International Journal of Computer Vision 54(1/2/3), 183–209, 2003
© 2003 Kluwer Academic Publishers. Manufactured in The Netherlands.

Learning the Statistics of People in Images and Video

HEDVIG SIDENBLADH*
Computational Vision and Active Perception Laboratory, Department of Numerical Analysis and Computer Science, KTH, SE-100 44 Stockholm, Sweden
hedvig@nada.kth.se

MICHAEL J. BLACK
Department of Computer Science, Box 1910, Brown University, Providence, RI 02912, USA
black@cs.brown.edu

Received May 1, 2001; Revised May 21, 2002; Accepted January 7, 2003

Abstract. This paper address the problems of modeling the appearance of humans and distinguishing human appearance from the appearance of general scenes. We seek a model of appearance and motion that is generic in that it accounts for the ways in which people's appearance varies and, at the same time, is specific enough to be useful for tracking people in natural scenes. Given a 3D model of the person projected into an image we model the likelihood of observing various image cues conditioned on the predicted locations and orientations of the limbs. These cues are taken to be steered filter responses corresponding to edges, ridges, and motion-compensated temporal differences. Motivated by work on the statistics of natural scenes, the statistics of these filter responses for human limbs are learned from training images containing hand-labeled limb regions. Similarly, the statistics of the filter responses in general scenes are learned to define a "background" distribution. The likelihood of observing a scene given a predicted pose of a person is computed, for each limb, using the likelihood ratio between the learned foreground (person) and background distributions. Adopting a Bayesian formulation allows cues to be combined in a principled way. Furthermore, the use of learned distributions obviates the need for hand-tuned image noise models and thresholds. The paper provides a detailed analysis of the statistics of how people appear in scenes and provides a connection between work on natural image statistics and the Bayesian tracking of people.

Keywords: human tracking, image statistics, Bayesian inference, articulated models, multiple cues, likelihood models

1. Introduction

The detection and tracking of humans in unconstrained environments is made difficult by the wide variation in their appearance due to clothing, illumination, pose, gender, age, etc. We seek a generic model of human appearance and motion that can account for the ways in which people's appearance varies and, at the same time, is specific enough to be useful for distinguishing people from other objects. Building on recent work in modeling natural image statistics, our approach exploits generic filter responses that capture information about appearance and motion. Statistical models of these filter responses are learned from training examples and provide a rigorous probabilistic model of the appearance of human limbs. Within a Bayesian framework, these object-specific models can be compared with generic models of natural scene statistics. The

*Present address: Department of Data and Information Fusion, Swedish Defence Research Agency (FOI), SE-172 90 Stockholm, Sweden. E-mail: hedvig@foi.se.

resulting formulation proves suitable for Bayesian tracking of people in complex environments with a moving camera.

Previous work on human motion tracking has exploited a variety of image cues (see Gavrila (1999) or Moeslund and Granum (2001) for recent reviews). In many cases, these cues are sequence-specific and capture local color distributions (Wren et al., 1997) or segment the person from the background using a known background model (Haritaoglu and Davis, 2000). While appropriate for some user interface applications, these sequence-specific approaches are difficult to extend to arbitrary image sequences.

Tracking approaches for generic scenes have typically used extracted edge information (Deutscher et al., 2000; Gavrila, 1996; Hogg, 1983; Isard and Blake, 1998; Rehg and Kanade, 1995; Rohr, 1994), optical flow (Bregler and Malik, 1998; Ju et al., 1996; Yacoob and Black, 1999) or both (DeCarlo and Metaxas, 1996; Sminchisescu and Triggs, 2001; Wachter and Nagel, 1999). Edges are first extracted using some standard technique and then a match metric is defined that measures the distance from predicted model edges (e.g. limb boundaries) to detected edges in the scene. Probabilistic tracking methods convert this match metric into an ad hoc "probabilistic" likelihood of observing image features given the model prediction.

Approaches that use image motion as a cue typically assume brightness constancy holds between pairs of adjacent frames (Bregler and Malik, 1998; Ju et al., 1996) or between an initial template and the current frame (Cham and Rehg, 1999). As with edges, an ad hoc noise model is often assumed (Gaussian or some more "robust" distribution) and is used to derive the likelihood of observing variations from brightness constancy given a predicted motion of the body.

These probabilistic formulations have recently been incorporated into Bayesian frameworks for tracking people (Cham and Rehg, 1999; Deutscher et al., 2000; Isard and Blake, 1998; Sidenbladh et al., 2000a; Sminchisescu and Triggs, 2001; Sullivan et al., 1999; Sullivan et al., 2000). These Bayesian methods allow the combination of various image cues, represent ambiguities and multiple hypotheses, and provide a framework for combining new measurements with the previous history of the human motion in a probabilistically sound fashion. The Bayesian methods require a temporal prior probability distribution and a conditional likelihood distribution that models the probability of observing image cues given a predicted pose or motion of the body.

In contrast to previous work, our goal is to formulate a rigorous probabilistic model of human appearance by learning distributions of image filter responses from training data. Given a database of images containing people, we manually mark human limb axes and boundaries for the thighs, calves, upper arm, and lower arm. Motivated by Konishi et al. (1999), probability distributions of various filter responses on human limbs are constructed as illustrated in Fig. 1. These filters are based on various derivatives of normalized Gaussians (Lindeberg, 1998) and provide some measure of invariance to variations in clothing, lighting, and background.

The boundaries of limbs often differ in luminance from the background resulting in perceptible edges. Filter responses corresponding to edges are therefore computed at the boundaries of the limbs. First derivatives of normalized Gaussian filters are steered (Freeman and Adelson, 1991) to the orientation of the limb and are applied at multiple scales. Note that an actual edge may or may not be present in the image depending on the local contrast; the statistics of this are captured in the learned distributions and vary from limb to limb.

In addition to boundaries, the elongated structure of a human limb can be modeled as a ridge at an appropriate scale. We employ a steerable ridge filter that responds strongly where there is high curvature of the image brightness orthogonal to the limb axis and low curvature parallel to it (Lindeberg, 1998).

Motion of the body gives rise to a third and final cue. We assume that the intensity pattern on the surface of the limb will change slowly over time. Given the correct motion of the limb, the image patch corresponding to it can be warped to register two consecutive frames. The assumption of brightness constancy implies that the temporal derivatives for this motion-compensated pair are small. Rather than assume some arbitrary distribution of these differences we learn the distribution for hand registered sequences and show that it is highly non-Gaussian.

These learned distributions can now form the basis for Bayesian tracking of people. While these models characterize the "foreground" object, reliable tracking requires that the foreground and background statistics be sufficiently distinct. We thus also learn the distribution of edge, ridge, and motion filter responses for general scenes without people. This builds upon

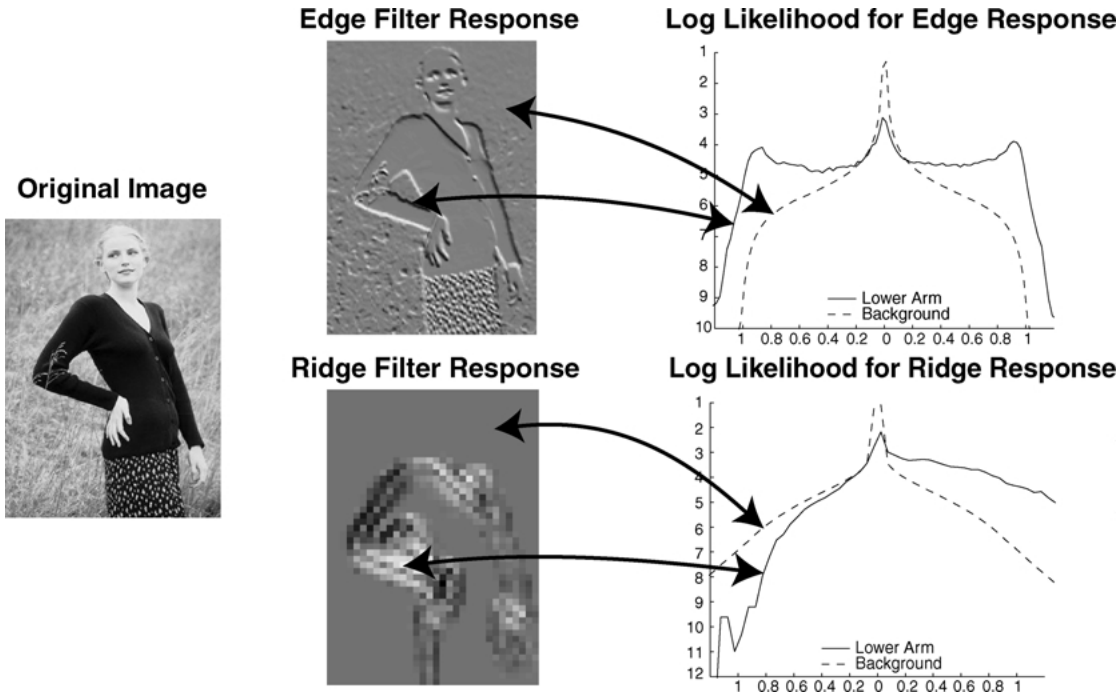


Figure 1. Learning the appearance of people and scenes. Distributions over edge and ridge filter response are learned from examples of human limbs and general scenes.

recent work on learning the statistics of natural scenes (Konishi et al., 1999; Lee et al., 2001; Olshausen and Field, 1996; Ruderman, 1994; Simoncelli, 1997; Zhu and Mumford, 1997) and extends it to the problem of people tracking. 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. Rittscher et al. (2000)):

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

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.

Additionally, we note that the absolute contrast between foreground and background is less important than the consistency of edge or ridge orientation. We therefore perform contrast normalization prior to filtering.[1] The formulation of foreground and background models provides a principled way of choosing the appropriate type of contrast normalization. For an optimal Bayesian detection task we would like the foreground and background distributions to be maximally distinct under some distance measure. We exploit an approach based on the Bhattacharyya distance between foreground and background distributions (Kaliath, 1951; Konishi et al., 1999).

This paper focuses on the the detailed analysis of the image statistics of people and only briefly describes the Bayesian tracking framework; details of the approach can be found in Sidenbladh and Black (2001) and Sidenbladh et al. (2000a). In the approach, the body is modeled as an articulated collection of 3D truncated cones. Using a particle filtering method (Gordon, 1993; Isard and Blake, 1998; Sidenbladh et al., 2000a), the posterior probability distribution over poses of the body model is represented using a discrete set of samples (where each sample corresponds to some pose of

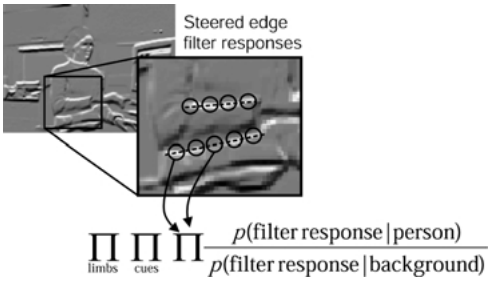


Figure 2. Steered edge responses. Edge responses are computed at the orientation of the limb and are sampled along the limb boundary. The ratio of the conditional probability that the filter responses were generated by a limb versus some generic background is related to the likelihood of observing the image. Assuming independence of the various cues and limbs, the overall likelihood is proportional to the product of the likelihood ratios.

the body). Each sample is projected into the image giving predicted limb locations and orientations in image coordinates. Image locations along the predicted limbs are sampled and the filter responses steered to the predicted orientation are computed. The learned distributions give the likelihood of observing these filter responses given the model. Assuming independence of the edge, ridge, and motion cues, the product of the individual terms provides the likelihood of observing these filter responses conditioned on the predicted pose (Fig. 2).

The approach extends 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 experiments in which the learned likelihood models are evaluated with respect to robustness and precision in spatial displacement of the limb models, and tracking examples that illustrate how the tracking benefits from a likelihood exploiting multiple cues.

2. Related Work

This paper applies ideas from work on the statistics of natural images to the task of Bayesian detection and tracking of humans. Both of these areas have attracted a considerable amount of interest in the recent years and are reviewed briefly here.

2.1. Appearance Models for Tracking Humans

Given the complexity of the appearance of a human, it is difficult to use bottom-up approaches to detect humans in images (Hogg, 1983). Most recent approaches to detection and tracking of humans employ some kind of model to introduce a priori information about the range of possible appearances of a human. These models vary in complexity from assemblies of 2D color blobs (Wren et al., 1997) or areas with a certain color distribution (Comaniciu et al., 2000), to layered 2D representations of articulated figures (Cham and Rehg, 1999; Ju et al., 1996), and, finally, to detailed 3D articulated structures (Bregler and Malik, 1998; Deutscher et al., 2000; Gavrila, 1996; Hogg, 1983; Rehg and Kanade, 1995; Rohr, 1994, 1997; Sidenbladh et al., 2000a; Sminchisescu and Triggs, 2001; Wachter and Nagel, 1999).

Tracking using articulated models involves (in the 3D case) projecting a certain configuration of the model into the image, and comparing the model features with the observed image features. In a probabilistic formulation of the problem, this corresponds to computing the likelihood of the observed image features, conditioned on the model configuration.

Depending on the application, many different techniques have been used to extract features for image-model comparison. Background subtraction (Deutscher et al., 2000; Haritaoglu and Davis, 2000; Rohr, 1994, 1997; Wren et al., 1997) gives an estimate of where the human is in the image, and the outline of the human, but does not provide information about the motion of the foreground. Furthermore, most background segmentation algorithms require a static camera and and slowly changing scenes and lighting conditions. While in many applications, these background assumptions are reasonable, the approach is difficult to extend to the general case of unknown, complex, and changing scenes.

