\cos^2\theta\, f_{xx}(\mathbf{x}, \sigma) + \sin^2\theta\, f_{yy}(\mathbf{x}, \sigma) + \\ & 2\sin\theta\,\cos\theta\, f_{xy}(\mathbf{x}, \sigma)| \,. \end{aligned} \tag{4}$$

Since ridges are highly scale-dependent [13] we do not expect a strong filter response at scales other than the one corresponding to the width of the limb in the image [19]. In training, we therefore only consider scales corresponding to the distance between the manually marked edges of the limb. For background training however, all four scales are considered as before.

We observe [19] that, as for edges, the distributions are independent of image scale, and hence a scale-independent empirical distribution is learned from responses at all levels.

Foreground. For each of the images in the training set (Figure 2), the scale σ_l and direction θ_l of each limb l are computed from the manually marked edges. We sample locations $\mathbf{x}$ on the limb foreground, with $\theta = \theta_l$ and $\sigma = \sigma_l$. A discrete probability function, $p^r_{\text{on}}(f_r \mid l)$ is constructed as for edges (Figure 5).

Background. As with edges, an empirical distribution of ridge responses, $f_r(\mathbf{x}, \theta, \sigma)$, in the background is learned for randomly sampled orientations, scales and image locations in the training set. The normalized distribution represents $p_{\text{off}}(f_r)$, the probability distribution over ridge filter responses of locations off human limbs (Figure 5).

While the background distribution looks similar to the edge response distribution, the foreground distributions is asymmetric about zero. Negative responses, corresponding to ridges orthogonal to the predicted orientation θ, are unlikely to come from limbs and, hence, the larger the filter response, the more likely it is to come from the foreground. This is reflected in the likelihood ratio (Figure 5).

3.3 Motion Cue

A measure of how well the model parameters, $\boldsymbol{\phi}$, fit the image data at time t is how well they predict the appearance of the human and the background given their appearance in the previous time step. In other words, we want to measure the error in predicting the image at time t based on the parameters $\boldsymbol{\phi}$ and the image at time $t-1$. The motion response at time t, $f_{m,t}$, is the pixel difference between the unfiltered image I_t, and the image I_{t-1}, warped according to the body parameters $\boldsymbol{\phi}_t$.

The 3D motion of the human model defines the 2D motion in the foreground portion, $\{\mathbf{x}_f\}$, of the image. Thus, the pixel $\mathbf{x}_{t-1}$ in $\{\mathbf{x}_{f,t-1}\}$ at time $t-1$ maps to some pixel location $\mathbf{x}_t$ in $\{\mathbf{x}_{f,t}\}$ if the limb surface point corresponding to both these image points is non-occluded at time $t-1$ and t. The pixel correspondences can be computed from $\boldsymbol{\phi}_t$.

Given two positions $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, corresponding to the same limb surface location, the motion response at time t and pyramid level σ is formulated as

$$f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma) = I_t(\mathbf{x}_t, \sigma) - I_{t-1}(\mathbf{x}_{t-1}, \sigma) \tag{5}$$

Note that this response function is only valid for positions $\mathbf{x}_t$ on the foreground (limb area). Also note that the standard brightness constancy assumption for optical flow assumes that these temporal differences are modeled by a Gaussian distribution [21].

Since the motion in the background is unknown, the background motion response is defined as $f_{m,t}(\mathbf{x}_t, \mathbf{x}_t, \sigma)$, i.e. the temporal difference between the un-warped images at time $t-1$ and t. By training on both stationary and moving sequences, this probability distribution models errors originating from moving texture as well as camera noise.

Temporal pixel differences are generally lower at coarse spatial scales since the effects of noise and aliasing are diminished. Therefore, unlike edge and ridge responses, temporal differences are not invariant to scale and we learn separate distributions for different image scales [19].

Note that it is not possible to pre-compute the ratio between the foreground and background likelihood distributions. This is because the filter responses are based on the underlying motion models which are different for foreground and background. Therefore, it is necessary to index into both the foreground and background distributions separately, and then take the ratio between the two likelihoods obtained.

