Robust Online Appearance Models for Visual Tracking

Allan D. Jepson, *Member*, *IEEE Computer Society*, David J. Fleet, *Member*, *IEEE Computer Society*, and Thomas F. El-Maraghi, *Member*, *IEEE Computer Society*

Abstract—We propose a framework for learning robust, adaptive, appearance models to be used for motion-based tracking of natural objects. The model adapts to slowly changing appearance, and it maintains a natural measure of the *stability* of the observed image structure during tracking. By identifying stable properties of appearance, we can weight them more heavily for motion estimation, while less stable properties can be proportionately downweighted. The appearance model involves a mixture of stable image structure, learned over long time courses, along with two-frame motion information and an outlier process. An online EM-algorithm is used to adapt the appearance model parameters over time. An implementation of this approach is developed for an appearance model based on the filter responses from a steerable pyramid. This model is used in a motion-based tracking algorithm to provide robustness in the face of image outliers, such as those caused by occlusions, while adapting to natural changes in appearance such as those due to facial expressions or variations in 3D pose.

Index Terms—Motion, optical flow, tracking, occlusion, EM algorithm, adaptive appearance models.

1 Introduction

One of the main factors that limits the performance of visual tracking algorithms is the lack of suitable appearance models. This is true of template-matching methods that do not adapt to appearance changes, and it is true of motion-based tracking where the appearance model can change rapidly, allowing models to drift away from targets.

This paper proposes a robust, adaptive appearance model for motion-based tracking of complex natural objects. The model adapts to slowly changing appearance, and it maintains a natural measure of the *stability* of the observed image structure during tracking. By identifying stable properties of appearance, we can weight them more heavily for motion estimation, while less stable properties can be proportionately downweighted.

The generative model for appearance is formulated as a mixture of three components, namely, a stable component that is learned with a relatively long time-course, a two-frame transient component, and an outlier process. The stable component adapts to slowly varying properties of image appearance, thereby encoding properties that remain reasonably stable over long time frames. This allows the stable model to identify the most reliable structure for motion estimation, while the two-frame constraints provide additional information when the appearance model is being initialized or when appearance is changing too quickly compared to the slow adaptation of the stable component. The parameters of the mixture model are learned efficiently with an online version of the EM algorithm.

- *A.D. Jepson and T.F. El-Maraghi are with the Department of Computer Science, University of Toronto, Toronto, M5S 3H5, Canada. E-mail: {jepson, tem}@cs.toronto.edu.*
- *D.J. Fleet is with the Palo Alto Research Center, 3333 Coyoto Hill Rd., Palo Alto, CA 94304. E-mail: fleet@parc.com.*

Manuscript received 29 Nov. 2001; revised 2 Oct. 2002; accepted 5 Mar. 2003. Recommended for acceptance by Z. Zhang.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 115465.

The appearance model and the tracker formulated here can be used with a wide variety of different types of image properties. These include image gradient, image features, and multiscale pyramid coefficients. Here, we consider an appearance model based on the complex-valued coefficients of a steerable pyramid. This wavelet-based model allows for stability at different scales or in different spatial neighborhoods to be assessed independently. This is useful for tracking objects where some regions of the object may be stable while others are not, like faces where the mouth may be less stable if someone is talking or changing expression during tracking. Moreover, the use of these complex-valued wavelet responses affords significant insensitivity to changes in lighting conditions.

Together, these components yield a robust motion estimator that naturally combines both stable appearance constraints and two-frame motion constraints. The approach is robust with respect to partial occlusions, significant image deformations, and natural appearance changes, like those occurring with facial expressions and clothing. The appearance model framework supports tracking and accurate image alignment for a variety of possible applications, such as localized feature tracking, and tracking models for which relative alignment and position is important, such as limbs of a human body.

2 Previous Work

Although, not always described as such, every motion estimation and tracking method embodies some representation of image appearance. The common, image-based appearance models include templates [12], [24], [25], [27], view-based subspace models [2], [6], [13], the most recent frame in two-frame flow estimation [28], [29], temporally filtered, motion-compensated images [14], [33], [35], and global statistics [1], [4]. There are several other approaches to visual tracking, such as 3D model-based methods (e.g., [3], [20], [29]) and curve-based methods (e.g., [15], [22], [26])

Fig. 1. These are cropped images from a 1,150 frame sequence that was taken with handheld video camera. The ellipse shows the tracking region in which the motion and appearance are estimated.

that are less directly relevant to the approach taken here, and we therefore omit a discussion of them below.

Tracking with fixed templates can be reliable over short durations, but it copes poorly with appearance changes over longer durations that occur in most applications. One can improve the robustness of such trackers by representing the variability of each pixel in the template [12], [18]. This allows one to track objects against cluttered backgrounds without specifying the detailed support of the object. A learning stage is required prior to tracking in which one estimates the variance of image brightness at each pixel over the training ensemble.

Robustness can be further enhanced with the use of subspace models of appearance [2], [6], [13]. Such view-based models, usually learned with Principal Component Analysis, have the advantage of modelling variations in pose and lighting. They can also be used for search as well as incremental tracking. But, they also have the disadvantage that they are object specific and they require training prior to tracking in order to learn the subspace basis.

The use of local and global image statistics, such as color histograms, have been popular for tracking [1], [4]. These methods offer robustness under image distortions and occlusions. Moreover, the models are fast to learn and can be used for search as well as tracking. Their primary disadvantage is their lack of expressiveness which limits their ability to accurately register the model to the image in many cases. Moreover, these coarse appearance models can also fail to accurately track regions that share similar statistics with other nearby regions.

Motion-based trackers integrate motion estimates through time. With two-frame motion estimation, the appearance model is, implicitly, just the most recently observed image. This has the advantage of adapting rapidly to appearance changes, but it suffers because models often drift away from the target. This is especially problematic when the motions of the target and background are similar.

Motion estimation can be improved significantly by accumulating an appearance model through time. Indeed, optimal motion estimation can be formulated as the estimation of both motion and appearance simultaneously [35]. In this sense, like the learned subspace approaches above, optimal motion estimation is acheived by registering the image against an appearance model that is acquired through time. For example, from the estimated motion, a stabilized image sequence can be formed to learn the appearance model. The stabilized sequence can be smoothed with an IIR low-pass filter, for example, to efficiently remove noise and to weight the most recent frames more heavily than past frames in constructing the appearance model [14], [33]. Our approach bears some similarity to this, but with online adaptation of a mixture model that captures the stable components of image appearance. The use of the adaptive mixture model for foreground appearance also bears some similarity to the Gaussian mixture model used by Stauffer and Grimson for background modeling [32].

This paper describes a robust appearance model that adapts to changes in image appearance. The three key contributions include:

1. an appearance model that identifies stable or slowly varying structure, and naturally combines this stable structure with more transient image information;
2. an online version of EM for adapting the model parameters; and
3. a tracking algorithm which incrementally estimates both motion and appearance.

Like all adaptive appearance models, there is a natural trade off that depends on the time-course of adaptation. Faster time courses allow rapid adaptation to appearance change, while slower time courses provide greater persistence of the model, which allow one to cope with occlusions and other outliers. Here, we find a balance between different time courses with a natural mixing of both two-frame motion information and stable appearance that is learned over many frames.

3 $\mathcal{WSL}$ Appearance Model Framework

As a motivational example, consider tracking a region, such as the face in Fig. 1 (see also [36]), using a simple parametric motion model. As the subject's head moves, the local appearance of the stabilized image can be expected to vary smoothly due to changes in 3D viewpoint and to changes in the subject's facial expression. We also expect the occasional burst of outliers caused by occlusion and sudden appearance changes, such as when the subject's glasses are removed.