To exploit more detailed information about the position of the individual limbs, researchers have also used detected image edges. Observing correlation between the boundaries of the human model and detected edges has proven to be successful in tracking, especially in indoor environments with little clutter (Deutscher et al., 2000; Gavrila, 1996). The common approach is to detect edges using a threshold on some image edge response (Fig. 3). After the detection, the distance from the limb boundaries to the detected image edges are used to determine the correlation between the model

Figure 3. Example of edge detection using the Canny filter. Left: Original image. Center: Typical Canny edges; too many edges in some regions, too few in others. Right: How should predicted limb edges be compared with the detected edge locations?

and the image. This can be computed using the Chamfer distance (Gavrila, 1996), or by enforcing a maximum distance between limb boundaries and image edges (Hogg, 1983; Wachter and Nagel, 1999). Alternatively, Isard and Blake (1998) define an edge distance measure that is converted into a conditional likelihood distribution. In this way, segmented edge information can be used in a Bayesian tracking framework, but the probabilistic model lacks formal justification.

Although successful, there are problems with these approaches; for example the segmentation typically depends on an arbitrarily set threshold. When thresholding, most information about edge strength is removed, leaving little information. Furthermore, it is not clear how to interpret the similarity between the edge image and the model in a probabilistic way.

The approach proposed here avoids these problems: Instead of first detecting edges in the image, using a threshold on an edge response, we observe the continuous edge response along the predicted limb boundary and compute the likelihood of observing the response using probability distributions learned from image data. Thus, more information about the edge response is taken into account, while enabling a principled formulation of the likelihood.

Edges provide a quite sparse representation of the world, since they only provide information about the location of limb boundaries. More information about the limb appearance can be derived from the assumption of temporal brightness constancy—that two image locations originating from the same scene location at two consecutive time instants have the same intensity. This assumption is used widely for tracking of humans (DeCarlo and Metaxas, 1996; Sidenbladh et al., 2000a; Wachter and Nagel, 1999). There are two problems with this assumption. First, since there is no absolute model of the limb appearance, any errors in the estimated motion will accumulate over time, and the model may drift off the tracking target, and eventually follow the background or some other object. To avoid this drift, brightness constancy is therefore often used in combination with edges (DeCarlo and Metaxas, 1996; Wachter and Nagel, 1999). Second, the assumption of brightness constancy never strictly holds and therefore one typically assumes that deviations from the assumption are distributed according to some distribution. This distribution is typically assumed to be Gaussian (Simoncelli et al., 1991) or some heavy-tailed, "robust", distribution (Black and Anandan, 1996). Within the framework proposed here, we learn this distribution from hand-registered sequences of human motion and show that it is, in fact, highly non-Gaussian. This learned distribution provides a rigorous probabilistic interpretation for the assumption of brightness constancy. Moreover, we show that the distribution is related to robust statistical methods for estimating optical flow (Black and Anandan, 1996).

The use of fixed templates also involves a brightness constancy assumption. However, instead of comparing corresponding image locations between two consecutive frames t and $t-1$, the image at time t is compared to a reference image at time 0. Templates have been used successfully for face tracking (Sullivan et al., 2000), and have also proven suitable for tracking of articulated structures in constrained cases (Cham and Rehg, 1999; Rehg and Kanade, 1995). One problem with templates for 3D structures is that the templates are view-based. Hence, if the object rotates, the tracking may fail since the system only "knows" what the object looks like from the orientation it had at time 0.

Black and Jepson (1998) addressed this problem by learning parameterized models of the appearance of an object from an arbitrary view given a few example views of the same object. This idea is extended in Sidenbladh et al. (2000b) for learning low-dimensional linear models of the appearance of cylindrical limb surfaces using principal component analysis. The

drawback of this approach is that the particular limb appearance of the people to be tracked must be learned in advance. Thus, these limb appearance models are only suitable for tracking people where the appearance varies little; for example, sports teams where the clothing is restricted. Recent work on tracking and learning appearance models (Jepson et al., 2001) may provide a principled way of adapting models of limb appearance over time.

The cues described above for comparing human models with images exhibit different strengths and weaknesses. Thus, none of the cues is entirely robust when used on its own. 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 (Darrell et al., 2000; Rasmussen and Hager, 2001; Wachter and Nagel, 1999), the formulation and combination of these cues is often ad hoc. The Bayesian approach presented in this paper enables combination of different cues in a principled way (for a related Bayesian method see Rasmussen and Hager (2001)). Moreover, by learning noise models and likelihood distributions from training data the problems of hand tuned noise models and thresholds are avoided.

2.2. *Statistics of Natural Images*

Recently there has been a large interest in learning the low-order spatial and temporal statistics of natural scenes. The statistics of grey-level values (Lee et al., 2000; Olshausen and Field, 1996; Ruderman, 1994, 1997; Zhu and Mumford, 1997) as well as first order (Lee et al., 2001; Konishi et al., 1999) and second order (Geman and Jedynak, 1996; Sullivan et al., 1999, 2000) gradients, and wavelet responses (Simoncelli, 1997) have been studied. These statistics have been used to aid image compression and restoration, and to model biological vision. The distributions over different kinds of filter responses have two notable things in common: The distributions are invariant over scale (Lee et al., 2001; Ruderman, 1994; Zhu and Mumford, 1997), and they are non-Gaussian, with a high kurtosis (Geman and Jedynak, 1996; Lee et al., 2001; Ruderman, 1994; Zhu and Mumford, 1997).

Most of the work on the statistics of images has focused on generic scenes rather than specific objects. Here we are interested in modeling the appearance of people and, hence, we would like to model the statistics of how people appear in, and differ from, natural scenes. This is similar in spirit to the work of Konishi et al. (1999). 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 (e.g., derivative of Gaussian) response for these edge locations. For our purposes we construct steerable image pyramids (Freeman and Adelson, 1991) using normalized Gaussian derivative filters (first and second order) (Lindeberg, 1998). With this representation, the filter response for any predicted limb orientation can be computed.

In our case, we model the empirical distribution of 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 the boundary of a limb 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. as a *generic* feature distribution while here we learn an *object-specific* distribution for people.

Konishi et al. (1999) also compute 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 modeling the likelihood of observing an image given a person in front of a generic, unknown, background. In related work, Nestares and Fleet (2001) use a steerable pyramid of quadrature-pair filters (Freeman and Adelson, 1991) and define the likelihood of an edge in terms of the empirical distribution over the amplitude and phase of these filter responses.

Finally, the absolute contrast between foreground and background is less important for detecting people than the orientation of the features (edges or ridges). We show that local contrast normalization prior to filtering enables better discrimination between foreground and background edge response. This would be less appropriate for the task of Konishi et al. (1999) and, as a result of normalization, the distributions we learn have a somewhat different shape.

Our work is also closely related to the tracking work of Sullivan et al. (1999, 2000) who model the distributions of filter responses for a general background and a particular foreground (using a generalized template). Given these distributions, they can determine if an image patch is background, foreground or on the boundary

by matching the distribution of filter responses in an image patch with the learned models for foreground, background, and boundary edges. 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 background and foreground in evenly distributed locations in the image. We use several different filter responses, and we use steerable filters (Freeman and Adelson, 1991) 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 the Filter Distributions

Scenes containing a single person can be viewed as consisting of two parts, the human (foreground) and the background, with pixels in an image of the scene belonging to one region or the other. A given configuration of the human model defines these foreground and background regions and, for a pixel $\mathbf{x}$ in the image, the likelihood of observing the filter responses at $\mathbf{x}$ can be computed given the appropriate learned models. The likelihood of the entire scene will then be defined in terms of the product of likelihoods at a sampling of individual pixels. The formulation of such a likelihood will be described in Section 4.

As stated in the introduction, the filter responses, $\mathbf{f} = [f_e, f_r, f_m]$ include edge responses f_e, ridge responses f_r and the motion responses f_m. Edges filter responses are only measured on the borders of the limb, while all positions on the limb are considered for the motion responses. Ridge responses are evaluated at pixels near the axis of the limb at the appropriate scale.

Probability distributions over these responses are learned both on human limbs and also for general background scenes. Let the probability distributions of foreground filter responses be $p^e_{\text{on}}(f_e)$, $p^r_{\text{on}}(f_r)$, $p^m_{\text{on}}(f_m)$ and the distributions over background filter responses be $p^e_{\text{off}}(f_e)$, $p^r_{\text{off}}(f_r)$, $p^m_{\text{off}}(f_m)$, following the notation of Konishi et al. (1999). Traditionally, it has been assumed that these distributions take a Gaussian shape. Studies on the statistics of natural images (Konishi et al., 1999; Lee et al., 2001; Olshausen and Field, 1996; Ruderman, 1994; Simoncelli, 1997; Zhu and Mumford, 1997) have shown that this is not the case—the distributions are highly non-Gaussian. To capture the actual shape of the distributions, we learn them from image training data.

This training set consists of approximately 150 images and short sequences of people in which the outline of limbs are marked manually. Examples of marked training images are given in Fig. 4. Since human heads and torsos generally do not display straight edges, nor clear ridges in the image, we do not consider these body parts for the edge and ridge cue. Distributions over edge and ridge response are only learned for upper and lower arms and legs. However, the head and torso are considered for the motion cue, and are therefore included in the images in Fig. 4.

In the figures below, we often display the logarithm of the ratio between the likelihood of the observed filter response on the foreground, and the likelihood of the same response on the background:

$$b^z(f_z(\mathbf{x})) = \log\left(\frac{p^z_{\text{on}}(f_z(\mathbf{x}))}{p^z_{\text{off}}(f_z(\mathbf{x}))}\right) \tag{1}$$

where z is either e (for edge filter response), r (for ridge filter response) or m (for motion filter response). Without any prior knowledge, if the log likelihood ratio b^z is negative, $\mathbf{x}$ is more likely to belong to the background, if it is positive, $\mathbf{x}$ is more likely to belong to the foreground. This ratio will be exploited in the formulation of the limb likelihood in Section 4.

The sub-sections below provide details of the statistical models for the various cues.

Figure 4. Example images from the training set with limb edges manually marked.

3.1. Edge Cue

To capture edge statistics at multiple scales, a Gaussian pyramid is created from each image, and filter responses are computed at each level of the pyramid. Level σ in the pyramid is obtained by convolving the previous level $\sigma - 1$ with a 5×5 filter window approximating a Gaussian with variance 1 and sub-sampling to half the size. The finest level, $\sigma = 0$ is the original image.

Let the edge response f_e be 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 response at the four finest pyramid levels, $\sigma = 0, 1, 2, 3$. 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:

$$f_e(\mathbf{x}, \theta, \sigma) = \sin\theta \; f_x(\mathbf{x}, \sigma) - \cos\theta \; f_y(\mathbf{x}, \sigma) \qquad (2)$$

where $f_x(\mathbf{x}, \sigma)$ and $f_y(\mathbf{x}, \sigma)$ are the image derivatives in the x and y image dimensions respectively at pyramid level σ and image position $\mathbf{x}$. Figure 5(b) shows examples of the steered edge response for a lower arm at different pyramid levels.

3.1.1. Learning Foreground and Background Distributions. For each of the images in the training set, the edge orientation θ_l, in image coordinates, for each limb l is computed from the manually marked edges. For all levels σ in the image pyramid, a number of locations $\mathbf{x}_i$ are sampled on the marked edges, with $\theta = \theta_l$.

For each limb l and each level σ, a separate histogram of steered edge responses, $f_e(\mathbf{x}, \theta_l, \sigma)$, is constructed using the sampled foreground edge locations $\mathbf{x}_i$. The normalized histograms represent $p^e_{\text{on}}(f_e \mid l, \sigma)$, the probability of edge response f_e conditioned on limb number l and pyramid level σ given that the model projects to an actual limb. Given a certain observed response $f_e(\mathbf{x}, \theta_l, \sigma)$, the likelihood of observing this response in the foreground (on limb l) is $p^e_{\text{on}}(f_e(\mathbf{x}, \theta_l, \sigma) \mid l, \sigma)$. Figure 6(a) shows the logarithm of p^e_{on} for the thigh, at pyramid levels 0, 1, 2 and 3.

The background edge distribution is learned from several hundred images with and without people. From these images, a large number of locations $\mathbf{x}$ are sampled uniformly over the image at all levels σ. We do

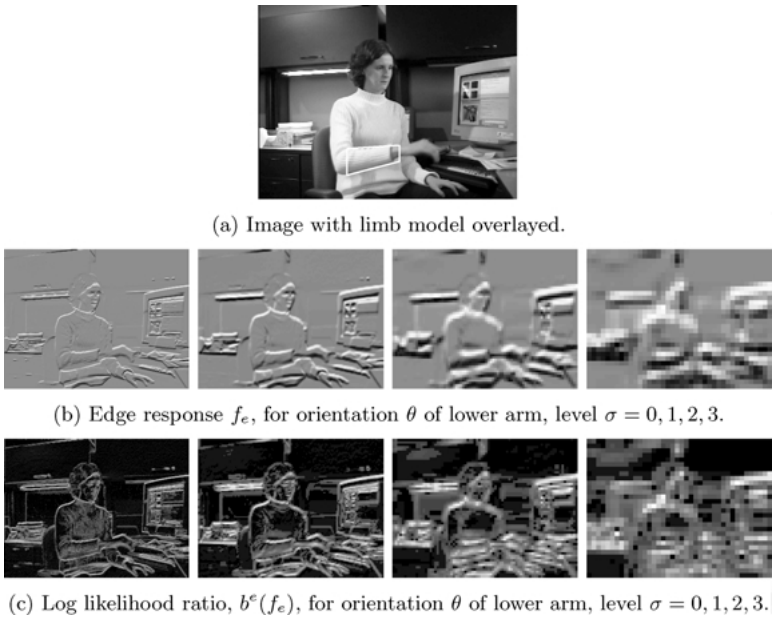


Figure 5. Computation of steered edge response. The original image with the overlayed model is shown in (a), while (b) shows the edge response $f_e(\mathbf{x}, \theta, \sigma)$ for the lower arm edges with angle θ corresponding to the orientation of the major axis of the projected limb. White denotes strong positive edge response, black strong negative response, grey weak response. The corresponding log likelihood ratio $b^e(f_e)$ for every image location is shown in (c). White denotes high (positive) log likelihood ratio, black low (negative) log likelihood ratio.