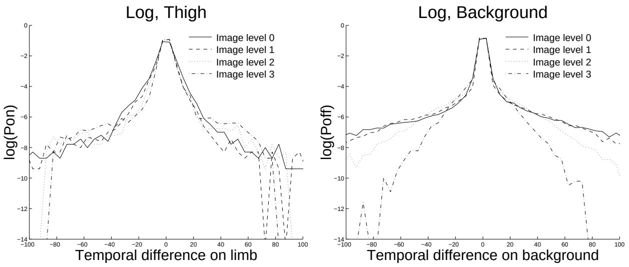

Figure 6: Motion responses. Learned log likelihoods for foreground pixel differences given known optical flow between two consecutive frames Left: Log likelihood for the thigh, $\log(p^m_{\text{on}}(f_m \mid thigh, \sigma))$ at image levels $\sigma = 0, 1, 2, 3$. Right: Log likelihood for the background, $\log(p^m_{\text{off}}(f_m \mid \sigma))$, at image levels $\sigma = 0, 1, 2, 3$. Note that we can not take the ratio between these distributions, since the response f_m is computed from different pixel differences in the background and on the limbs, originating from different motion models.

Foreground. Given training sequences, approximately 35 pairs of consecutive frames are used to learn the distributions of temporal foreground differences. Foreground locations $\mathbf{x}_t$ are sampled randomly, and the corresponding locations $\mathbf{x}_{t-1}$ in the frame before are computed from $\mathbf{x}_t$ and the marked edges. The differences at all image levels σ, $f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma)$, are collected into histograms separately for each limb l and each image scale σ. The normalized histograms represent $p^m_{\text{on}}(f_m \mid l, \sigma)$.

Given a certain observed response $f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma)$, the likelihood of observing this response in the foreground (on limb l and level σ) is $p^m_{\text{on}}(f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma) \mid l, \sigma)$. The log likelihood, $\log(p^m_{\text{on}})$, for the thigh is shown in Figure 6.

Background. Consecutive frames from sequences containing moving objects and either a static or moving camera are used as training data for the background distributions. Locations $\mathbf{x}_t$ are sampled, and histograms of $f_{m,t}(\mathbf{x}_t, \mathbf{x}_t, \sigma)$ are computed, one for each scale σ. The normalized histograms represent $p(f_m \mid \sigma)$, the probability distribution over temporal differences in general backgrounds (Figure 6).

4 Bayesian Tracking

Human tracking is formulated as an inference problem [12]. We adopt a Bayesian formulation and estimate the parameters of the body model over time using particle filtering [7, 8]. We briefly sketch the method here; for details the reader is referred to [18].

At each time instant t, the configuration of the human model is given by $\boldsymbol{\phi}_t$. Prior knowledge about the dynamics of the human body is used to generate hypotheses about the configuration at time t given the configuration at the previous time instant. The hypotheses are then compared with the sequence of filter images up to time t, $\vec{\mathbf{f}}_t$. By Bayes' rule, the posterior probability of the model parameters, $\boldsymbol{\phi}$, given $\vec{\mathbf{f}}_t$, $p(\boldsymbol{\phi}_t \mid \vec{\mathbf{f}}_t) =$

$$\kappa_2 \, p(\mathbf{f}_t \mid \boldsymbol{\phi}_t) \int p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}) p(\boldsymbol{\phi}_{t-1} \mid \vec{\mathbf{f}}_{t-1}) d\boldsymbol{\phi}_{t-1} \qquad (6)$$

where κ_2 is a normalizing constant that does not depend on the state variables.

The posterior distribution $p(\boldsymbol{\phi}_t \mid \vec{\mathbf{f}}_t)$ is modeled using a large number of samples where samples correspond to possible poses of the body, ϕ^i_t, and their normalized likelihood. We employ between 10^3 and 10^4 samples to represent the posterior distribution.