These phenomena motivate the components of our appearance model, which we introduce in the context of a single real-valued data observation, say d_t, at each frame t. The first component of the appearance model is the stable model, $\mathcal{S}$. The purpose of this component is to capture the behavior of temporally stable and slowly varying image observations, when and where they occur. Conditioned on the generation of observation d_t by the stable component, we model the probability density for d_t by the Gaussian density $p_s(d_t \,|\, \mu_{s,t}, \sigma^2_{s,t})$. Here, $\mu_{s,t}$ and $\sigma^2_{s,t}$ are piecewise,

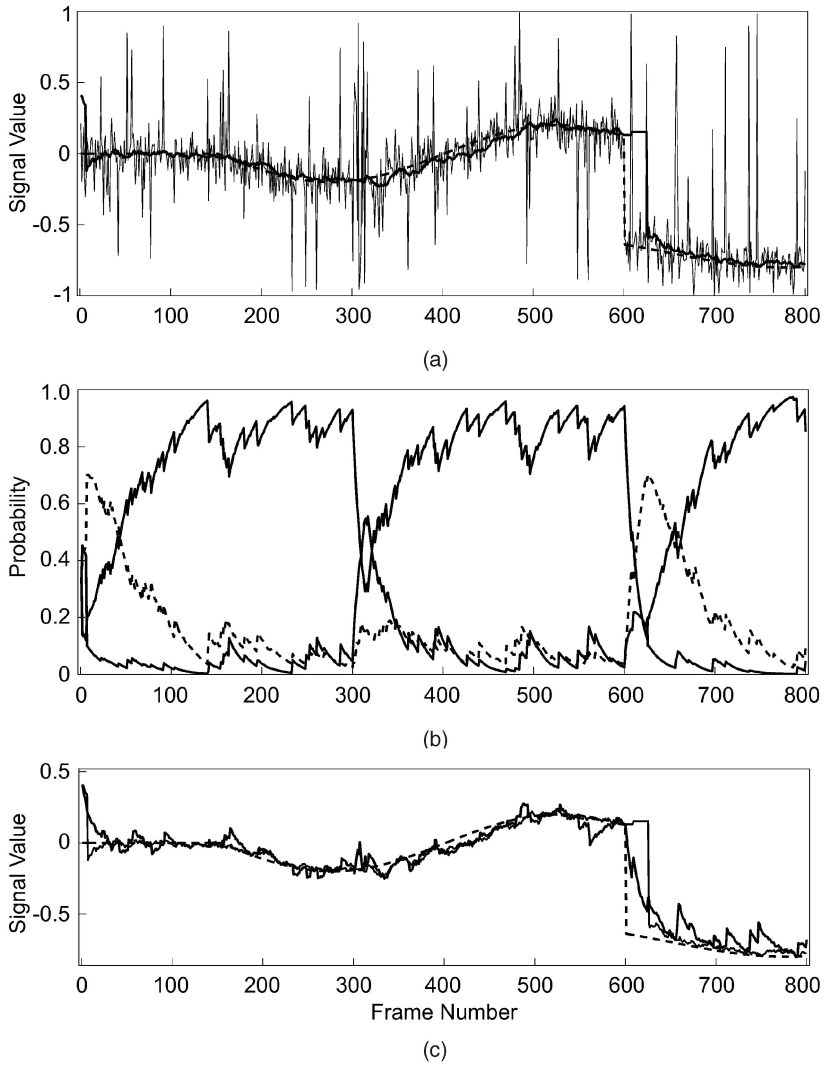

Fig. 2. Estimation using online EM. (a) The original data (thin noisy curve) with true signal (dashed) and the estimated mean of the stable process (solid thick line). The noise is a mixture of Gaussian and uniform densities, with mixing probabilities (0.9, 0.1), except for 15 frames at 300 which are pure outliers. (b) Mixing probabilities for $\mathcal{S}$ (black), $\mathcal{W}$ (dashed), and the $\mathcal{L}$ (solid light gray). (c) The output of an adaptive IIR filter (solid thick curve) applied to the data is shown in addition to the noiseless signal and the stable mean estimate. The time constant of the IIR filter was identical to that of the temporal window used for estimation of the stable component of the mixture model.

slowly varying functions of time, specifying the mean and variance of the Gaussian model.

The second component of the model accounts for data outliers, which are expected to arise due to failures in tracking, occlusion, or noise. We refer to the corresponding random process as the "lost" component, and denote it by $\mathcal{L}$. The probability density for $\mathcal{L}$, denoted by $p_l(d_t)$, is taken to be a uniform distribution over the observation domain.

The synthetic signal depicted in Fig. 2a provides an idealized example of these generative processes. The smooth (dashed) curve represents the piecewise slowly varying appearance signal. The observed data (thin noisy curve) has been corrupted by long-tailed noise, formed from a mixture of the Gaussian density $p_s(d_t|\mu_{s,t}, \sigma^2_{s,t})$ and the broad distribution $p_l(d_t)$ for the lost component. In accordance with our discussion of Fig. 1, we have also included an appearance discontinuity at frame 600, and a burst of outliers representing an occluder between frames 300 and 315.

The third component of our appearance model is motivated by the desire to blend the appearance model with an image-based tracking algorithm. In this context, we wish to be able to track a given image region even before we are able to learn a model for the dominant, stable image structure within the target region. We also want to be able to cope with sudden changes of image appearance like those in Figs. 1 and 2. The problem exists because initially, or after a sudden appearance change, we have neither a good stable appearance model, nor a model for how the object is expected to move. In these cases, rather than rely on the stable component, it makes sense to have a third component of the appearance model which adapts with a short time-course, as in a two-frame tracker. This third component of the appearance model is called the wandering component $\mathcal{W}$. In effect, the wandering component permits the tracker described in Section 6 to gracefully degrade to a two-frame motion tracker when the appearance model does not account for enough past data observations.

The wandering component needs to allow both for more rapid temporal variations and shorter temporal histories than are required for the reliable estimation of the stable model parameters. It should therefore adapt to the local properties of image appearance on a much shorter time course than the stable model. As such, we model the probability density for d_t, given that it is generated by $\mathcal{W}$, to be the Gaussian density $p_w(d_t \,|\, d_{t-1})$. Here, the mean is simply the observation from the previous frame, d_{t-1}, and the variance is fixed at σ_w^2.

The three components, $\mathcal{W}$, $\mathcal{S}$, and $\mathcal{L}$, are combined in a probabilistic mixture model for d_t,

$$p(d_t \,|\, \mathbf{q}_t, \mathbf{m}_t, d_{t-1}) = m_w p_w(d_t|d_{t-1}) + m_s p_s(d_t|\mathbf{q}_t) + m_l\, p_l(d_t), \tag{1}$$

where $\mathbf{m} = (m_w, m_s, m_l)$ are the mixing probabilities, and $\mathbf{q}_t = (\mu_{s,t}, \sigma_{s,t}^2)$ contains the mean and variance parameters of the stable component of the model. Our goal is to use this probabilistic mixture model as a prediction density for new observations d_t, given the recent observation history (under a sliding window of temporal support). As such, the mixing probabilities reflect the fraction of recent observations explained by the different model components.

4 Parameter Estimation with Online EM

Our goal is to estimate the parameters of the generative model in (1), namely, the mean and variance of the data prediction by the stable process, $\mathbf{q} = (\mu_s, \sigma_s^2)$, and the mixing probabilities $\mathbf{m} = (m_w, m_s, m_l)$. Since we plan to apply this mixture model estimation scheme to high-dimensional appearance data, like the responses of a wavelet filter bank, it is very important that we find an efficient computational algorithm that requires a small amount of memory for each temporal stream of data observations.

Anticipating a recursive formulation, and allowing for temporal adaptation of the model parameters, we consider data observations under an exponential envelope located at the current time, $S_t(k) = \alpha e^{-(t-k)/\tau}$, for $k \le t$. Here, $\tau = n_s / \log 2$, where n_s is the half-life of the envelope in frames, and $\alpha = 1 - e^{-1/\tau}$, so the envelope weights $S_t(k)$ sum to 1. With this envelope, we can express the log-likelihood of the observation history, $\mathbf{d}_t = \{d_k\}_{k=0}^t$, according to the mixture model density in (1) as

$$L(\mathbf{d}_t \,|\, \mathbf{m}_t, \mathbf{q}_t) = \sum_{k=t}^{-\infty} S_t(k)\, \log p(d_k \,|\, \mathbf{m}_t, \mathbf{q}_t, d_{k-1}), \tag{2}$$

where $\mathbf{m}_t$ and $\mathbf{q}_t$ denote parameters relevant to the data under the temporal support envelope $S_t(k)$. Although these parameters change slowly through time, we first consider an EM-algorithm [5] for estimating $\mathbf{m}_t$ and $\mathbf{q}_t$, that assumes they are approximately constant under the temporal window. The form of these EM-updates provides the basis for an online nonlinear estimator.