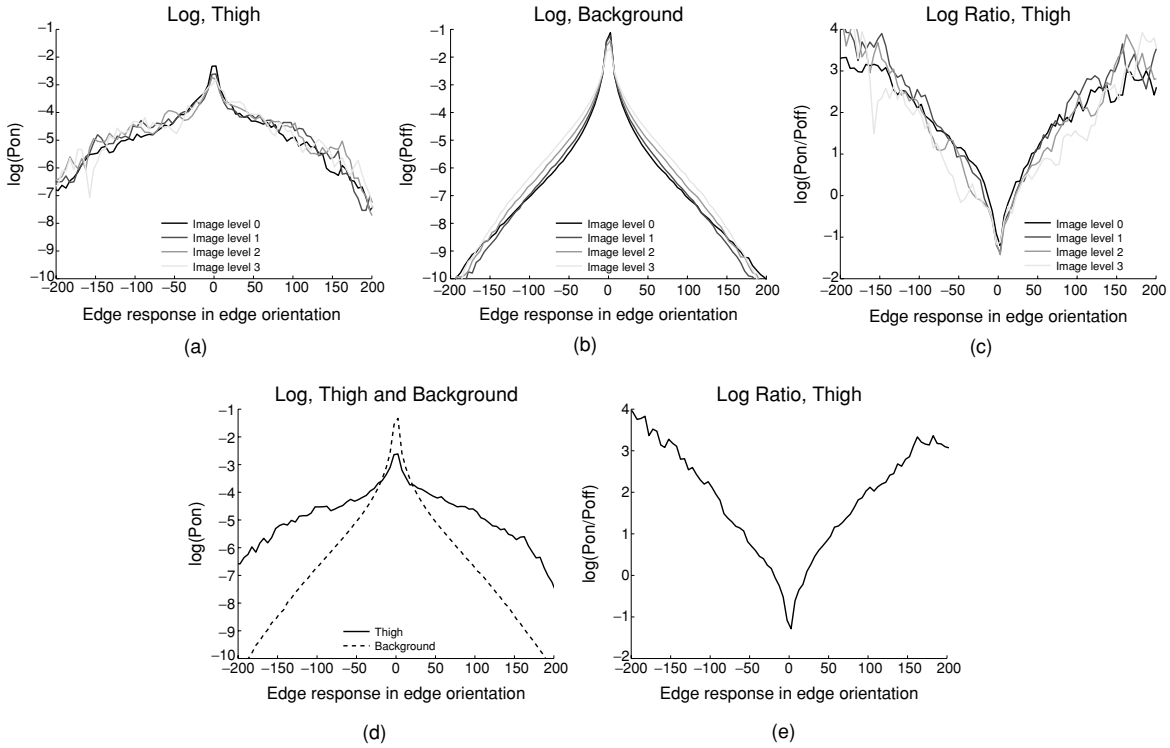


Figure 6. Foreground and background distributions. The empirical distribution (log probability) for Thigh is shown. The horizontal axis corresponds to the edge filter response given the correct limb location and orientation. (a) The thigh log likelihood for different image levels. (b) The background log likelihood for the same levels. (c) The log likelihood ratio for different levels. (d) Log likelihoods integrated over pyramid level. (e) Final log likelihood ratio.

not assume any prior information on edge directions in general scenes, and thus orientations for edge response directions θ are also sampled uniformly between 0 and 2π. The normalized version of the histograms over edge responses $f_e(\mathbf{x}, \theta, \sigma)$ at the sampled locations, orientations and levels represent $p^e_{\text{off}}(f_e \mid \sigma)$, the probability of edge responses conditioned on pyramid level, given that we look at locations and orientations that *do not correspond to* the edges of human limbs. According to this function, the likelihood of observing a certain edge response $f_e(\mathbf{x}, \theta, \sigma)$ in the background is $p^e_{\text{off}}(f_e(\mathbf{x}, \theta, \sigma) \mid \sigma)$. Figure 6(b) shows the logarithm of p^e_{off} for pyramid levels 0, 1, 2 and 3.

Both the background and foreground distributions have maxima at 0. This means that it is more likely to observe low edge filter responses both in the foreground and the background. However, the probability of responses around 0 is higher for the background distributions. This means that if a low filter response is observed, it is more likely to be observed on the background than on the foreground.

This information is captured by the log likelihood ratio (Eq. (1), Fig. 6(c)). It has a minimum at filter response 0, and grows for larger negative or positive filter responses f_e. For small values of f_e, the log likelihood ratio is negative—these filter responses are more likely to be observed in the background. For larger positive or negative values of f_e, the log likelihood ratio is positive, which means that these filter responses are more common in the foreground than in the background (assuming no a priori information for now).

Studying the distributions over foreground (Fig. 6(a)) and background (b), as well as the ratio between them (c), we note that they are very similar at different scales. This is also found by Ruderman (1994, 1997) and Zhu and Mumford (1997)—edge response is consistent over scale. Based on the assumption that the underlying distributions for different levels are the same, the learned distributions for all levels are all represented by integrating over the scale variable. This marginal distribution, $p_{\text{on}}(f_e \mid l)$, is based on more training data, and therefore more representative of the

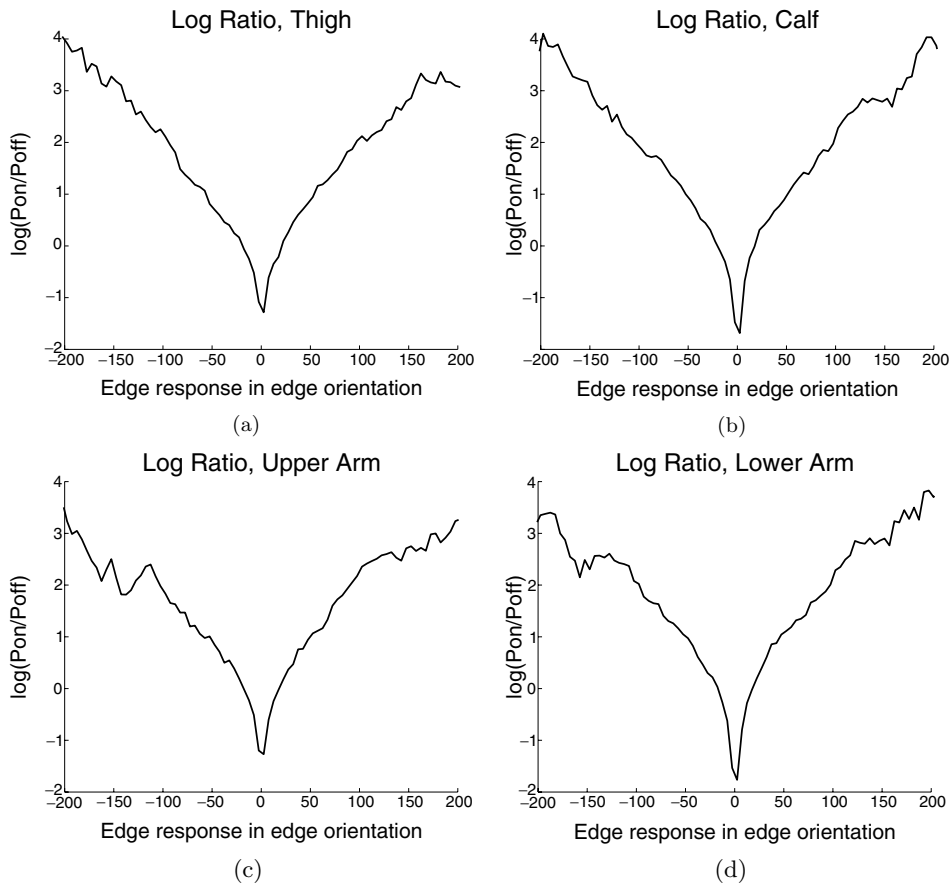


Figure 7. Learned log likelihood ratios for edges. No contrast normalization.

true distribution (Konishi et al., 1999). The likelihood of edge responses for different pyramid levels will be computed using this marginal distribution (Fig. 6(d) and (e)).

The scale-independent log likelihood ratios for the thigh, calf, upper arm and lower arm were learned from the training set and are shown in Fig. 7. The ratios for calf and lower arm have more pronounced valleys near zero than the ones for thigh and upper arm. This implies that edges generally are more pronounced at calfs and lower arms. This corresponds to intuition, since thighs often are viewed together, and upper arms often are viewed next to the torso, which usually have the same clothing as the arm.

3.1.2. "Distance" Between Foreground and Background. The "shape" of the likelihood ratio plot is related to "distance" between the distributions p^e_{on} and p^e_{off}. If the distributions p^e_{on} and p^e_{off} are very similar, the log likelihood ratio b^e is very close to 0 for all filter responses—the distributions cannot be used to determine if a pixel with a certain filter response belongs to the foreground or the background. Distinguishing people from non-people will be easier the more these foreground and background distributions are dissimilar.

The Bhattacharyya distance (Kaliath, 1951) provides one measure of similarity. Given two distributions p_{on} and p_{off} over the variable y, the Bhattacharyya distance between them is

$$\delta_B(p_{\text{on}}, p_{\text{off}}) = -\log \int \sqrt{p_{\text{on}}(y)\, p_{\text{off}}(y)}\, dy. \quad (3)$$

Alternatively, the Kullback-Leibler (1951) divergence between p_{on} and p_{off} is given by

$$\delta_{KL}(p_{\text{on}}, p_{\text{off}}) = \int p_{\text{on}}(y) \log \frac{p_{\text{on}}(y)}{p_{\text{off}}(y)}\, dy. \quad (4)$$

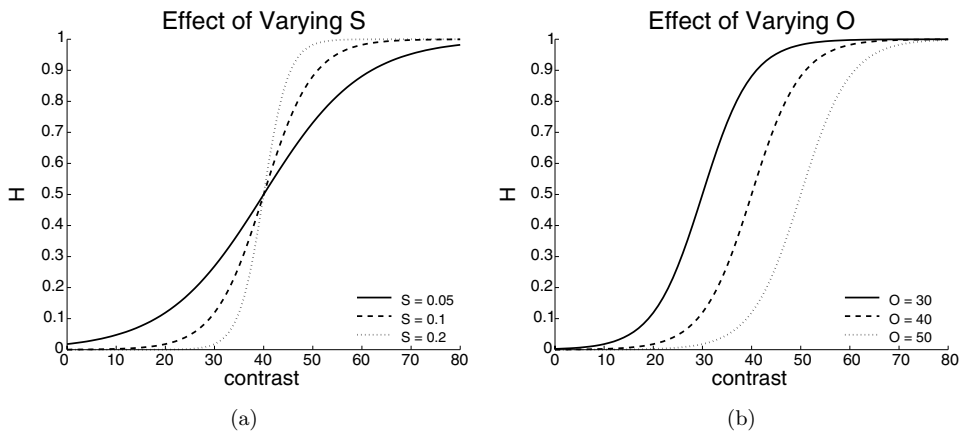


Figure 8. *The function* $H(contrast)$ for different values of offset O and slope S.

Note, the Kullback-Leibler divergence is asymmetric, which strictly means that it is not a distance metric, but it still provides a measure of the difference between the two distributions.

Below we use these measures to chose contrast normalization parameters that maximize the difference between p_{on} and p_{off}.

3.1.3. Normalization of Image Gradients. Filter response is effected by the image contrast between foreground and background which varies due to clothing, illumination, shadows, and other environmental factors. What does not vary is that the maximal filter response should be obtained at the predicted edge orientation (if the edge is visible). Reliable tracking therefore requires filter responses that are relatively insensitive to contrast variation. This can be achieved by normalizing image contrast prior to the computation of the image derivatives.

Two different normalization schemes are explored; one normalizes the image contrast locally and the other one normalizes the image values globally. These two methods, and the obtained results, are described below. The distributions using the normalization techniques are compared with the unnormalized distributions using the distance measures described above.

Local Contrast Normalization. Local contrast normalization can be obtained using a hyperbolic tangent nonlinearity that involves scaling the image derivatives at pixel $\mathbf{x}$ by a weighting factor

$$h(contrast) = \frac{1 + \tanh(S\, contrast - O)}{2\, contrast} \qquad (5)$$

where *contrast* is the maximum absolute pixel difference in a 3×3 window around $\mathbf{x}$, and S and O are parameters determining the slope and offset of the hyperbolic tangent function. For display, $H(contrast) = h(contrast)\, contrast$ is is plotted for different values of S and O in Fig. 8. H maps the original contrast to the normalized window contrast on which the gradient computation is based.