The posterior distribution can be propagated and updated over time using Equation (6) [7, 8]. This is done by drawing samples $\boldsymbol{\phi}^i_{t-1}$ according to their posterior probability at time $t-1$. These samples are then propagated in time by sampling from the temporal prior $p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1})$, which is either an activity dependent model (e.g. walking motion [18]) or a general model of smooth motion where all angles in the body are assumed independent. Details of the human motion models are described in [15, 18].

The posterior distribution is extremely peaked (i.e. the difference between the smallest and largest likelihood is very large) and thus, only a few of the samples from the posterior will be selected multiple times. Sampling from a broader distribution would result in more stable tracking, since more samples survive in each time step. Hence, instead of drawing samples from the posterior at time $t-1$, we sample from a proposal distribution that is a smoothed (approximate) version of the posterior. Using importance sampling [9], the samples are reweighted by a factor representing the probability that this particle could have been generated by the true posterior at time $t-1$, divided by the probability with which it was generated by the smoothed posterior at time $t-1$.

For each sample $\boldsymbol{\phi}^i_t$ in the propagated distribution, the likelihood is evaluated. A set of points $\{\mathbf{x}^i_{f,t}\}$ is randomly chosen from the model foreground, and the likelihood ratio between each point being foreground and background (Equation (2)) is computed. For the edge and motion cues, this is performed at several scales.

Since all configuration parameters in $\boldsymbol{\phi}$ are sampled together rather than hierarchically, occluded areas are automatically computed and removed from consideration in the likelihood evaluation.

5 Tracking Results

The performance of the likelihood model using the learned distributions was tested for different tracking tasks. A general smooth motion model was used [18]. The test sequences contained clutter, no special clothing, and no special backgrounds. The experiments use monocular grayscale sequences with both static and moving cameras. With 5000 samples and all cues, the Java implementation takes