As with a typical EM iteration, given an initial guess for the state variables $\mathbf{m}_t$ and $\mathbf{q}_t$, the E-step provides the ownership probabilities for each observation d_k:

$$o_{i,t}(d_k) = \frac{m_{i,t}\, p_i(d_k;\, \mathbf{q}_t, d_{k-1})}{p(d_k;\, \mathbf{m}_t, \mathbf{q}_t, d_{k-1})}, \tag{3}$$

for $i \in \{w, s, l\}$ (see [5]). Conditioned on these ownerships, the M-step then computes new maximum likelihood estimates for the parameters $\mathbf{m}_t$ and $\mathbf{q}_t$. To this end, first the updated mixture probabilities, $\mathbf{m}_t$, are given by

$$m_{i,t} \;=\; \sum_{k=t}^{-\infty} S_t(k)\, o_{i,t}(d_k), \tag{4}$$

for $i \in \{w, s, l\}$ (we have reused the notation $m_{i,t}$ to denote the updated values). Similarly, the M-step for the mean and variance are

$$\mu_{s,t} = \frac{M_{1,t}}{M_{0,t}}\,, \quad \sigma_{s,t}^2 = \frac{M_{2,t}}{M_{0,t}} - \mu_{s,t}^2, \tag{5}$$

where $M_{j,t}$ are the ownership weighted, jth-order data moments defined by

$$M_{j,t} = \sum_{k=t}^{-\infty} S_t(k)\, d_k^j\, o_{s,t}(d_k). \tag{6}$$

It is worth noting here that the zeroth data moment, the time averaged ownerships of the stable process, is precisely the mixing probability for the stable component of the appearance model, $M_{0,t} = m_{s,t}$. The standard EM-algorithm then consists of iterating the steps outlined in (3), (4), (5), and (6).

This EM-algorithm requires that the data from previous times be retained to compute $o_{s,t}(d_k)$, which is impractical for an online approach. Instead, we adopt an approximation to (3), (4), (5), and (6). To this end, we first exploit a recursive expression for the exponential support $S_t(k)$ to obtain,

$$\begin{aligned} M_{j,t} &= S_t(t)\, d_t^j\, o_{s,t}(d_t) \;+\; \sum_{k=t-1}^{-\infty} S_t(k)\, d_k^j\, o_{s,t}(d_k)\,, \\ &= \alpha\, d_t^j\, o_{s,t}(d_t) + (1-\alpha) \sum_{k=t-1}^{-\infty} S_{t-1}(k)\, d_k^j\, o_{s,t}(d_k)\,, \end{aligned} \tag{7}$$

where, as above, $\alpha = 1 - e^{-1/\tau}$. In order to avoid having to retain past data, we approximate the current ownership of past data by the ownerships at the times the data were first observed. That is, we replace $o_{s,t}(d_k)$ by $o_{s,k}(d_k)$, to obtain the approximate moments

$$\begin{aligned} \hat{M}_{j,t} &= \alpha\, d_t^j\, o_{s,t}(d_t) + (1-\alpha) \sum_{k=t-1}^{-\infty} S_{t-1}(k)\, d_k^j\, o_{s,k}(d_k)\,, \\ &= \alpha\, d_t^j\, o_{s,t}(d_t) + (1-\alpha)\, \hat{M}_{j,t-1}. \end{aligned} \tag{8}$$

We also approximate the mixing probabilities the same way:

$$\hat{m}_{i,t} = \alpha\, o_{i,t}(d_t) \;+\; (1-\alpha)\, \hat{m}_{i,t-1}, \tag{9}$$

for $i \in \{s, w, l\}$. One further deviation from (3), (4), (5), and (6) is used to avoid singular situations; i.e., we impose a nonzero lower bound on the mixing probabilities and $\sigma_{s,t}$. We could also achieve the same effect straightforwardly with priors on $\mathbf{m}$ and $\mathbf{q}$.

In this approximation to the batch EM algorithm in (3), (4), (5), and (6), as mentioned above, we do not update the data ownerships of the past observations. Therefore, when the model parameters change rapidly this online approximation is poor. Fortunately, this typically occurs when the

data are not stable, which usually results in a low mixing probability and a broad variance for $\mathcal{S}$, indicating that the data is not well explained by the stable component. Conversely, when the mean and variance drift slowly, the online approximation is typically very good (see Fig. 2).

As with any nonlinear online estimator, this procedure requires an initial guess for the optimization. It also requires occasional restart guesses for times when the estimator becomes trapped at local extrema. This occurs, for example, when the stable component, $\mathcal{S}$, loses track of the underlying mean state because of sudden changes in appearance, or because of unstable data over relatively long time periods. Fortunately, we can detect such cases by monitoring the stable mixing probability, $m_{s,t}$, which measures the fraction of the recent observations that are successfully explained by the stable component. This mixing probability normally remains high when there is stable structure. But, when $\mathcal{S}$ fails to track the stable structure, or when there is no stable structure, then $m_{s,t}$ usually becomes very small (e.g., see Fig. 2). While there are many ways to detect the need for restarts of the nonlinear estimator, in all the experiments below, we simply initiate restarts (i.e., new initial guesses) whenever $m_{s,t}$ falls below a threshold of 0.1. We find that thresholds between 0.05 and 0.4 yield similar estimator behavior.

To restart the estimator, we simply set the values of all state variables to the initial guess. As with many nonlinear estimators, the exact choice of the initial guess is not critical. In the experiments reported below, we set the mixing probabilities $m_{i,t}$ to 0.4, 0.15, and 0.45 for $i = w, s, l$, respectively. The small value for $m_{s,t}$ reflects an initial uncertainty for the $\mathcal{S}$ model. In particular, initially, there is no history of stable observations that the stable model has explained. To the contrary, conditioned on a recent restart, we should assume initially that the stable component has not captured stable structure, and that the appearance observations may be changing rapidly, or they may be outliers. For this reason, our initial guess uses a larger value for the wandering and lost mixing probabilities, $m_{w,t}$ and $m_{l,t}$. Finally, the initial values for the moments $M_{j,t}$ for $j = 0, 1, 2$ are taken to be $m_{s,t}$, $d_t m_{s,t}$, and $\sigma_{s,0}^2 m_{s,t}$, respectively. In effect, this starts the stable model with a mean given by the current observation d_t, and a variance given by the constant $\sigma_{s,0}^2$. Here, we use $\sigma_{s,0} = \sigma_w/1.5$ so that there is always some prior preference for optima in which the stable model explains the coherent observations. These same values are used for initialization in the first frame. The performance of the estimator was found to be robust to variations in these restart constants, and to various alternative choices for the restart criterion [7].

Fig. 2 illustrates the EM procedure on our 1D example with half-life $n_s = 8$. As shown in Fig. 2b, initially, the $\mathcal{W}$ component owns most of the data because the mixing probability of the stable component $\mathcal{S}$ remains small until it has seen enough stable observations to have a high probability of explaining subsequent observations. Since the first 150 observations are constant, with occasional outliers, the mixing probability of the stable component grows steadily. Beyond frame 150, the signal is slowly varying. Note that the mean of the stable component shown Fig. 2a accurately tracks this slowly varying signal.

During the outlier burst that occurs at frame 300, Fig. 2b shows that the outlier $\mathcal{L}$ component begins to own a much greater share of the data, while the mixing probability for $\mathcal{S}$ decays quickly. Similarly, after the step change in the signal at frame 600, the stable component fails to provide good predictions for the signal, so the mixing probability for $\mathcal{S}$ again decays quickly. In this case, the stable component exhibits the persistence we expect from a stable model, so that it can remain stable when outliers are observed. However, unlike the situation immediately after frame 300, the signal undergoes a step change at frame 600, after which it is slowly varying. Because the signal is coherent after frame 600 and the predictions from the stable process are poor, it is the wandering $\mathcal{W}$ component that best predicts (explains) the data. Therefore, its mixing probability increases rapidly. By frame 625, the $\mathcal{S}$ mixing probability drops sufficiently low that a restart occurs, after which the $\mathcal{S}$ component locks back onto the true state, and its mixing probability rises.