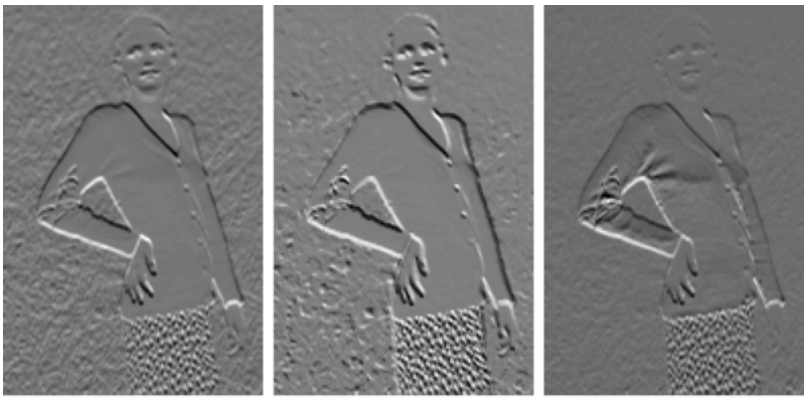
The scaling nonlinearity causes areas of low contrast to be normalized to zero contrast and areas of high contrast to be normalized to unit contrast. The horizontal and vertical derivatives of the normalized image are then either 0, or cosine functions of the angle between the gradient direction and the horizontal or vertical direction—the edge response is now more dependent on orientation information than contrast information. Figure 9(b) shows the first derivative in vertical direction, using the local normalization scheme. This can be compared to Fig. 9(a), which shows the corresponding un-normalized derivative image.

The shape of the tanh function is determined by S and O. The Bhattacharyya distance and Kullback-Leibler divergence can be used to select the optimal values of these parameters that maximize the distance between the learned distributions. The distributions p^e_{on} and p^e_{off} for different limbs are learned from normalized gradient images, obtained with different values of S and O. As seen in Fig. 10, the mean distance for all limbs is maximized for $O = 45$ and $S = 0.05$. The maximized Bhattacharyya distance and Kullback-Leibler divergence are compared to the distances for other kinds of normalization in Table 1.

In Fig. 11, the log likelihood ratios for all limbs, using local contrast normalization with the optimal values

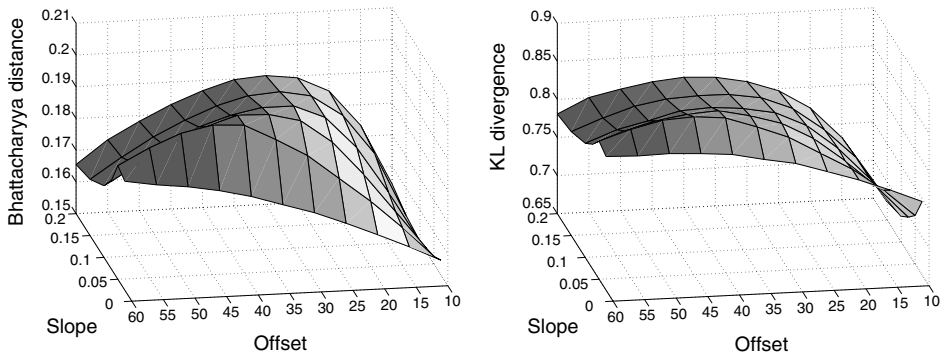
Table 1. Comparison of Bhattacharyya distance and Kullback-Leibler divergence with different types of normalization. Local contrast normalization proves the best at maximizing the difference between foreground and background distributions.

	Bhattacharyya distance				Kullback-Leibler divergence			
Limb	Thigh	Calf	U arm	L arm	Thigh	Calf	U arm	L arm
No normalization	0.15	0.20	0.14	0.20	0.71	0.84	0.63	0.83
Local normalization	0.16	0.22	0.16	0.22	0.80	0.96	0.74	0.97
Global normalization	0.13	0.15	0.11	0.15	0.60	0.68	0.46	0.63



(a) No contr. norm. (b) Local contr. norm. (c) Global contr. norm.

Figure 9. *Image gradient in vertical direction*, comparison between different normalization techniques.



(a) (b)

Figure 10. *Finding optimal normalization parameters.* The plots show the Bhattacharyya distance (a) and Kullback-Leibler divergence (b) for different values of offset O and slope S, averaged over all limbs. For both measures, the distance between p^e_{on} and p^e_{off} is maximized when $O = 45$ and $S = 0.05$.

of S and O, are shown. Note that the shape differs from the shape of the unnormalized ratios shown in Fig. 7 due to the nonlinear transfer function H.

Global Contrast Normalization. We also test the global contrast normalization used by Lee et al. (2001) and Ruderman (1994). As opposed to the local normalization technique this global method normalizes the contrast in the whole image instead of local areas. Before computing filter responses, the image intensities are normalized as

$$I_{\text{norm}} = \log\left(\frac{I}{\hat{I}}\right)$$

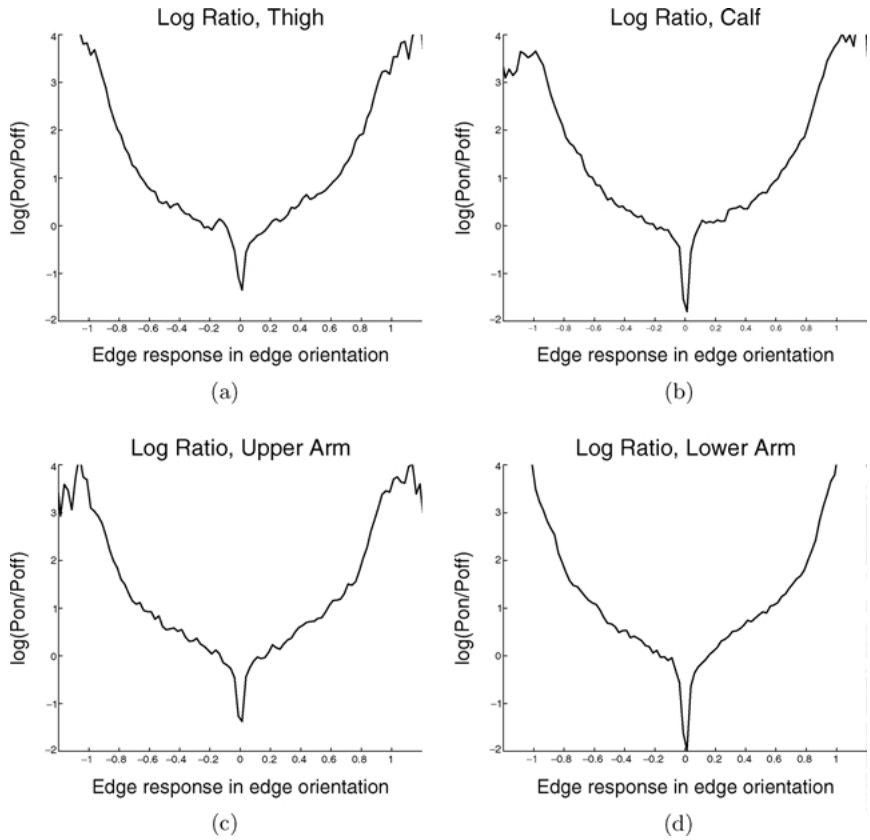


Figure 11. Local contrast normalization. Learned log likelihood ratios for edge response.

where $\hat{I}$ is the mean image intensity in the image I. I_{norm} can be interpreted as representing deviations from the image mean intensity.

The Bhattacharyya distance and Kullback-Leibler divergence of distributions using this normalization scheme are listed in Table 1. Given the greater distance between foreground and background using local contrast normalization, all further analysis below will use the local contrast normalization scheme.

3.2. Ridge Cue

In the same spirit as with edges, we use the response of second derivatives filters steered to the predicted orientation of the limb axis. These filter responses, f_r, are a function of $[f_{xx}, f_{xy}, f_{yy}]$, the second derivatives of the image brightness function in the horizontal and vertical directions.

Following Lindeberg (1998), we define ridge response as the second derivative of the image perpendicular to the ridge ($|f_{\theta\theta}|$), minus the second derivative parallel to the ridge ($|f_{(\theta-\frac{\pi}{2})(\theta-\frac{\pi}{2})}|$). This will suppress 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:

$$\begin{aligned} f_r(\mathbf{x}, \theta, \sigma) = |&\sin^2\theta\, f_{xx}(\mathbf{x}, \sigma) + \cos^2\theta\, f_{yy}(\mathbf{x}, \sigma) \\ &- 2\sin\theta\, \cos\theta\, f_{xy}(\mathbf{x}, \sigma)| \\ &- |\cos^2\theta\, f_{xx}(\mathbf{x}, \sigma) + \sin^2\theta\, f_{yy}(\mathbf{x}, \sigma) \\ &+ 2\,\sin\theta\, \cos\theta\, f_{xy}(\mathbf{x}, \sigma)|. \end{aligned} \tag{6}$$

Figure 12 shows an example of a steered ridge response for a lower arm.

3.2.1. Relation Between Limb Width and Image Scale. Since ridges are highly dependent on the size of the limb in the image we do not expect a strong filter response at scales other than the one corresponding to the projected width of the limb. In training, we therefore only consider scales corresponding to the distance between the manually marked edges of the limb.

(a) Image with limb model overlayed (b) Ridge response f_r, orientation θ of lower arm, level $\sigma = 3$ (c) Log likelihood ratio $b^r(f_r)$, orientation θ of lower arm, level $\sigma = 3$

Figure 12. Computation of steered ridge response. The original image with the overlayed model is shown in (a), while (b) shows the ridge response $f_r(\mathbf{x}, \theta, \sigma)$ for the lower arm with angle θ and pyramid level $\sigma = 3$ (for scale selection see Section 3.2.1). White denotes strong positive ridge response, black strong negative response, grey weak response. The corresponding log likelihood ratio $b^r(f_r)$ for every image location is shown in (c). White denotes high (positive) likelihood ratio, black low (negative) likelihood ratio.

To determine the relationship between image scale and width of limb in the image, a dense scale-space is constructed from each image in the training set. Scale s is constructed by convolving the image at scale $s-1$ with a 5×5 window, approximating a Gaussian of variance 1. Scale 0 is the original image. For each limb, N points within the limb area (determined by the hand-marked edges) are selected, and the ridge response according to Eq. (6) is computed for each point. The sum of these responses is a measure of how visible the limb ridge is.

To be able to compare ridge response at different scales, normalized derivatives (Lindeberg, 1998) are used to compute the filter responses. The normalized filters are denoted $f^s_{xx} = s^{2\gamma} f_{xx}$, $f^s_{xy} = s^{2\gamma} f_{xy}$ and $f^s_{yy} = s^{2\gamma} f_{yy}$, where s is the scale,[2] and $\gamma = 3/4$, which is optimal for ridge response detection (Lindeberg, 1998). The scale corresponding to the maximum (normalized) filter response is found for each limb. If the maximum response is above a certain level, the tuple (limb width, scale) is saved. The image scale with the maximal response is plotted as a function of projected limb diameter in Fig. 13.

We can assume that the function relating limb width and scale is linear, since the scale can be viewed as a length measure in the image—a linear function of the radius or length of the structures visible at that scale. We can also assume that the slope of the linear function is positive—that larger limbs are visible on coarser scales. With these restrictions, a straight line is fitted to the measured limb-width-scale tuples using RANSAC (Fischler and Bolles, 1981). The linear relationship is shown in Fig. 13 and is given by

$$s = -24 + 4.45\, w \tag{7}$$

where w is the limb width and s the image scale.[3]

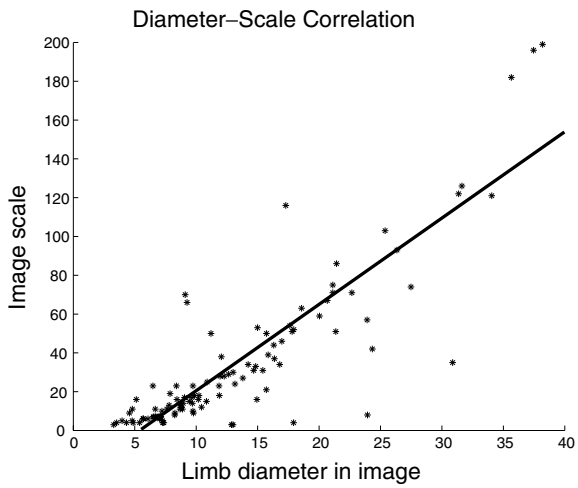


Figure 13. Relationship between image scale s of the maximal ridge response and limb diameter w. The scale of the maximal filter response is plotted versus the limb diameter in image pixels. Best linear fit: $s = -24 + 4.45\, w$.

For Bayesian tracking we use a more compact image pyramid rather than the full scale space. This lowers the complexity of the image processing operations, but also has the drawback that the scale resolution is limited. Since each level of the pyramid is a sub-sampled version of the level below the scales s in the dense scale-space relate to levels σ in the pyramid by

$$s = \begin{cases} 0 & \text{if } \sigma = 0 \\ \sum_{i=1}^{\sigma} 4^{i-1} & \text{otherwise} \end{cases} \tag{8}$$

The appropriate pyramid level σ for a certain limb width w is computed using Eqs. (7) and (8).

3.2.2. Learning Foreground and Background Distributions. For each of the images in the training set (Fig. 4), the ridge orientation θ_l and ridge pyramid level σ_l (Eqs. (7) and (8)) of each limb l are computed from the manually marked edges. Then, a set of locations $\mathbf{x}_i$ are sampled on the area spanned by the marked limb edges, at level σ_l with $\theta = \theta_l$. For each limb l and each level σ, we construct a separate discrete probability distribution of steered ridge responses $f_r(\mathbf{x}, \theta_l, \sigma_l)$, for the sampled foreground locations $\mathbf{x}_i$. The normalized empirical distributions represent $p^r_{\text{on}}(f_r \mid l, \sigma)$, which is analogous to $p^e_{\text{on}}(f_e \mid l, \sigma)$ described above. Figure 14(a) shows the logarithm of p^r_{on} for the thigh, at pyramid levels 2, 3 and 4.