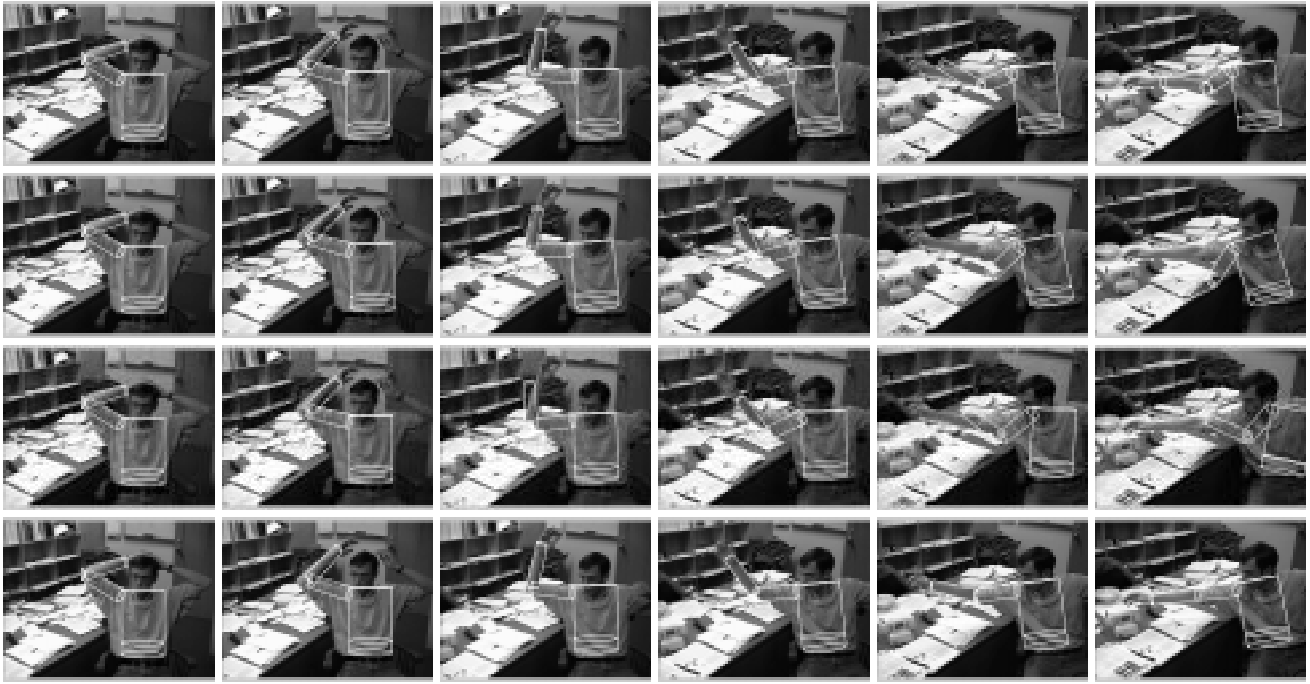

Figure 7: Tracking an arm, moving camera, 5000 samples, with different cues. The columns show frames 0, 10, 20, 30, 40 and 50 of the sequence. In each frame, the expected value of from the posterior distribution over ϕ is projected into the image. Row 1: Only flow cue. Row 2: Only edge cue. Row 3: Only ridge cue. Row 4: All cues.

approximately 7 minutes per frame on a 400 MHz Pentium III processor.

We first show how the tracking benefits from combining different image cues. Figure 7 shows four different tracking results for the same sequence. The model is initialized with a Gaussian distribution around a manually selected set of start parameters ϕ. Camera translation during the sequence causes motion of both the foreground and the background.

The first row shows tracking results using only the motion cue. As shown in [18] motion is an effective cue for tracking, however, in this example, the 3D structure is incorrectly estimated due to drift. The edge cue (row 2), does not suffer from the drift problem, but the edge information at the boundaries of the arm is very sparse and the model is caught in local maxima. The ridge cue is even less constraining (row 3) and the model has too little information to track the arm properly.

Row 4 shows the tracking result using all three cues together. We see that the tracking is qualitatively more accurate than when using any of the three cues separately. While the use of more samples would improve the performance of the individual cues, the benefit of the combined likelihood model is that it constrains the posterior and allows the number of samples to be reduced.

Next, we show an example of tracking two arms (Figure 8). In this example, the right arm is partly occluded by the left arm. Since each sample represents a generative prediction of the limbs in the scene, it is straightforward to predict occluded regions. The likelihood computations are then performed only on the visible surface points.

6 Conclusions

The two main contributions of this paper are the learning of image statistics of people and scenes, in terms of motion, edge, and ridge filter responses, and the application of these models to tracking of humans in cluttered environments. By modeling the likelihood of observing a human in terms of a foreground-background ratio we are able to track human limbs in scenes with both clutter and a moving camera.

The edge and ridge cues provide fairly sparse information about the limb appearance. To capture richer information, the likelihood model could be extended to represent statistical models of color and texture. One approach is to match distributions on the human using the Bhattacharyya distance [2].

We also would like a more explicit background model. Modeling the motion of the background would substantially constrain the tracking of the foreground. We are currently exploring the estimation of background motion using global, parametric, models such as affine or planar motion. We will need to learn background motion distributions for stabilized sequences of this form.

While the Bayesian formulation provides a framework for combining different cues, the issue of their relative

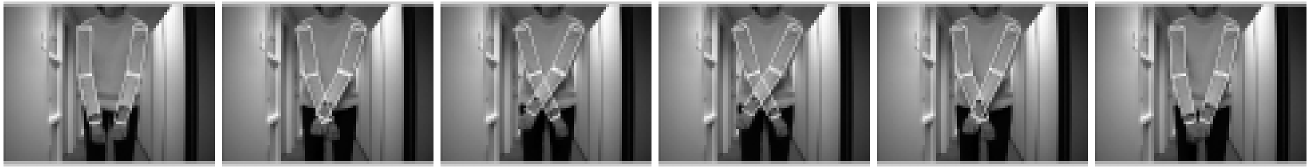

Figure 8: Tracking two crossing arms, 3000 samples. The images show the expected value of the posterior distribution in frame 0, 20, 40, 60, 80 and 100 of the sequence.

weighting requires further investigation. The issue is related to the spatial dependance of filter responses and here the weighting is implicitly determined by the number of samples chosen for each cue.