This behavior illustrates an important property of the model. Namely, the stable component adapts over a relatively long time course and, therefore, it exhibits persistence through bursts of outliers. Naturally, this persistence comes at the cost of not being able to react quickly to step changes in the appearance. In these cases, it is useful to have another process that adapts over short time courses, such as our $\mathcal{W}$ process, so it can quickly adapt to coherence in the signal, giving the stable model more time to adapt.

Finally, by way of comparison, conventional adaptive templates typically use a recursive (IIR) linear filter to average the data observations. It is well-known that such filters are very sensitive to outliers and, therefore, we expect them to produce noisier signals as compared to the mean of the stable component of the mixture model. For example, Fig. 2c compares the mean of the stable component to the reponses of an IIR filter applied to the data observations. Computed over the first 600 frames (i.e., before the step change in the signal), the RMS error of the IIR estimate is approximately 65 percent higher than that of the stable mean. Furthermore, when there are sudden changes in appearance, like that at frame 600, adaptive IIR filters will tend to produce a signal that averages the signal before and after the step, representing neither accurately. By comparison, with the mixture model used here, with a long time-course for the $\mathcal{S}$ component and a short time-course for the $\mathcal{W}$ component, the $\mathcal{S}$ component exhibits persistence, continuing to represent the stable component of the signal before the step, while the $\mathcal{W}$ component quickly begins to track the coherent signal structure immediately after the step. In this way, the $\mathcal{WSL}$ mixture model achieves our goal of capturing the structure of slowly varying signals having both outliers and occasional sudden changes in appearance.

5 Wavelet-Based Appearance Model

There are many properties of image appearance that one could learn for tracking and object search. Examples include local color statistics, multiscale filter responses, and localized edge fragments. Here, we wanted to be able to detect stable properties of appearance that might be localized spatially, or restricted to certain scales or orientations. This will be useful for objects in which some local regions are stable while others are not, or with motion blur in which coarse scales might be more stable than finer scales, for example. Accordingly, we used the $\mathcal{WSL}$ appearance modeling framework to adaptively model the responses of a steerable pyramid [31], [30]. In particular, we focus on the time-varying behavior of the phase structure of the filter responses.

We chose phase as the basis for the appearance model because of its useful properties for estimating optical flow and stereo disparity [8], [9], [10]. Phase provides a significant degree of amplitude and illumination independence. It allows for robust image matching when there are significant scale changes between views, and it has been shown that phase gradients can be used for estimating significantly larger disparities and image velocities than gradients of images smoothed with similar spatial support [8], [9]. A final benefit of phase is that, based on a scale-space singularity analysis of band-pass signals, one can detect probable outliers in phase responses [9], much like intensity gradient magnitudes can be used to detect unstable gradient-based motion constraints. Clearly, it is preferable to identify outliers explicitly where possible, rather than to rely solely on the robustness of the estimator. The primary disadvantage of using only the phase of the complex-valued wavelet responses are: 1) we are discarding the amplitude of the responses and therefore not using all the available information, and 2) the extra computational overhead involved in computing the wavelet pyramid transform, compared to a gradient-based approach, for example. While phase offers this trade off, it is important to note that the online $\mathcal{WSL}$ estimator could be applied to other image properties such as image intensity or image gradients.[1]

In summary, given an image pyramid and a target region $\mathcal{N}_t$, let $\{d(\mathbf{x},t)\}_{\mathbf{x}\in\mathcal{N}_t}$ denote the set of phase observations, one from each filter within the region, at time t. We then apply a 1D $\mathcal{WSL}$ appearance model to the phase signal from each filter output. Accordingly, let $\mathcal{A}_t=\{(\mathbf{m}(\mathbf{x},t),\,\mathbf{q}(\mathbf{x},t)\}_{\mathbf{x}\in\mathcal{N}_t}$ denote the collective appearance model of the phase at each orientation, scale, and spatial location in $\mathcal{N}_t$. For the experiments below, we used a steerable pyramid based on the G_2 and H_2 filters of [11]. In particular, we used these filters to decompose each image into two scales (tuned to wavelengths of 8 and 16 pixels, subsampled by factors of 2 and 4), each of which is further decomposed into 4 orientations. With the use of phase, the $\mathcal{WSL}$ model then involves four parameters:

- The half-life of the exponential temporal support, $S_t(k)$, set to $n_s = 20$ frames for all experiments reported below. More generally, we have used a wide range of time-constants, from 4 to 40, with very good success.
- The phase outlier density in the mixture model is taken to uniform on $[-\pi,\,\pi)$, and, hence, the probability is simply $1/2\pi$.
- The standard deviation of the $\mathcal{W}$ process on phase differences, based on our experience with phase-based optical flow and disparity estimation [8], [9], is taken to be mean-zero Gaussian with $\sigma_w = 0.35\pi$.
- The minimum standard deviation of the stable process is set to $\sigma_{s,0} = 0.1\pi$. This reflects our prior experience with phase variability under near ideal conditions with natural images [8].

1. Since our initial publication of this research at CVPR 2001, the method has been successfully applied to image brightness [21]. Nevertheless, we do not expect brightness to yield as robust an appearance model as phase information because of the many sources of significant brightness variations in natural images.

6 Motion-Based Tracking

We demonstrate the behavior of the adaptive, phase-based appearance model in the context of tracking nonrigid objects. Beginning with an elliptical region, $\mathcal{N}_0$, at time 0, the tracking algorithm estimates the image motion and the appearance model as it tracks the stable image structure within the convected target region $\mathcal{N}_t$ over time. The motion is represented in terms of frame-to-frame parameterized image warps, $\mathbf{w}(\mathbf{x};\mathbf{c})$. In particular, given the warp parameters $\mathbf{c}_t$, a pixel $\mathbf{x}$ at frame $t-1$ corresponds to the image location $\mathbf{x}_t = \mathbf{w}(\mathbf{x};\mathbf{c}_t)$ at time t. We use similarity transforms here, so $\mathbf{c}_t = (\mathbf{u}_t, \theta_t, \rho_t)$ is a 4-vector describing translation, rotation, and scale changes, respectively. We specify translations in pixels, rotations in radians, and the scale parameter denotes a multiplicative factor, so $\vec{\xi} \equiv (0,0,0,1)$ is the identity warp. By way of tracking, the target neighborhood is convected forward at each frame by the motion parameters. That is, given the parameter vector $\mathbf{c}_t$, $\mathcal{N}_t$ is just the elliptical region provided by warping $\mathcal{N}_{t-1}$ by $\mathbf{w}(\mathbf{x};\mathbf{c}_t)$. Other parameterized image warps and other parameterized region representations could also be used (e.g., see [17], [23], [33], [24]).

Our goal is to determine the optimal image warp from the stable properties of image appearance. These are the properties that we believe will enable accurate alignment of coherent structure over long durations. Toward that end, we need to identify stable structure, and we need to use that structure to estimate image warp parameters $\mathbf{c}_t$.

For notational convenience in what follows, let the appearance (phase) data from the previous frame be denoted by $D_{t-1} \equiv \{d_{\mathbf{x},t-1}\}_{\mathbf{x}\in\mathcal{N}_{t-1}}$, where an individual datum is $d_{\mathbf{x},t-1} \equiv d(\mathbf{x},t-1)$. Similarly, given the warp parameters c_t, let the current data D_t warped back to the previous frame of reference be denoted by $\hat{d}_{\mathbf{x},t} \equiv d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)$. Then, conditioned on the correct warp, $\mathbf{c}_t$, and on the appearance model from the previous time, $\mathcal{A}_{t-1}$, the log observation density for D_t is simply

$$L(D_t \mid \mathcal{A}_{t-1}, D_{t-1}, \mathbf{c}_t) = \sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \log\Big[m_s p_s(\hat{d}_{\mathbf{x},t} \mid \mathbf{q}) + m_w p_w(\hat{d}_{\mathbf{x},t} \mid d_{\mathbf{x},t-1}) + m_l p_l\Big]. \quad (10)$$

(Here, we have abused the notation in omitting the explicit dependence of m_s, m_w, m_l, and q on both $\mathbf{x}$ and $t-1$.) Then, as in a standard E-step of an EM iteration [16], the expected ownership probabilities for a single datum $\hat{d}_{\mathbf{x},t}$ are given by

$$o_i(\hat{d}_{\mathbf{x},t}) = \frac{m_i p_i(\hat{d}_{\mathbf{x},t} \mid \mathbf{q}_{t-1})}{p(\hat{d}_{\mathbf{x},t} \mid \mathcal{A}_{\mathbf{x},t-1}, d_{\mathbf{x},t-1})}, \quad (11)$$

for $i = w, s, l$. These quantities provide the probabilities that the current phase response $\hat{d}_{\mathbf{x},t}$ arises from the corresponding components of the $\mathcal{WSL}$ model, and can be used to assess the stability of the corresponding motion constraints.