Proceeding analogously to the learning of edge background distribution, we learn a distribution $p^r_{\text{off}}(f_r \mid \sigma)$, the probability distribution over ridge responses in general scenes, conditioned on pyramid level. For a certain response $f_r(\mathbf{x}, \theta, \sigma)$, $p^r_{\text{off}}(f_r(\mathbf{x}, \theta, \sigma) \mid \sigma)$ is the probability that $\mathbf{x}$ is explained by the background. Figure 14(b) shows the logarithm of p^r_{off} for pyramid levels 2, 3 and 4. As with edges, we draw the conclusion from Fig. 14 that the distributions over ridge response in general scenes are invariant over scale and thus represent likelihoods at all levels by integrating out the scale variable (Fig. 14(d)).

As with edges, the likelihood ratios are computed from the foreground and background distributions (Fig. 15). The ratio is roughly linear; the greater the response, the more likely it is to have come from a human limb. Furthermore, responses close to zero are very unlikely to come from a human limb.

3.3. Motion Cue

Human motion gives rise to predictable changes in the projected brightness pattern in the image. The motion cue used here is based on the temporal brightness derivative, $f_{m,t}$, of the image at time t. Given the change

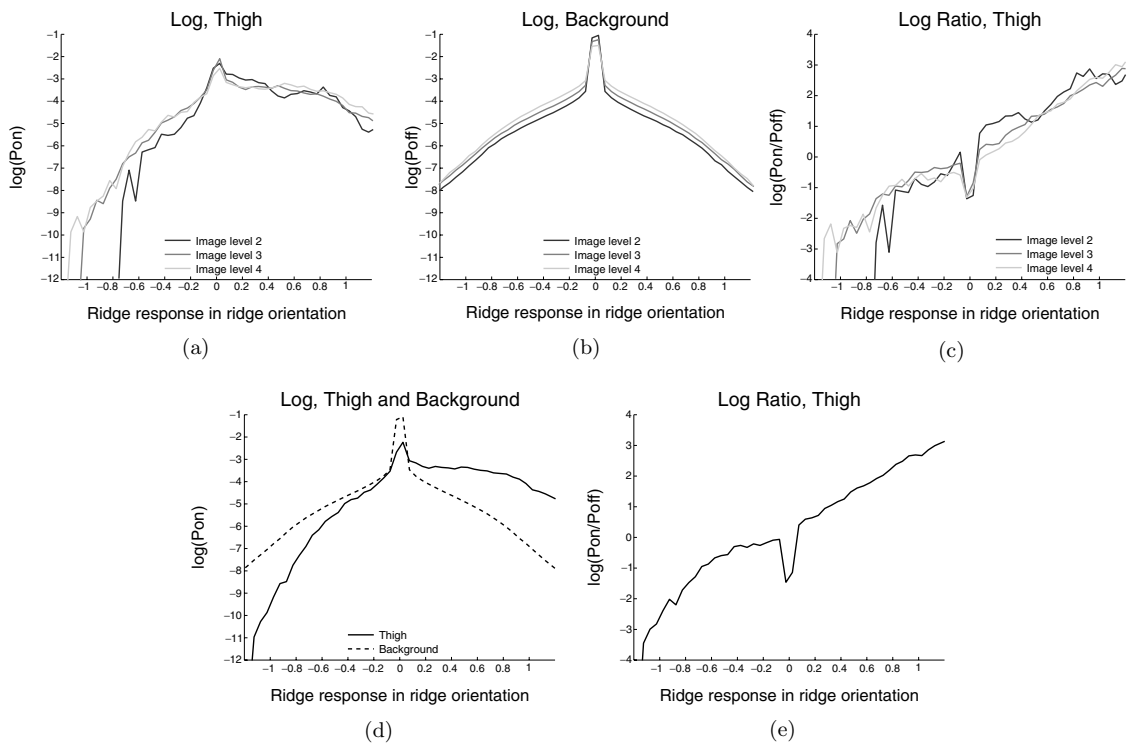


Figure 14. Foreground and Background distributions (local contrast normalization). The distribution for Thigh is shown. For foreground distributions, only the level corresponding to the width (Fig. 13) of the limb in each training example is considered. (a) The thigh log likelihood distributions for different image levels. (b) The background log likelihood distributions for the same levels. (c) The log likelihood ratio for different levels. (d) The marginals over pyramid level. (e) Final log likelihood ratio. The shape of the distributions are slightly different from the shape of the unnormalized distributions. This is due to the non-linear normalization function.

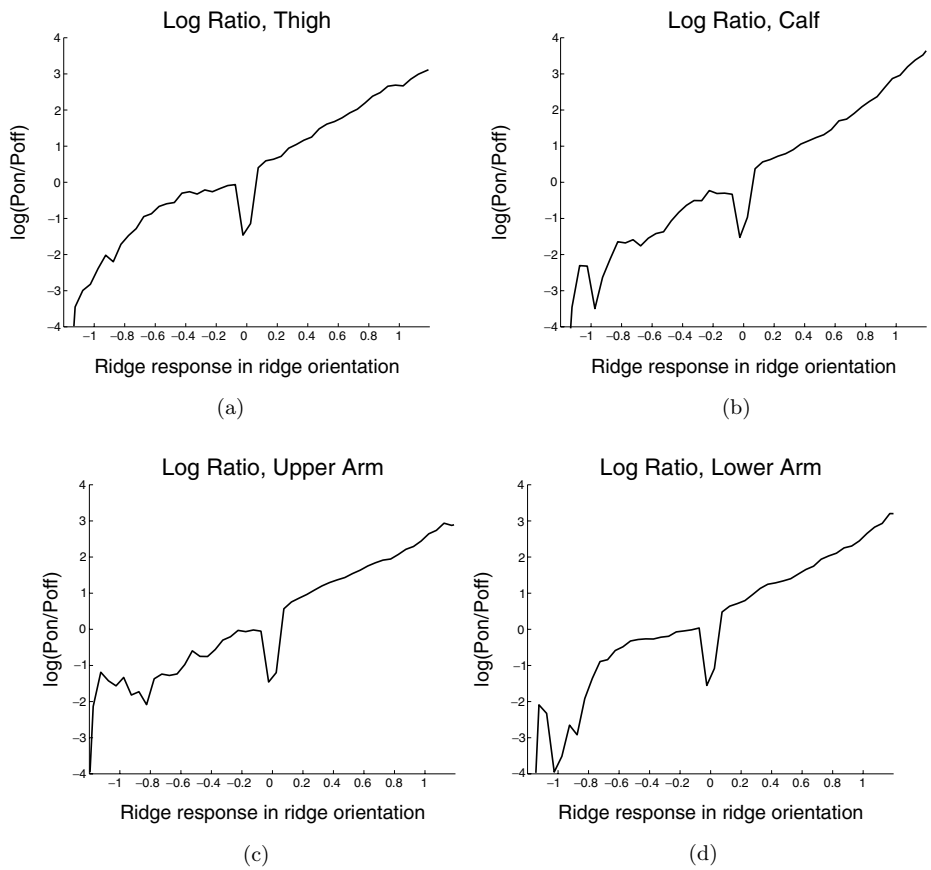


Figure 15. Learned log likelihood ratios for ridge response. Local contrast normalization.

in 3D pose of the body between time $t-1$ and t, the 2D displacement of limb regions in the image can be computed. This is used to register (or warp) the image at time $t-1$ towards the image at time t. If the 3D motion is correct, the magnitude of the temporal differences between the registered limb regions should be small.

Let $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, correspond to the same limb surface location at time $t-1$ and t respectively; the motion response at time t and pyramid level σ is then formulated as:

$$f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma) = I_t(\mathbf{x}_t, \sigma) - I_{t-1}(\mathbf{x}_{t-1}, \sigma). \quad (9)$$

Note that this response function is only valid for positions $\mathbf{x}_t$ on the foreground (limb area).

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. In the case of moving background, there will be large responses in textured areas or around edges, but not in homogeneous regions. Furthermore, all responses from static backgrounds will be low. If the motion of the background was modeled, this could be used to compute a warping between images at time $t-1$ and t, in the same way as for the foreground model. This would help to better explain image changes, and discriminate between foreground and background.

3.3.1. Learning Foreground and Background Distributions. The probability distributions over motion response in the foreground and the background, p^m_{on} and p^m_{off}, are learned from a set of short sequences in the training set, with hand-marked limb locations at each frame. Two of the sequences contain cluttered scenes shot with a moving camera, two contain outdoor scenes with moving foliage, and all scenes contain moving humans. Due to the difficulty of obtaining "ground truth" motions, the training set for the motion distributions is more limited than for edges and ridges.

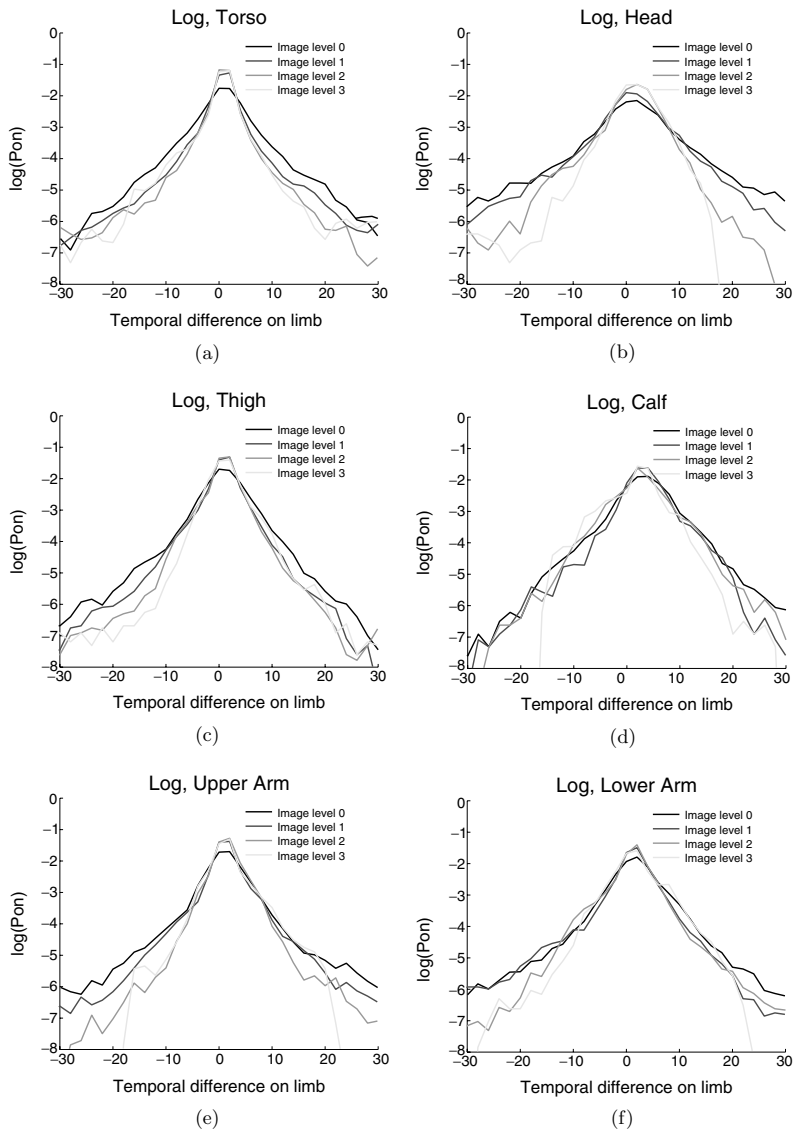


Figure 16. Learned log likelihood distributions for foreground pixel difference given model flow between two consecutive frames.

For each pair of consecutive frames, and for each limb of the human, the limb area at the first frame is warped to align it with the same area in the second frame. The difference between the two areas is computed, and a number of image locations $\mathbf{x}_i$ are randomly selected from the area. The differences are collected into a normalized histogram representing $p^m_{\text{on}}(f_m \mid l, \sigma)$, the probability distribution over motion response f_m, conditioned on limb number l and pyramid level σ, given that the motion of the limb model explains the image change. Given a certain observed response $f_{m,t}(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma)$, the likelihood of observing this response on limb l is $p^m_{\text{on}}(f_m(\mathbf{x}_{t-1}, \mathbf{x}_t, \sigma) \mid l, \sigma)$. Figure 16 shows the logarithm of p^m_{on} for all limbs, at pyramid levels 0, 1, 2 and 3.

For each pair of consecutive frames, the un-warped difference image is also computed. A number of image locations $\mathbf{x}_i$ are chosen uniformly over the difference image, and the differences are collected into a normalized histogram representing $p^m_{\text{off}}(f_m \mid \sigma)$, the probability distribution over motion response f_m, conditioned on pyramid level σ, given that the image change is explained by a static background. Given a certain observed response $f_m(\mathbf{x}_t, \mathbf{x}_t, \sigma)$, the

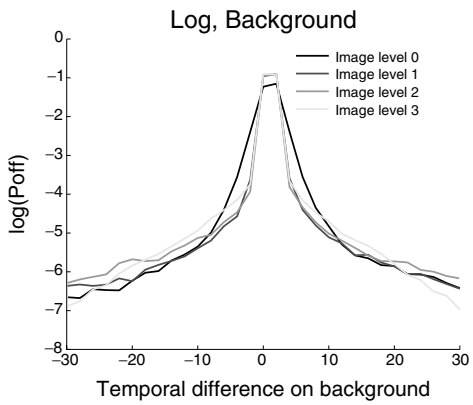


Figure 17. Learned log likelihood distributions for background pixel difference given model flow between two consecutive frames. Note that the distributions assume no motion in the background. The heavy tails are due to violations of the brightness constancy assumption.

likelihood of observing this response in the background is $p^m_{\text{off}}(f_m(\mathbf{x}_t, \mathbf{x}_t, \sigma) \mid \sigma)$. Figure 17 shows the logarithm of p^m_{off}, at pyramid levels 0, 1, 2 and 3.