While preliminary, our experimental results suggest that learned models of object-specific and general image statistics can be exploited for Bayesian tracking. In contrast to the situation in the speech recognition community, data collection with ground truth remains a significant hurdle for learning in applications such as people tracking. We believe that building on the careful analysis of image statistics currently under way in the literature [10, 11, 14, 20, 26] will lead to more robust algorithms. Towards that end, training data and ground truth used here can be downloaded from `http://www.nada.kth.se/~hedvig/data.html`.

Acknowledgments. This work was partially sponsored by the Foundation for Strategic Research under the "Center for Autonomous Systems" contract. This support is gratefully acknowledged. We thank David Fleet who developed an early edge likelihood and provided many valuable insights. We are grateful to Allan Jepson for discussions on foreground/background modeling and Bayesian tracking. We would also like to thank Jan-Olof Eklundh, Tony Lindeberg, and Josephine Sullivan for helpful discussions on filters and likelihood models.

References

[1] T-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, vol. 1, pp. 239–245, 1999.

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

[3] T. Darrell, G. Gordon, M. Harville, and J Woodfill. Integrated person tracking using stereo, color, and pattern detection. *IJCV*, 37(2):175–185, 2000.

[4] J. Deutscher, A. Blake, and I. Reid. Articulated motion capture by annealed particle filtering. *CVPR*, vol. 2, pp. 126–133, 2000.

[5] W. Freeman and E. Adelson. The design and use of steerable filters. *PAMI*, 13(9):891–906, 1991.

[6] D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *PAMI*, 18(1):1–14, 1996.

[7] N. Gordon. A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar, Sonar & Navigation*, 140(2):107–113, 1993.

[8] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV*, pp. 343–356, 1996.

[9] M. Isard and A. Blake. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. *ECCV*, pp. 893-909, 1998.

[10] S. Konishi, A. Yuille, J. Coughlan, and S. Zhu. Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues. submitted: *PAMI*.

[11] A. Lee, J. Huang, and D. Mumford. Random-collage model for natural images. to appear: *IJCV*.

[12] M. Leventon and W. Freeman. Bayesian estimation of 3-D human motion from an image sequence. TR–98–06, Mitsubishi Electric Research Lab, 1998.

[13] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *IJCV*, 30(2):117–156, 1998.

[14] B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network Computation in Neural Systems*, 7(2):333-339, 1996.

[15] D. Ormoneit, H. Sidenbladh, M. Black, and T. Hastie. Learning and tracking cyclic human motion. *NIPS*, 2000

[16] C. Rasmussen and G. Hager. Joint probabilistic techniques for tracking multi-part objects. *CVPR*, pp. 16–21, 1998.

[17] J. Rittscher, J. Kato, S. Joga, and A. Blake, A probabilistic background model for tracking. *ECCV*, pp. 336–350, 2000.

[18] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, vol. 2, pp. 702–718, 2000.

[19] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. submitted: *IJCV*.

[20] E. Simoncelli. Statistical models for images: Compression, restoration and optical flow. *Asilomar Conf. Signals, Systems & Computers*, pp. 673–678, 1997.

[21] E. Simoncelli, E. Adelson, and D. Heeger. Probability distributions of optical flow. *CVPR*, pp. 310–315, 1991.

[22] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by Bayesian correlation. *ICCV*, pp. 1068–1075, 1999.

[23] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. *ECCV*, pp. 307–323, 2000.

[24] S. Wachter and H. Nagel. Tracking of persons in monocular image sequences. *CVIU*, 74(3), 1999.

[25] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. *PAMI*, 19(7):780–785, 1997.

[26] S. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *PAMI*, 19(11), 1997.