The stable motion constraints are those arising from $\mathcal{S}$ components in (10) with high ownership probabilities o_s. In particular, a high ownership probability means that the filter channel has a good history of stable observations, and that the current datum is consistent with the current stable model. Accordingly, to determine the optimal warp parameters, we introduce an energy function for $\mathbf{c}_t$ that emphasizes constraints from $\mathcal{S}$ components of the $\mathcal{WSL}$ model with large ownership probabilities.

Fig. 3. Each row shows, from left to right, the tracking region, the stable component's mixing probability $m_s(\mathbf{x},t)$, mean $\mu_s(\mathbf{x},t)$, and ownership probability $o_s(\mathbf{x},t)$. In each case, of the mixing probability, the mean phase and the ownership probabilities, the images show the orientation channel with the highest ownership by the stable process. The different rows of images correspond to frames 244, 259, 274, and 289, top to bottom. Note both the model adaptation and persistence, along with the drop in data ownership within the occluded region.

Considering the basic dynamic phenomena illustrated in Fig. 2, we also expect that there are sometimes not enough stable constraints to reliably determine the warp parameters. This occurs at frame 0 and during rapid appearance changes for example. To deal with these cases, we want to ensure that the stable appearance-based tracker degrades gracefully to a two-frame tracker and, therefore, we wish to include motion constraints from the $\mathcal{W}$ components of the appearance model. Finally, when the target is completely occluded the motion will not be sufficiently constrained by the $\mathcal{S}$ or the $\mathcal{W}$ components (i.e., most of the data are owned by the $\mathcal{L}$ process, see Fig. 2). In this most desperate situation, we need to resort to a priori expectations of the motion in order to regularize the motion estimates.

In combination, these considerations suggest an objective function that is the sum of three energy terms, one for each of the $\mathcal{S}$ and $\mathcal{W}$ components, and one for the prior motion expectations. We consider each of these three terms below.

For the contribution of $\mathcal{S}$ components to the objective function, we use the negative log likelihood of warp parameters $\mathbf{c}_t + \delta_{\mathbf{c}_t}$, for small updates $\delta_{\mathbf{c}_t}$ to the current estimate $\mathbf{c}_t$,

$$E_s(\delta_{\mathbf{c}_t} \mid \mathbf{c}_t) \equiv - \sum_{\mathbf{x} \in \mathcal{N}_{t-1}} o_s(\hat{\boldsymbol{d}}_{\mathbf{x},t}) \log p_s(d(\mathbf{w}(\mathbf{x}; \mathbf{c}_t + \delta_{\mathbf{c}_t}), t) \mid \mathbf{q}). \tag{12}$$

Here, the log-likelihood terms $\log p_s$ are weighted by the data ownerships $o_s(\hat{\boldsymbol{d}}_{\mathbf{x},t})$ derived for the current warp

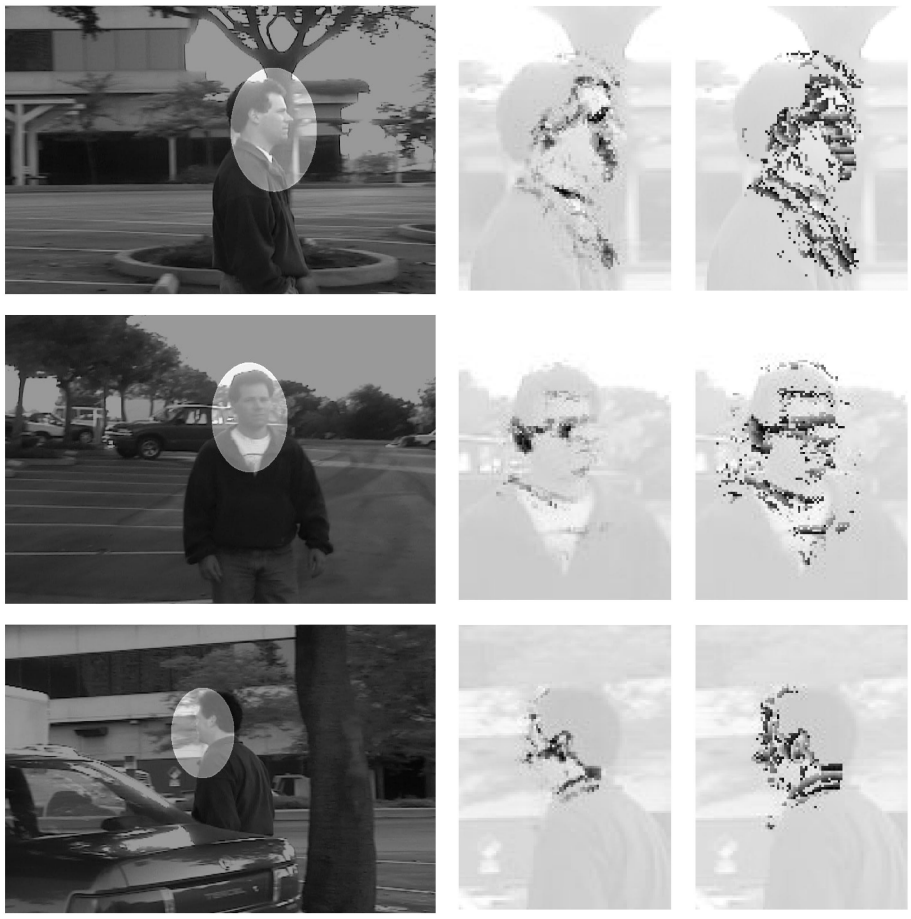

Fig. 4. Adaptation of the model during tracking is illustrated at frames 200, 300, and 480. (a) The three columns show the target region, (b) the mixing probability of the stable component, and (c) the mean of the stable component for the selected frames.

estimate $\mathbf{c}_t$. A natural condition on the optimal warp parameters $\mathbf{c}_t$ is for the expected log-likelihood of the $\mathcal{S}$ constraints to be maximal for the current $\mathbf{c}_t$. That is, we want to converge to a $\mathbf{c}_t$ such that there is no nonzero update, $\delta_{\mathbf{c}_t}$, that will decrease the energy $E_s(\delta_{\mathbf{c}_t} \,|\, \mathbf{c}_t)$.

Similarly, the contribution from the $\mathcal{W}$ components is taken to be

$$E_w(\delta_{\mathbf{c}_t} \,|\, \mathbf{c}_t) \equiv \\ -\epsilon \sum_{\mathbf{x} \in \mathcal{N}_{t-1}} o_w(\hat{\boldsymbol{d}}_{\mathbf{x},t}) \log p_w(d(\mathbf{w}(\mathbf{x}; \mathbf{c}_t + \delta_{\mathbf{c}_t}), t) \,|\, \mathbf{d}_{\mathbf{x},t-1}). \quad (13)$$

The scaling factor $\epsilon \in [0, 1]$ controls the relative influence of the $\mathcal{W}$ motion constraints. When ϵ is zero, they have no influence, leaving the motion estimation to the stable constraints and the prior. When $\epsilon = 1$, $\mathcal{W}$ constraints are taken to be as informative as $\mathcal{S}$ constraints. In practice, there may be many more $\mathcal{W}$ constraints within the tracking region $\mathcal{N}_t$ than there are $\mathcal{S}$ constraints, which tends to bias the motion estimate. For example, in the complex 3D motions in Figs. 3, 4, 5, 6, 7, 8, and 9, regions that are well fit by a similarity transform and stable through time may be relatively small. In these cases, one should expect that the $\mathcal{W}$ constraints in other neighborhoods of $\mathcal{N}_t$ will then act as a source of structured noise that will usually bias the motion away from the optimal warp dictated by the $\mathcal{S}$ constraints.[2] Accordingly, one should use a value of ϵ less than 1 to reduce the effect of this bias. In all experiments in Section 7, we use a value of $\epsilon = 1/20$ which roughly matches the ratio of $\mathcal{S}$ constraints to $\mathcal{W}$ constraints, and we leave the automatic selection of ϵ to future research.