The distributions appear to be largely scale-independent yet, given the limited size of the motion training set, conclusions about the scale-independence of the temporal differences would be premature. This remains an open question for further research.

It is worth noting that the distributions over temporal differences can be approximated analytically (see Sidenbladh (2001) for more details). In particular, the heavy-tailed nature of these distributions can be well approximated by a Cauchy or t-distribution. This provides an interesting connection with work on robust optical flow estimation (Black and Anandan, 1996) where violations of the brightness constancy assumption are dealt with using a robust error term. The common robust error functions have analogous heavy-tailed distributions when adopting a probabilistic interpretation. The success of robust optical flow methods may well be explained by the fact that the ad hoc robust error terms are precisely the appropriate functions for dealing with the actual distribution of brightness differences in natural images.

4. Using the Filter Distributions

This section presents the formulation of the probabilistic framework in which the filter distributions are employed. The human is modeled as an articulated assembly of limbs with the appearance of each limb considered conditionally independent of the others. The configuration of the limbs is represented by a set of joint angle parameters ϕ. Without loss of generality, below we consider the appearance of a single limb.

4.1. Likelihood Formulation

Tracking is viewed in a Bayesian framework as the problem of estimating the posterior probability, $p(\phi \mid \mathbf{f})$, that the body has a pose ϕ given the observed filter responses $\mathbf{f}$. By Bayes' rule, the posterior distribution can be written as

$$p(\phi \mid \mathbf{f}) = \kappa_1 \, p(\mathbf{f} \mid \phi) \, p(\phi) \tag{10}$$

where κ_1 is a constant independent of ϕ, $p(\mathbf{f} \mid \phi)$ is the likelihood of $\mathbf{f}$ given ϕ, and $p(\phi)$ the prior distribution over ϕ. Here we do not address the prior distribution; for examples of generic and event-specific priors, the reader is referred to Ormoneit et al. (2001) and Sidenbladh et al. (2000a, 2002).

4.1.1. Combining Responses at Different Pixels. Pixels in the image belong either to the background or the foreground (person). The body pose parameters, ϕ determine $\{\mathbf{x}_f\}$, the set of image locations corresponding to the foreground. Let the set of background pixels be $\{\mathbf{x}_b\} = \{\mathbf{x}\} - \{\mathbf{x}_f\}$, where $\{\mathbf{x}\}$ is the set of all pixels.[4] Let $p(\mathbf{f} \mid \phi)$ be the likelihood of observing filter responses $\mathbf{f}$ given the parameters, ϕ, of the foreground object (e.g., the joint angles of a human body model). 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

$$\begin{aligned} p(\mathbf{f} \mid \phi) &= \prod_{\mathbf{x} \in \{\mathbf{x}_b\}} p_{\text{off}}(\mathbf{f}(\mathbf{x})) \prod_{\mathbf{x} \in \{\mathbf{x}_f\}} p_{\text{on}}(\mathbf{f}(\mathbf{x}, \phi)) \\ &= \frac{\prod_{\mathbf{x} \in \{\mathbf{x}\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))}{\prod_{\mathbf{x} \in \{\mathbf{x}_f\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))} \prod_{\mathbf{x} \in \{\mathbf{x}_f\}} p_{\text{on}}(\mathbf{f}(\mathbf{x}, \phi)) \end{aligned} \tag{11}$$

since $\{\mathbf{x}_b\} = \{\mathbf{x}\} - \{\mathbf{x}_f\}$.

Note that $\prod_{\mathbf{x} \in \{\mathbf{x}\}} p_{\text{off}}(\mathbf{f}(\mathbf{x}))$ is independent of ϕ; we call this constant term κ_2 and simplify the likelihood as

$$p(\mathbf{f} \mid \phi) = \kappa_2 \prod_{\mathbf{x} \in \{\mathbf{x}_f\}} \frac{p_{\text{on}}(\mathbf{f}(\mathbf{x}, \phi))}{p_{\text{off}}(\mathbf{f}(\mathbf{x}))} \, . \tag{12}$$

This is the normalized ratio of the likelihood that the foreground pixels are explained by the person model

versus that the same pixels are explained by a generic background model. Note that this is simply a scaled version of the likelihood ratio plotted throughout the paper (Eq. (1)).

4.1.2. Combining Cues. We assume the responses for edges, ridges and motion can be considered independent. This means that the likelihood can be formulated as

$$p(\mathbf{f} \mid \phi) = p(f_e \mid \phi)\, p(f_r \mid \phi)\, p(f_m \mid \phi). \tag{13}$$

4.1.3. Combining Responses over Scale. Responses for edges and motion can be observed at several levels σ in the image pyramid. We model the responses at different levels as uncorrelated. This is a simplified model of the world, in reality there exists a high degree of correlation. The effect of treating the levels as uncorrelated is that the combined probability will take the same information into regard more than once, which will make the distribution more "peaked"; correlation across scale requires further study.

The motivation for combining edge and motion response over scales, is that high response from the true limb location will be present at all scales, while "false" maxima due to image noise are unlikely to appear at all scales. The real maximum is thus enforced by combination over scales.

With the independence assumption, the likelihoods for edge and motion are

$$p(f_e \mid \phi) = \prod_{\sigma=0}^{n} p(f_e(\sigma) \mid \phi) \tag{14}$$

$$p(f_m \mid \phi) = \prod_{\sigma=0}^{n} p(f_m(\sigma) \mid \phi) \tag{15}$$

where n is the highest level in the pyramid.

4.1.4. Learned Likelihood Ratios. The effect of treating filter responses from different cues and different scales as independent is that $p_{\text{on}}(\mathbf{f})$ is the product of foreground likelihoods for all cues and scales, and, equivalently, $p_{\text{off}}(\mathbf{f})$ is the product of all background likelihoods. Thus, Eqs. (12), (13) and (15) give

$$p(f_e \mid \phi) = \kappa_2^e \prod_{\sigma=0}^{n} \prod_{\mathbf{x}\in\{\mathbf{x}_e\}} \frac{p_{\text{on}}^e(f_e(\mathbf{x}, \theta(\phi), \sigma))}{p_{\text{off}}^e(f_e(\mathbf{x}, \theta(\phi), \sigma))} \tag{16}$$

$$p(f_r \mid \phi) = \kappa_2^r \prod_{\mathbf{x}\in\{\mathbf{x}_r\}} \frac{p_{\text{on}}^r(f_r(\mathbf{x}, \theta(\phi), \sigma(\phi)))}{p_{\text{off}}^r(f_r(\mathbf{x}, \theta(\phi), \sigma(\phi)))} \tag{17}$$

$$p(f_{m,t} \mid \phi_t) = \kappa_2^m \prod_{\sigma=0}^{n} \prod_{\mathbf{x}_t\in\{\mathbf{x}_{m,t}\}} \frac{p_{\text{on}}^m(f_{m,t}(\mathbf{x}_{t-1}(\mathbf{x}_t, \phi_t), \mathbf{x}_t, \sigma))}{p_{\text{off}}^m(f_{m,t}(\mathbf{x}_t, \mathbf{x}_t, \sigma))} \tag{18}$$

where $\kappa_2^{\{e,r,m\}}$ are normalizing constants such that $\kappa_2 = \kappa_2^e \kappa_2^r \kappa_2^m$, $n = 3$ scales in our experiments, 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 $\{\mathbf{x}_f\}$.[5] Note that the cardinalities of the sets for each feature define an implicit weighting of the likelihood terms of each cue.

5. Experimental Results

Two different experiments described below illustrate the learned likelihood model.

5.1. Studying the Likelihood for One Limb

To illustrate the discriminative power of the likelihood measure, we plot the log of the ratio of unnormalized likelihoods for one limb (the lower arm) as the predicted limb location is displaced spatially from its correct position. The orientation of the limb in the experiments below is held constant; therefore, the filter responses f_e and f_r can be pre-computed. Figures 18 and 19 show the pre-computed filter images for f_e and f_r, the edge and ridge response in the orientation of the limb, for the three images used in the experiments.

In this experiment, the true edges of the lower arm are manually determined. The positions of the edges are then varied vertically, maintaining the relative distance between the model edges. For each position, the likelihood is computed. If the likelihood discriminates well between actual limbs and general background, there should be a peak around translation 0. Thus the variation in likelihood as a function of translation provides insight into the robustness and precision of the likelihoods for different cues and combinations of cues.

5.1.1. Edge Likelihood. In Fig. 20 the unnormalized edge log likelihood ratio (Eq. (16)) over vertical translation for three different images is shown. In Fig. 20(a) and (c), there is a clear maximum at translation

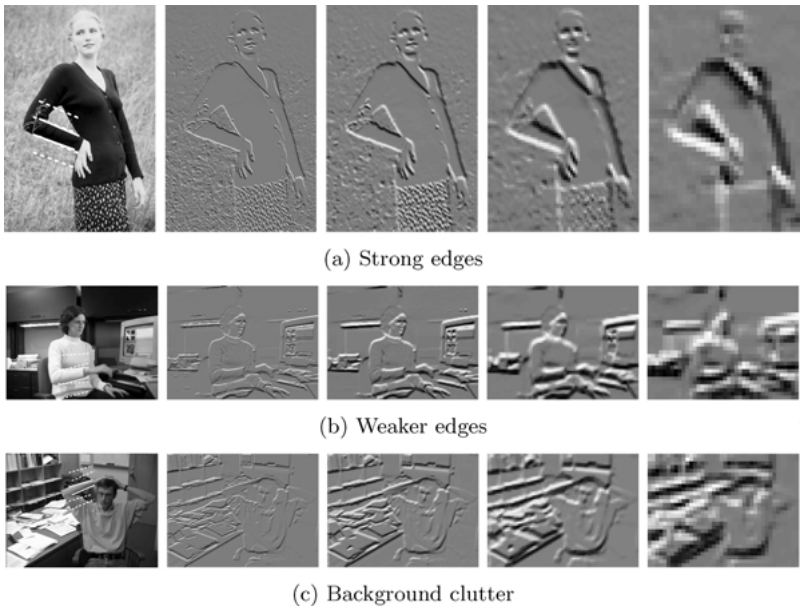


Figure 18. Edge response in the lower arm orientation, θ, at image pyramid level $\sigma = 0, 1, 2, 3$.

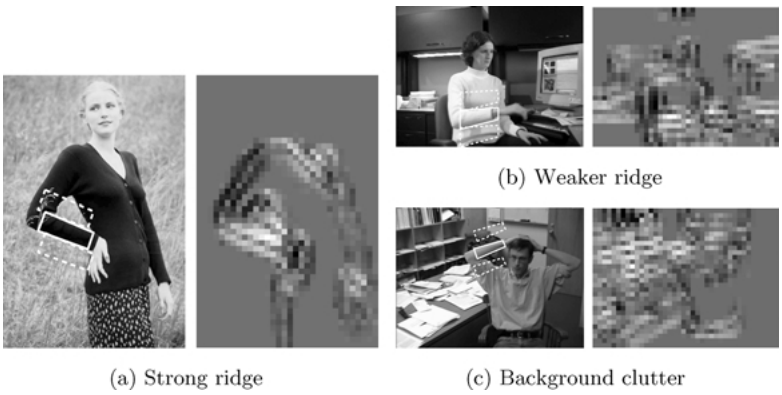


Figure 19. Ridge response in lower arm orientation, θ, at the pyramid level corresponding to the lower arm size ($\sigma = 3$ in all three cases).

error 0; this means that the edge term discriminates well between limb edges and general background in these two images. In Fig. 20(b), there are strong local maxima at translation error -11 and 18. At these translations, the model encounters edges (wrinkles on the shirt and shadow boundaries) that the model takes for limb edges. The effect of aliasing is also clearly visible in all the plots as the lower edge of the arm matches the upper edge and vice versa.

The distribution in case b is multi-modal, there are three large peaks, two "false" and one "true". If this distribution were the basis for temporal propagation of the limb configuration in a Bayesian tracker, this would cause problems if the multiple maxima were not taken into account. A uni-model tracker that maintains a maximum a posteriori estimate would not well represent the inherent uncertainties present in the likelihood distribution. This suggests that a tracking scheme that models the whole distribution, such as particle filtering (e.g., CONDENSATION (Isard and Blake, 1998; Sidenbladh et al., 2000a; Sidenbladh and Black, 2001)) is more appropriate and may lead to more robust tracking.