Finally, the third term in the objective function represents the negative log prior probability for $\mathbf{c}_t + \delta_{\mathbf{c}_t}$,

$$E_0(\delta_{\mathbf{c}_t} \,|\, \mathbf{c}_t, \mathbf{c}_{t-1}) \equiv -\log p(\mathbf{c}_t + \delta_{\mathbf{c}_t} \,|\, \mathbf{c}_{t-1}). \quad (14)$$

For this prior model, we take the stochastic dynamics of the motion parameters, $\mathbf{c}_t = (\mathbf{u}_t, \theta_t, \rho_t)$, to be a simple random walk with a small bias toward slow motions [34]. Conditioned on the previous state, $\mathbf{c}_{t-1}$, the prior density over states at frame t is given by a product of Gaussian densities:

$$p(\mathbf{c}_t \,|\, \mathbf{c}_{t-1}) = G(\mathbf{c}_t; \vec{\boldsymbol{\xi}}, \mathbf{V}_1) G(\mathbf{c}_t; \mathbf{c}_{t-1}, \mathbf{V}_2). \quad (15)$$

2. For example, to understand the source of the bias, consider tracking a 3D object such as a human head, with stable features being identified within a relatively small patch on the head. As the head rotates around an axis perpendicular to the viewing direction, the motion of this patch can be approximated by a similarity transform. However, the image motion of the rest of the the head will be biased toward one side of the axis of rotation.

(a) (b) (c)

Fig. 5. Adaptation to changes of expression is illustrated in frames 420, 455, and 490. (a) The three columns show the target region, (b) the mixing probability for the stable component, and (c) the mean of the stable component for the selected frames. Note how the regions around the mouth and eyebrows adapt, while others remain stable.

The first Gaussian factor prefers slow motions, with its mean equal to the identity warp $\vec{\xi}$ and its covariance denoted by $\mathbf{V}_1$. The second Gaussian has its mean centered at $\mathbf{c}_{t-1}$ to prefer slow changes in motion (i.e., smooth motion), with its covariance denoted by $\mathbf{V}_2$. This prior density (15) corresponds to a stochastic dynamical model for $\mathbf{c}_t$ in the form of a Brownian random walk in an energy well:

$$\mathbf{c}_t = \mathbf{c}_{t-1} + \vec{\eta}(\mathbf{c}_{t-1}), \tag{16}$$

where the noise $\vec{\eta}$ is Gaussian, independent from one time to the next, but biased away from $\mathbf{c}_{t-1}$ toward the identity warp $\vec{\xi}$. Since we want the bias in the stochastic dynamics to be relatively weak, we want a broad energy well, and so we set the variances in $\mathbf{V}_1$ to be significantly larger than those in $\mathbf{V}_2$. While the specific values do not greatly affect the behavior of the tracker, in the experiments reported below, we set the covariances to be $\mathbf{V}_1 \equiv \mathrm{diag}(8^2, 8^2, 0.05^2, 0.01^2)$ and $\mathbf{V}_2 \equiv \mathrm{diag}(1, 1, 0.02^2, 0.01^2)$.

Given an initial estimate for $\mathbf{c}_t$, we compute an update $\delta_{\mathbf{c}_t}$ which approximately minimizes the sum of the three energy terms,

$$E(\delta_{\mathbf{c}_t}) = E_s(\delta_{\mathbf{c}_t}|\mathbf{c}_t) + E_w(\delta_{\mathbf{c}_t}|\mathbf{c}_t) + E_0(\delta_{\mathbf{c}_t}|\mathbf{c}_t, \mathbf{c}_{t-1}). \tag{17}$$

This step is described in detail in the Appendix. It is a straightforward variant of the maximization step used in the EM-algorithm for optical flow described in [16]. Given an update $\delta_{\mathbf{c}_t}$, we recompute the ownership probabilities using the updated warp parameters, $\mathbf{c}_t + \delta_{\mathbf{c}_t}$, in place of the previous values $\mathbf{c}_t$. Given these new ownership probabilities, we return to compute a new update, by minimizing the new energy function (17). This process is iterated until the warp update $\delta_{\mathbf{c}_t}$ is sufficiently small. Finally, in order to increase the domain of convergence of this iterative method, we have also found it useful to use a coarse-to-fine framework with deterministic annealing (see the Appendix for details).

Once the warp parameters $\mathbf{c}_t$ have been determined, we convect the appearance model $\mathcal{A}_{t-1}$ forward to the current time t using the warp specified by $\mathbf{c}_t$. To perform this warp, we use a piecewise constant interpolant for the $\mathcal{WSL}$ state variables $\mathbf{m}(\mathbf{x}, t-1)$ and $\sigma_s(\mathbf{x}, t-1)$. This interpolation was expected to be too crude to use for the interpolation of the mean $\mu(\mathbf{x}, t-1)$ for the stable process, so instead, the mean is interpolated using a piecewise linear model. The spatial phase gradient for this interpolation is determined from the gradient of the filter responses at the nearest pixel to the desired location $\mathbf{x}$ on the image pyramid sampling grid [10].

Fig. 6. Robust tracking despite occlusion. Tracking results for frames 200, 205, 210, and 215 are shown, top to bottom. The elliptical tracking region, and the stable model's mixing probability, mean, and ownership are arranged left to right. Note that the model is misaligned during the occlusion (see the second and third images on the second row), but that it promptly realigns. Also, note the stability and model persistence (left three columns), along with the reduced data ownership on the hand (right column).

7 Experiments

The phase-based $\mathcal{WSL}$ appearance model and tracker have been implemented and tested on a wide variety of natural image sequences. Without optimizing the code in any significant way, the current implementation takes approximately 10 sec/frame on a 400MHz SUN workstation when applied to images of size 720×480. Approximately half that time is used to compute the wavelet transform. While the appearance model is currently computed over the entire image, only the data obtained within the convected elliptical region is used to compute the motion. The user initializes the tracker manually by placing an elliptical region in the image at frame 0.

The behavior of the tracking algorithm is illustrated in Fig. 3, where we plot the elliptical target region $\mathcal{N}_t$, the mixing probability $m_s(\mathbf{x}, t)$, the mean $\mu_s(\mathbf{x}, t)$, and the data ownership $o_{s,t}(\mathbf{x}, t)$ for the stable component, each overlaid on the original images. In these and the following images, we only show responses where $m_s(\mathbf{x}, t)$ is greater than a fixed threshold. Thus, blank areas indicate that the appearance model has not found stable structure. As is expected, the significant responses (shown in black) for the $\mathcal{S}$ component occur around higher contrast image regions.

For Fig. 3, the processing was started roughly 70 frames prior to the frame shown in the top row [36]. The significant responses for $m_s(\mathbf{x}, t)$ and $o_s(\mathbf{x}, t)$ demonstrate that the appearance model successfully identified stable, slowly varying structure, typically inside the object boundary. On the second and third rows of Fig. 3, where the person is partially occluded by the sign, note that $m_s(\mathbf{x}, t)$ decays smoothly in the occluded region due to the absence of data support, while the mean $\mu_s(\mathbf{x}, t)$ remains roughly fixed until $m_s(\mathbf{x}, t)$ falls below the

Fig. 7. Tracking with partial occlusion along with variable lighting, appearance, and size. The camera was stationary and the sequences are each roughly 250 frames. The highlighted target region for selected frames is superimposed on the last frame.

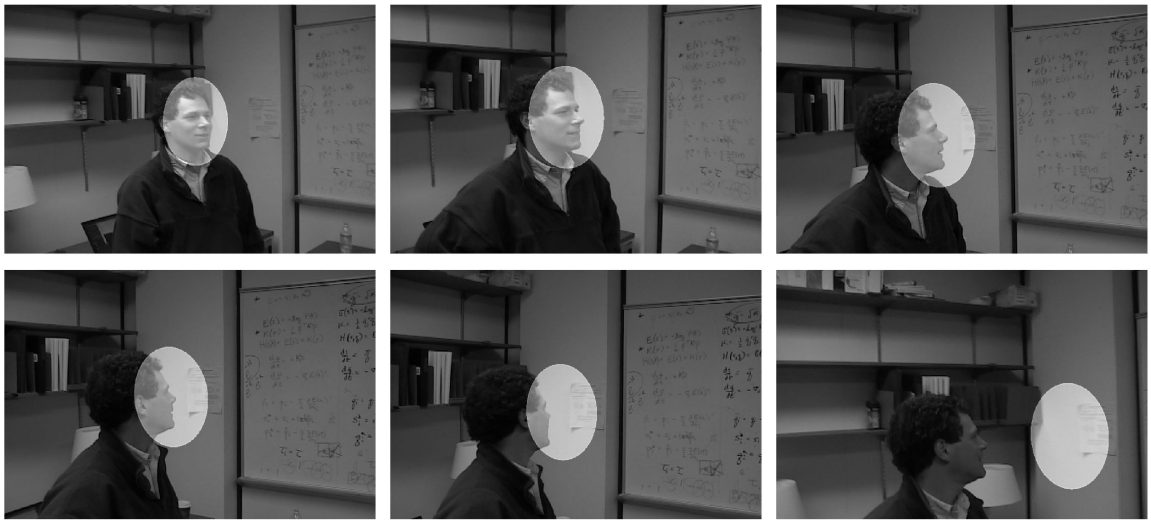

Fig. 8. Tracking failure (frames 440, 480, 520, 560, 600, and 640): When learning stable structure it is possible for the model to learn the structure of the background whenever the background moves consistently with the foreground over a relatively long time. In this case, the model can drift off of the target object, as occurs in this case when the observer slowly rotates their head to leave the target neighborhood stable on the background surface.

plotting threshold. This clearly demonstrates the persistence of the appearance model. The third row depicts the model after roughly 20 frames of occlusion (recall the half-life of the model is $n_s = 20$), by which time the weaker components in $\mathcal{S}$ have disappeared. However, the model continues to track through this occlusion event and maintains the stable model on the visible portion of the subject. When the person emerges from behind the occluder, the appearance model rebuilds the dissipated stable model.

The ability to adapt to changing appearance is demonstrated in Fig. 4 [36]. Here, despite the person turning to walk in the opposite direction (at frame 300), the $\mathcal{S}$ component maintains a reasonable model for the stable image structure.

One of our goals was to track and identify stable properties in images of nonrigid objects, such as in the example shown in Fig. 5. From the images of m_s in Fig. 5a (bottom), notice that the mouth region was initially identified as stable, but after the person smiles, the stability is weakened significantly. Once the new expression has been held for about 20 frames, the structure is again identified as stable. Other parts of the face, such as the eyebrows, show similar behavior. Conversely, the values of m_s near the hairline and on nose continue to increase through these events, indicating that they are consistently stable and, overall, the head is being accurately tracked.

The behavior during a brief occlusion event is shown in Fig. 6, where the person's hand reaches up to brush his hair back. The model persists, with m_s and μ_s remaining essentially constant despite the occlusion. By contrast, notice how the data ownerships $o_{s,t}$ reflect the presence of the occluder. Also, note that the data ownerships are not perfect; there are some false matches to the appearance model in the area of the occluder. Presumably, these are a result of "accidental" alignments of the phase responses from the occluder with those of the appearance model. Given that the minimum standard deviation for σ_s is 0.1π (see Section 5), we should expect the false target rate to be reasonably high. In fact, these false targets appear to drag the model into misalignment during the occlusion (see the caption in Fig. 6 for a pointer to this), but that the appearance model is subsequently able to lock back on. Such a misalignment would persist in any two-frame tracking algorithm.

Fig. 7 shows the stability of the joint estimation of motion and appearance, despite significant changes in size and

Fig. 9. The ground truth points (*) are shown for frames 0, 361, 540, 700, 795, and 1,140. The approximations provided by the points fit using the $\mathcal{WSL}$ tracker (+) results, and those corresponding to the optimal similarity transforms (o), are also shown. Note the nonsimilarity deformations of the ground truth data.

lighting conditions. Even more challenging for the current method are the (at times) small target regions, and the small separation of the object and background motions (about a pixel per frame). Also, roughly half the target region is occluded by the bushes at times. The two runs in Fig. 7 are close to the limit of our approach in terms of these latter sources of difficulty.

7.1 Quantitative Performance

In order to quantiatively assess the accuracy of the tracking, we generated ground truth data by manually locating seven facial feature points (see Fig. 9) in each of the 1,145 frames in the Dudek face sequence [36]. We estimate that the RMS error in "mousing in" this ground truth data is 1.1 ± 0.1 pixel per frame for each facial feature point.[3] This error represents the minimum tracking error we can expect to resolve using this ground truth data.

In addition to inherent ground truth noise, we must also quantify errors in the ground truth data due to modeling assumptions. Specifically, the current tracker uses similarity transformations to model the object poses. This assumes that the positions of fixed object points in any given frame can be approximated by mapping points from a single canonical reference frame to corresponding points in each image frame using 2D similarity transforms. In the Dudek sequence, for example, this assumption is clearly violated by 3D rotations of the head and changes in facial expression. To compute the magnitude of the error due to this modeling assumption, we jointly estimate the optimal least squares positions for the seven image points in a fixed canoncial frame, along with a similarity transformation for each of the 1,145 frames of the sequence. The optimal RMS error between the ground truth points and the positions of the canonical points, was found to give a baseline error of 3.1 pixels per frame per feature point. This baseline error combines both the ground truth noise and the positional variations that cannot be explained by similarity transformations (see Fig. 9). Notice that this 3.1 pixel error must be a lower bound for the RMS error evaluated on the ground truth data for *any* tracker based on similarity transformations.

Remember that the $\mathcal{WSL}$ tracker is required to automatically identify and select stable image features for tracking. Moreover, it is free to slowly adapt these features, and to switch from one set of features to another. Therefore, the $\mathcal{WSL}$ tracker is unlikely to be tracking just the image data in the neighborhoods of the ground truth points. This puts the $\mathcal{WSL}$ tracker at a disadvantage relative to the baseline RMS error and, therefore, we should not expect our appearance-based tracker to achieve this baseline error. Nevertheless, we feel the baseline serves as a useful yardstick.

In order to compare the results from the $\mathcal{WSL}$ tracker with this baseline, we used the similarity transformations computed by the $\mathcal{WSL}$ tracker and located the positions of seven feature points in a canonical frame which, when mapped under these similarity transformations, provide the optimal RMS fit to the ground truth points.[4] The results for selected frames are shown in Fig. 9. The resulting tracking error over the entire sequence for the $\mathcal{WSL}$ tracker is 5.2 pixels per frame per feature, which compares well with the baseline error of 3.1.

3. This RMS error estimate was obtained by fitting a PCA model to the feature point positions and examining the decay rate of the residual RMS error with increasing model dimension.

4. This measure is similar, but not identical to, the standard deviation of the ground truth points when mapped to the canonical frame using the computed similarity transformations. The reason for the difference is due to the scale variations between the image data and the canonical frame. The distance in the original image is a more natural metric than distance in a canonical frame and, thus, write all errors in terms of distance in the original images.

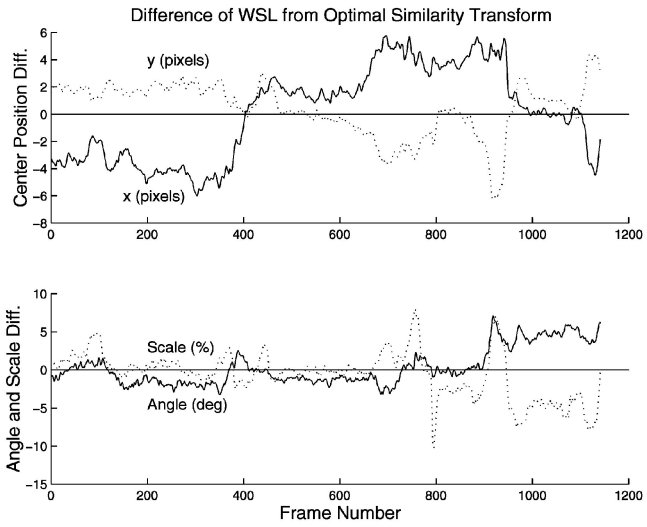

Fig. 10. The difference between the similarity transform parameters computed by the $\mathcal{WSL}$ tracker and those for the optimal fit to the moused in points.