We can conclude from this experiment that the edge cue by itself provides a strong but not sufficient cue for discriminating between limbs and background, and

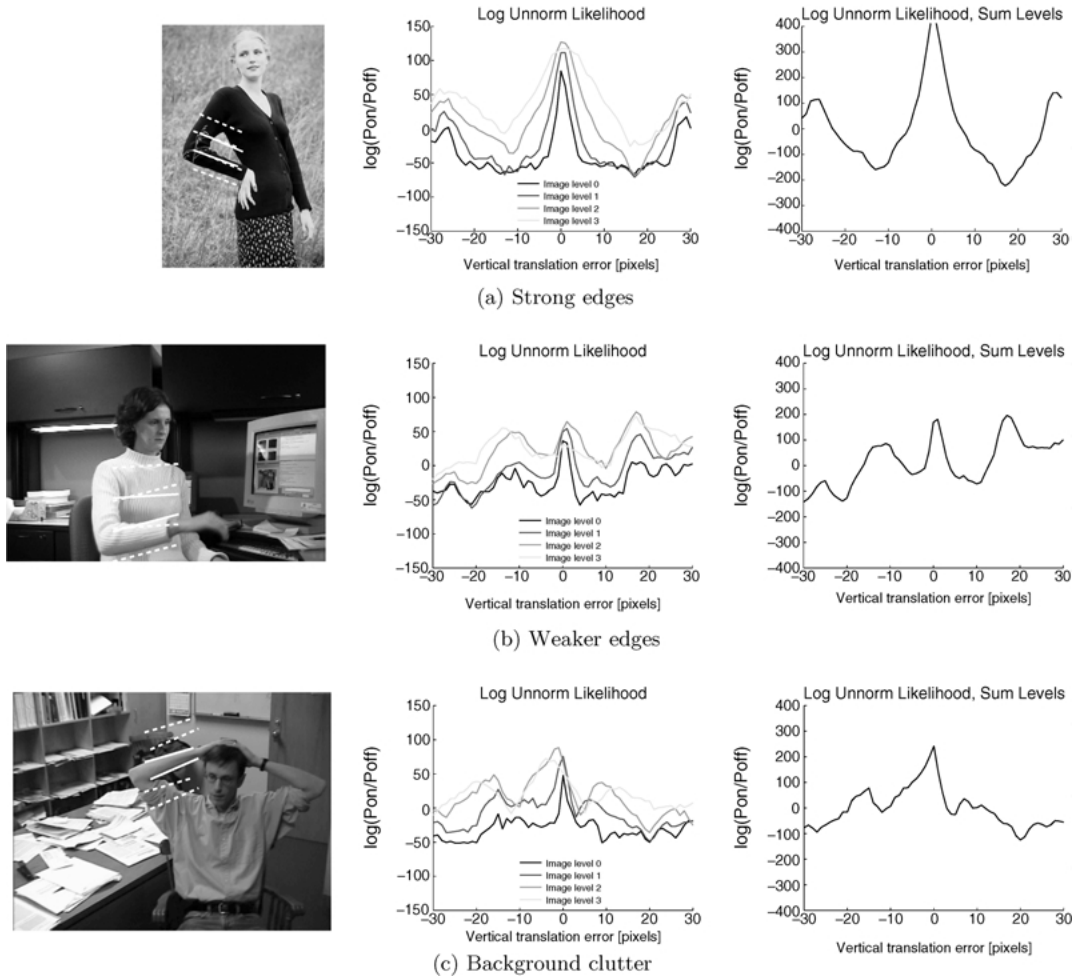


Figure 20. Edge cue: Lower arm log likelihood as a function of displacement. The original image with the correct edges (solid), and the two translation extrema (dashed) is shown left. The left plot shows the likelihoods w.r.t. vertical displacement for each pyramid level separately, while the right plot shows the sum of log likelihoods for different pyramid levels.

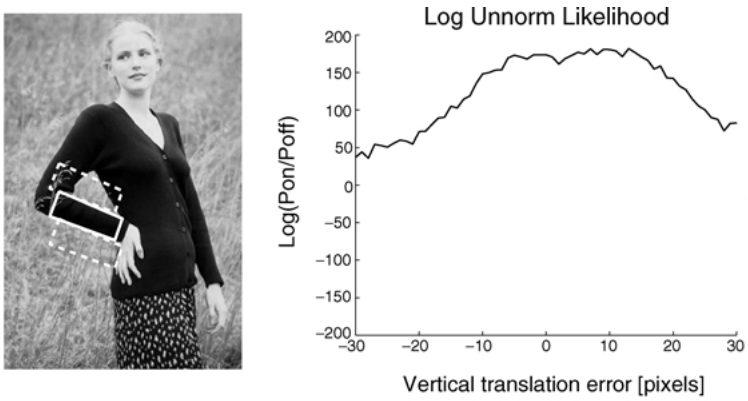
that multi-modal distributions occur in tracking and detection of human limbs.

5.1.2. Ridge Likelihood. The experiment is repeated for the ridge likelihood (Eq. (17)) and the results are displayed in Fig. 21. The ridge likelihood varies much more smoothly as a function of translation than does the edge likelihood. This means that a limb ridge is "visible" from a larger spatial displacement. Furthermore, there are fewer false maxima than in the edge experiment.

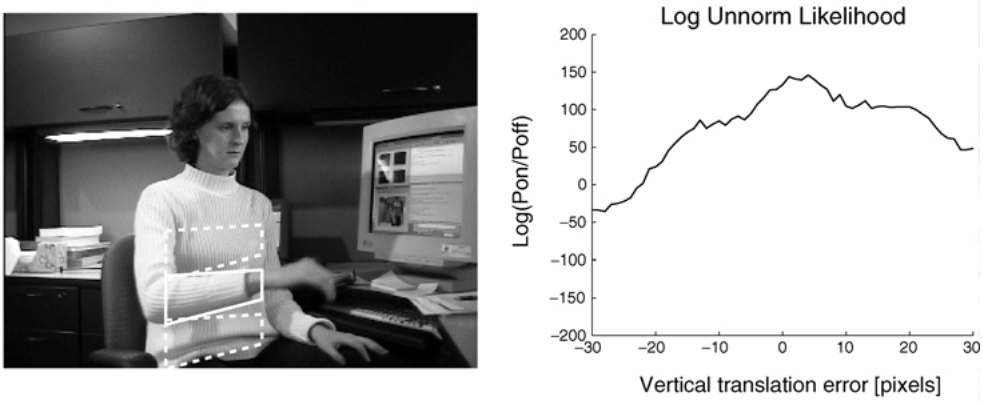
When the likelihoods from the two cues are combined, the ridge cue will suppress the false maxima from the edge cue, while the edge cue will help to discriminate between slightly misplaced limb locations and correct ones.

5.1.3. Motion Likelihood. We also test the effect of displacement on the motion response likelihood (Eq. (18)). Given the correct location of the limb at time $t-1$, the position at time t is varied as in the two previous experiments (Fig. 22(a)). There is a clear peak at 0, as expected. It is broader than the peak for edge likelihood, but there are no false maxima.

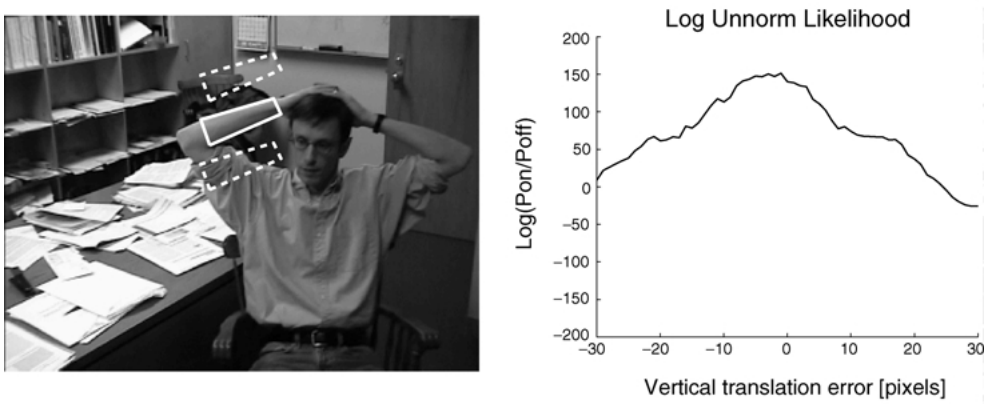
To see how drift effects the cue, in the next experiment, the position at time $t-1$ is chosen to be incorrect; the initial limb model is moved five pixels in the negative vertical direction (up) from its correct



(a) Strong ridge. Image level 3 chosen for ridge response evaluation.



(b) Weaker ridge. Imagelevel 3 chosen for ridge response evaluation.



(c) Background clutter. Image level 3 chosen for ridge response evaluation.

Figure 21. *Ridge cue:* Lower arm log likelihood as a function of displacement. The original image with the correct limb area (solid), and the two translation extrema (dashed) is shown left. The plot shows the likelihood w.r.t. vertical displacement, using the filter images at the pyramid level that corresponds to the limb width according to Eqs. (7) and (8).

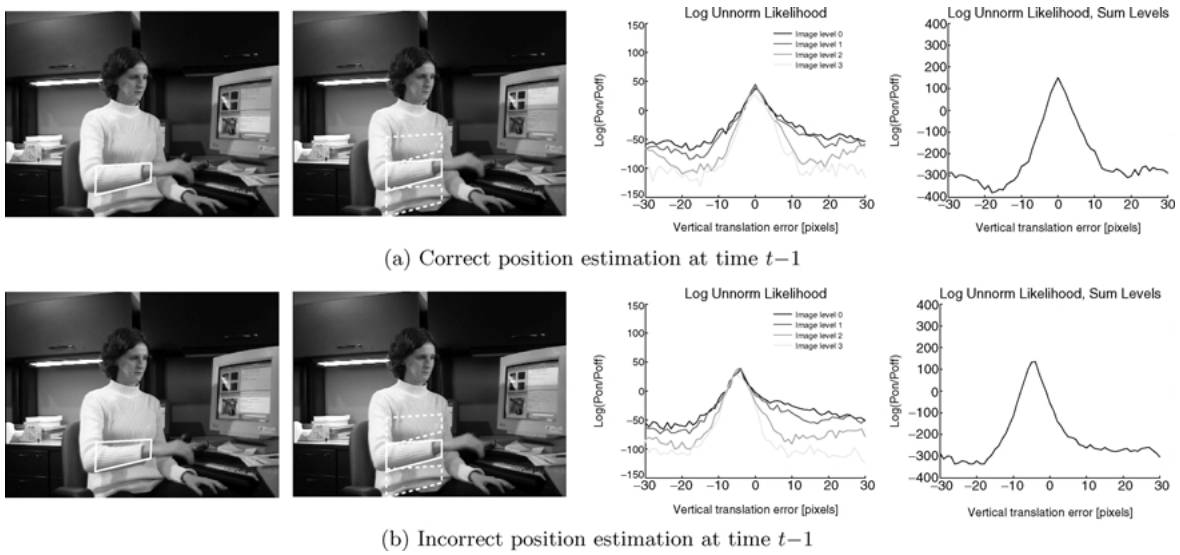


Figure 22. Motion cue: Log likelihood as a function of displacement. The images at time $t-1$ and t, overlayed with the correct limb area (solid), and the two translation extrema (dashed), are shown left. The left plot show the likelihoods w.r.t. vertical displacement, while the right plot show the sum of log likelihoods for different levels. In (a) the limb area at time $t-1$ is correctly estimated, while it is translated 5 pixels in (b).

position. Consequently, the peak in the likelihood moves five pixels in negative vertical direction (Fig. 22(b)). This is expected, since the pattern on the limb model at time $t-1$ corresponds best to this location at time t. This means that the tracking using only the motion cue will generally not recover from errors since the cue is relative and, hence, prone to "drift".

5.1.4. Combining the Cues. The likelihood with multiple cues is achieved by by summing the log likelihoods from the different cues (Eq. (13)). In Fig. 23(a) the effect of vertical translation, using combined edge, flow and motion likelihood with the correct limb position at time $t-1$, is shown. The combination of the edge cue results in a sharper peak than in the case of the motion cue alone (shown in the left plot in a). The false maxima of the edge cue are suppressed by the motion cue and the ridge cue, while the true maximum is present with all three cues. This means that the motion and ridge cues can make the tracking less prone to tracking incorrect parallel edges that happen to look like limbs.

Even when the position at time $t-1$ is wrongly predicted (Fig. 23(b)) the combined graph has a maximum at displacement 0, due to the edge cue. This means that the edge cue can help the tracking recover from the accumulation of errors that can result from the drift of the motion cue.

These experiments suggest that tracking can benefit from likelihood measures using multiple cues, since the cues have different properties and are effected by different kinds of noise (cf. Rasmussen and Hager (2001)).

5.2. Tracking an Arm

The likelihood is now tested as part of a particle filtering tracking framework (Sidenbladh and Black, 2001; Sidenbladh et al., 2000a). The human is modeled as a 3D assembly of truncated cones. The configuration of the cones at each time step t are determined by the parameters ϕ_t. For the experiments here we consider a simplified body model representing only the torso and the right arm. The configuration ϕ_t includes the arm angles, the global torso position and rotation, and their respective velocities. The particle filtering framework represents the posterior probability over the possible configurations with a discrete set of sample configurations and their associated normalized likelihoods. Here $N = 5000$ hypotheses (samples, or particles), ϕ_t^s, $s = 1 \ldots N$, are maintained, and are propagated in time with a linear motion model (see Sidenbladh and Black (2001) for details).

The likelihood of the image cues (filter responses) conditioned on sample ϕ_t^s is evaluated as $p(\mathbf{f} \mid \phi^s)$,

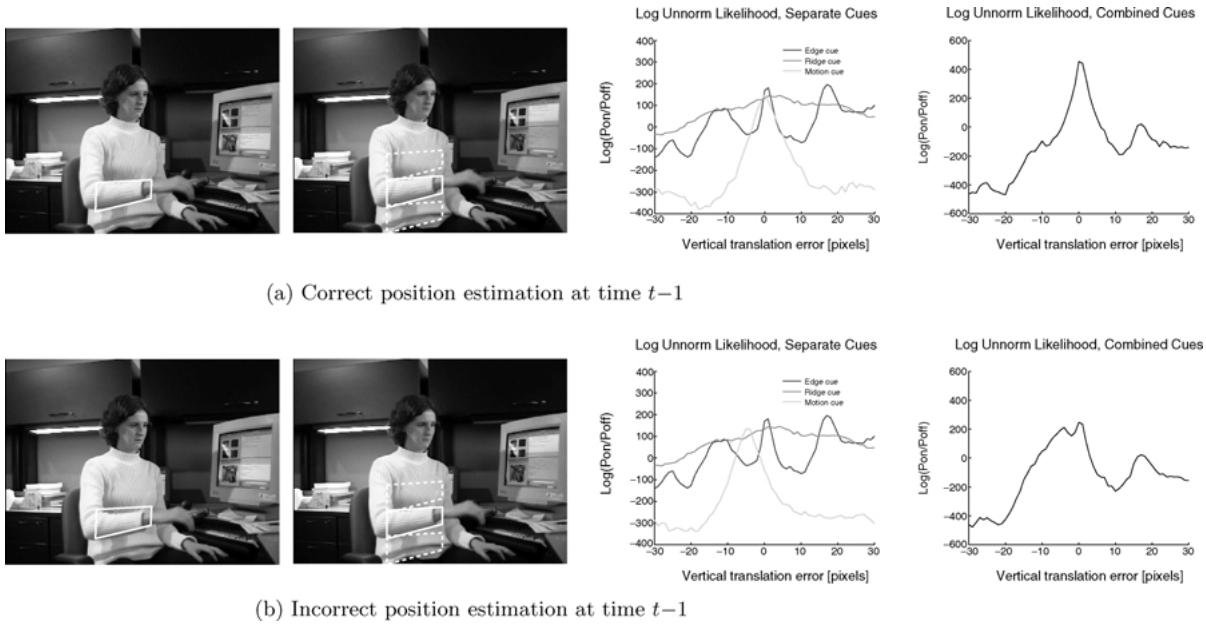


Figure 23. *Multiple cues:* Log likelihood as a function of displacement. The likelihood responses w.r.t. edges, ridges and motion are assumed to be independent. Thus, the log likelihood for all cues are summed. The left graph shows the likelihood w.r.t. vertical displacement separately, the right graph shows the combined likelihood. In the combined likelihood, there is a maximum at displacement 0, both in the case when the initial position at time $t-1$ is correct (a) and incorrectly displaced (b).

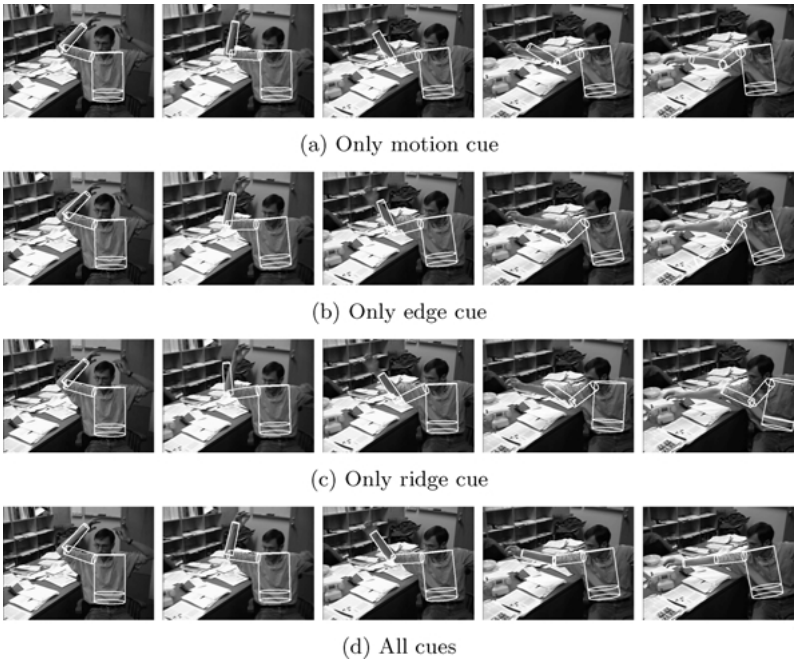


Figure 24. *Tracking an arm, moving camera, 5000 samples.* The sub-figures show frames 10, 20, 30, 40 and 50 of the sequence. In each frame, the expected value from the posterior distribution over ϕ is projected into the image. (a) Only motion cue. (b) Only edge cue. (c) Only ridge cue. (d) All cues.

according to Eqs. (11)–(18). It should be noted that, in Fig. 23, the difference between the highest and lowest log likelihood is large; this means that the actual probability distribution approaches a delta function. A particle filter tracking system using this distribution could be very brittle, since finding such a sharp peak with discrete particles is difficult. To overcome this problem, a re-sampling approach is used (Sidenbladh and Black, 2001) that essentially smoothes the likelihood and damps the highest peaks.

Figure 24 shows four different tracking results for a sequence of a cluttered scene containing both human motion and camera motion. The model is initialized with a Gaussian distribution around a manually selected set of start parameters ϕ_0. Camera translation during the sequence causes motion of both the foreground and the background.

Figure 24(a) shows tracking results using only the motion cue. Generally, motion is an effective cue for tracking, however, in this example, the 3D structure is incorrectly estimated due to drift. The edge cue (Fig. 24(b)), 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 (Fig. 24(c)) and the model has too little information to track the arm properly.

Figure 24(d) 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 particles would improve the tracking performance with the individual cues, the benefit of the combined likelihood model is that it constrains the likelihood and allows the number of particles to be reduced.

6. Conclusions

This paper has presented a framework for learning statistical models for human appearance in images and image sequences. We have shown that a likelihood model, robust to both image clutter and small errors in limb position, can be constructed from probabilistic models of filter responses at individual pixels learned from training data. For a moderate number of images of people, the positions of the humans were manually marked and steered filter responses for edges, ridges and motion were extracted from positions on and off the humans in the images. Given a certain image position in an unknown image, the learned distributions of filter responses can be used to determine the probability that this location is best explained by the foreground (human) or some general background. Experiments showed that local contrast normalization improves the ability to discriminate between background and foreground filter responses.

Section 4 described how these learned empirical distributions can be exploited for tracking. The learned models are used to define the likelihood of observing edge, ridge and motion filter responses given the predicted pose of a limb. Experiments with a cluttered image sequence illustrate how the the learned likelihood is used for tracking human limbs in the Bayesian framework described in Sidenbladh et al. (2000a) and Sidenbladh and Black (2001).

There remain a number of important directions for future work. First, to diminish the effects of overlearning and incomplete data, analytic functions could be fitted to the learned distributions. In contrast to previous work, our local contrast normalization scheme means that the distributions do not have a simple form (e.g. Cauchy). It may be necessary to employ a mixture of distributions to approximate the likelihoods accurately.

Furthermore, the learning framework presented in this paper is not restricted to responses for edges, ridges and motion. Different statistical measures of texture, or distributions over color, can be extracted and learned in similar ways. The Bayesian formulation of the framework enables several cues to be combined in a mathematically grounded way.

Further work needs to be performed to model correlations across scale and among cues. Additionally, filter responses along a limb are assumed constant while, in practice they vary. For example the ridge response is greater in the center of the limb than it is at either end. Additional filters might be employed to cope with termination of the limb at joints or extremities. Similarly, responses are not view-independent as they are assumed here. From a given viewpoint, some poses of the body are much more likely to result in limbs being viewed against the similarly clothed torso resulting in lower filter responses than when they are viewed against a background. We have not attempted to model this view dependence.

While the Bayesian formulation provides a way of combining different cues, the issue of their relative weighting requires further investigation. The issue is related to the spatial dependence of filter responses and

here the weighting is implicitly determined by the number of samples chosen for each cue.

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.

Finally, a more extensive training set, particularly for the motion cue, should be developed. To encourage comparisons of different likelihood models, the current training data, ground truth, and learned models used in this paper can be downloaded from: http://www.nada.kth.se/~hedvig/data.html.

Acknowledgments

HS was sponsored by the Foundation for Strategic Research under the "Center for Autonomous Systems" contract. MJB was supported by the DARPA HumanID Project (ONR contract N000140110886) and by a gift from the Xerox Foundation. This support is gratefully acknowledged.

We thank David Fleet who developed an early edge likelihood model 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.

Notes

1. See recent work by Nestares and Fleet (2001) for a related approach that uses the phase of complex-valued filter responses to achieve similar contrast insensitivity.
2. s corresponds to the scale parameter $\sqrt{t}$ in Lindeberg (1998).
3. Since image scale is never negative, the scale is in reality computed as $s = \max(0, -24 + 4.45\,w)$.
4. The spatial and temporal statistics of neighboring pixels are unlikely to be independent (Sullivan et al., 2000). 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, regardless of pose, and covers the visible parts of the human model.
5. 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. In general, issues of spatial correlation deserve further study (c.f. Sullivan et al. (1999, 2000)).

References

Black, M.J. and Anandan, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104.

Black, M.J. and Jepson, A.D. 1998. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84.

Bregler, C. and Malik, J. 1998. Tracking people with twists and exponential maps. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 8–15.

Cham, T.-J. and Rehg, J.M. 1999. A multiple hypothesis approach to figure tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 239–245.

Comaniciu, D., Ramesh, V., and Meer, P. 2000. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. 142–149.

Darrell, T., Gordon, G., Harville, M., and Woodfill, J. 2000. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185.

DeCarlo, D. and Metaxas, D. 1996. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 231–238.

Deutscher, J., Blake, A., and Reid, I. 2000. Articulated motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. 126–133.

Fischler, M.A. and Bolles, R.C. 1981. RANSAC random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 26:381–395.

Freeman, W.T. and Adelson, E.H. 1991. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906.

Gavrila, D.M. 1996. Vision-based 3-D tracking of humans in action. Ph.D. thesis, University of Maryland, College Park, MD.

Gavrila, D.M. 1999. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Geman, D. and Jedynak, B. 1996. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14.

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

Hogg, D.C. 1983. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20.

Haritaoglu, I., Harwood, D., and Davis, L.S. 2000. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830.

Isard, M. and Blake, A. 1998. Condensation—Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28.

Jepson, A.D., Fleet, D.J., and El-Maraghi, T.F. 2001. Robust on-line appearance models for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. I, pp. 415–422.

Ju, S.X., Black, M.J., and Yacoob, Y. 1996. Cardboard people: A parameterized model of articulated motion. In *International Conference on Automatic Face and Gesture Recognition*, pp. 38–44.

Kaliath, T. 1951. The divergence and Bhattarcharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, COM-15(1):52–60.

Konishi, S.M., Yuille, A.L., Coughlan, J.M., and Zhu, S.C. 1999. Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 573–579.

Kullback, S. and Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Lee, A.B., Mumford, D., and Huang, J. 2001. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1/2):35–59.

Lindeberg, T. 1998. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156.

Moeslund, T.B. and Granum, E. 2001. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 18:231–268.

Nestares, O. and Fleet, D.J. 2001. Probabilistic tracking of motion boundaries with spatiotemporal predictions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. II, pp. 358–365.

Olshausen, B.A. and Field, D.J. 1996. Natural image statistics and efficient coding. *Computation in Neural Systems*, 7(2):333–339.

Ormoneit, D., Sidenbladh, H., Black, M.J., and Hastie, T. 2001. Learning and tracking cyclic human motion. In *Advances in Neural Information Processing Systems 13*, T.K. Leen, T.G. Dietterich, and V. Tresp (Eds.), pp. 894–900.

Rasmussen, C. and Hager, G. 2001. Probabilistic data association methods for tracking complex visual objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):560–576.

Rehg, J. and Kanade, T. 1995. Model-based tracking of self-occluding articulated objects. In *IEEE International Conference on Computer Vision, ICCV*, pp. 612–617.

Rittscher, J., Kato, J., Joga, S., and Blake, A. 2000. A probabilistic background model for tracking. In *European Conference on Computer Vision, ECCV*, D. Vernon (Ed.), pp. 336–350.

Rohr, K. 1994. Towards model-based recognition of human movements in image sequences. *CVGIP—Image Understanding*, 59(1):94–115.

Rohr, K. 1997. Human movement analysis based on explicit motion models. In *Motion-Based Recognition*, M. Shah and R. Jain (Eds.), pp. 171–198.

Ruderman, D.L. 1994. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548.

Ruderman, D.L. 1997. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3395.

Sidenbladh, H. 2001. Probabilistic tracking and reconstruction of 3D human motion in monocular video sequences. Ph.D. Thesis, KTH, Sweden. TRITA-NA-0114.

Sidenbladh, H. and Black, M.J. 2001. Learning image statistics for Bayesian tracking. In *IEEE International Conference on Computer Vision, ICCV*, vol. 2, pp. 709–716.

Sidenbladh, H., Black, M.J., and Fleet, D.J. 2000a. Stochastic tracking of 3D human figures using 2D image motion. In *European Conference on Computer Vision, ECCV*, D. Vernon (Ed.), vol. 2, pp. 702–718.

Sidenbladh, H., Black, M.J., and Sigal, L. 2002. Implicit probabilistic models of human motion for synthesis and tracking. In *European Conference on Computer Vision, ECCV*, Copenhagen.

Sidenbladh, H., De la Torre, F., and Black, M.J. 2000b. A framework for modeling the appearance of 3D articulated figures. In *International Conference on Automatic Face and Gesture Recognition*, pp. 368–375.

Simoncelli, E.P. 1997. Statistical models for images: Compression, restoration and optical flow. In *Asilomar Conference on Signals, Systems and Computers*.

Simoncelli, E.P., Adelson, E.H., and Heeger, D.J. 1991. Probability distributions of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 310–315.

Sminchisescu, C. and Triggs, B. 2001. Covariance scaled sampling for monocular 3D body tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 447–454.

Sullivan, J., Blake, A., Isard, M., and MacCormick, J. 1999. Object localization by Bayesian correlation. In *IEEE International Conference on Computer Vision, ICCV*, vol. 2, pp. 1068–1075.

Sullivan, J., Blake, A., and Rittscher, J. 2000. Statistical foreground modelling for object localisation. In *European Conference on Computer Vision, ECCV*, D. Vernon (Ed.), vol. 2, pp. 307–323.

Wachter, S. and Nagel, H. 1999. Tracking of persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192.

Wren, C., Azarbayejani, A., Darrel, T., and Pentland, A. 1997. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785.

Yacoob, Y. and Black, M.J. 1999. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 73(2):232–247.

Zhu, S.C. and Mumford, D. 1997. Learning generic prior models for visual computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250.