The differences in the parameters of the similarity transformations computed by the $\mathcal{WSL}$ tracker and the optimal transforms for the seven ground truth points are displayed in Fig. 10. Despite the adaptive, open-loop nature of the $\mathcal{WSL}$ appearance model, the drift in the tracking appears to be well controlled. In particular, it is seen from Fig. 10 that, even after 1,000 frames, there is no clearly consistent bias in the center position, while the angle parameter appears to consistently differ by about 5 degrees, and the scale parameter differs by about 5 percent.

7.2 Failure Modes

In our tests, we have identified three failure modes [36]. The first is caused by acceptance of nontarget constraints as being consistent with the appearance model (see discussions of Figs. 3, 6, and 7). These erroneous constraints perturb the alignment of the model and, if this effect is sufficiently large, a tracking failure can occur.

Second, when the tracked object consistently moves with its background, then the appearance model also learns the background structure (as in Fig. 8). Tracking can fail if the object then moves independently. Failures of this sort can usually be avoided if one also includes a background appearance model, as in [17]. Unfortunately, simple background models are not always available, as in situations in which the depth structure of the background is complex. Therefore, in the experiments reported here, we chose to focus attention on the $\mathcal{WSL}$ appearance model and tracker alone.

Finally, tracking with this model of appearance will certainly fail in most situations in which the object becomes entirely occluded (see Fig. 11). This is caused by the $\mathcal{W}$ constraints identifying the coherent motion of the occluding object. In the absence of $\mathcal{S}$ constraints, the estimated motion will be determined by the $\mathcal{W}$ constraints that remain and, therefore, there is a tendency to begin tracking the occluder. An interesting topic for future work concerns the use of longer term memory and visual search that would allow one to recover from such tracking failures.

8 Conclusions

This paper proposes a robust, adaptive appearance model for motion-based tracking of complex natural objects. The model adapts to slowly changing appearance, and it maintains a natural measure of the *stability* of the observed image structure during tracking. By identifying stable properties of appearance, we can weight them more heavily for motion estimation, while less stable properties can be proportionately downweighted. The key contributions in this paper include:

1. the $\mathcal{WSL}$ mixture model that combines predictive density models of appearance with components that adapt over long and short time courses,
2. an online version of EM for learning the parameters of the $\mathcal{WSL}$ model,
3. an application in which we learn the time-varying phase behavior of a steerable pyramid, and
4. a tracking algorithm which exploits this appearance model to simultaneously estimate both motion and appearance.

Possible topics for future work include the incorporation of color and brightness data into the appearance model [7], and the use of the stable appearance model for global image matching to recover from tracking failures caused by total occlusion.

Appendix

Motion-Based Tracking Using $\mathcal{WSL}$

Here, we present in detail the process for fitting motion parameters $\mathbf{c}_t$ when the $\mathcal{WSL}$ appearance model is used to identify stable structure during tracking.

Our first task is to explain how approximate local minima are computed for the objective function $E(\delta_{\mathbf{c}_t})$ introduced in (17). Expanding the conditional distributions p_s, p_w, and p_o yields

$$E(\delta_{\mathbf{c}_t}) = \kappa + \sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \Bigg[\epsilon\, o_w(\hat{d}_{\mathbf{x},t})\, \frac{(d_c(\mathbf{x},t) - d(\mathbf{x},t-1))^2}{2\sigma_w^2} + o_s(\hat{d}_{\mathbf{x},t}) \frac{(d_c(\mathbf{x},t) - \mu_s(\mathbf{x},t-1))^2}{2\sigma_s(\mathbf{x},t-1)^2} \Bigg] + \frac{1}{2}(\mathbf{c}_t + \delta_{\mathbf{c}_t} - \vec{\xi})^T \mathbf{V}_1^{-1} (\mathbf{c}_t + \delta_{\mathbf{c}_t} - \vec{\xi}) + \frac{1}{2}(\mathbf{c}_t + \delta_{\mathbf{c}_t} - \mathbf{c}_{t-1})^T \mathbf{V}_2^{-1} (\mathbf{c}_t + \delta_{\mathbf{c}_t} - \mathbf{c}_{t-1}), \tag{18}$$

where κ is a constant independent of $\delta_{\mathbf{c}_t}$, $d_c(\mathbf{x},t) \equiv d(\mathbf{w}(\mathbf{x};\mathbf{c}_t + \delta_{\mathbf{c}_t}),t)$, and $\mu_s(\mathbf{x},t-1)$, $\sigma_s(\mathbf{x},t-1)$ are the mean and standard deviation estimates for the stable component of the appearance model at $(\mathbf{x},t-1)$. The terms on the second line in (18) arise from the logarithm of the prior (15). From (18) it is clear that, due to the nonlinear dependence of d_c on $\delta_{\mathbf{c}_t}$, the objective function $E(\delta_{\mathbf{c}_t})$ is not a quadratic function of $\delta_{\mathbf{c}_t}$.

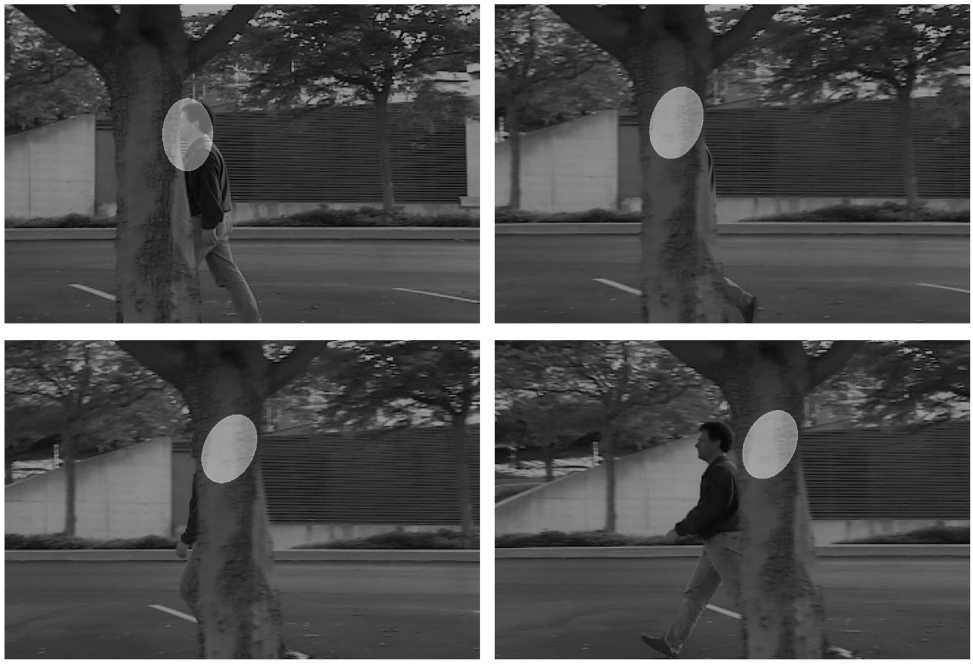

Fig. 11. Tracking failure (frames 635, 640, 645, and 650 of the sequence used in Fig. 4): When the person becomes fully occluded by the tree, tracking becomes impossible with the weak priors used here since there are no remaining $\mathcal{S}$ constraints, and the $\mathcal{W}$ constraints are consistent with the motion of the tree.

Instead of dealing directly with $E(\delta_{\mathbf{c}_t})$, we follow a standard practice in optical flow estimation by approximating this objective function with a quadratic functional in $\delta_{\mathbf{c}_t}$, namely, $\tilde{E}(\delta_{\mathbf{c}_t})$. To get this quadratic functional, we linearize the current observations about the current guess for the motion parameters $\mathbf{c}_t$. In particular, we approximate $d(\mathbf{w}(\mathbf{x};\mathbf{c}_t+\delta\mathbf{c}),t)$ by its first order Taylor series taken about $\delta_{\mathbf{c}_t}=0$. More formally,

$$\begin{aligned} d_c(\mathbf{x},t) &\equiv d(\mathbf{w}(\mathbf{x};\mathbf{c}_t+\delta\mathbf{c}),t) \\ &= d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)+\nabla d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)\,W\,\delta_{\mathbf{c}_t}+O(||\delta_{\mathbf{c}_t}||^2), \end{aligned} \tag{19}$$

where $\nabla d(\mathbf{x},t)\equiv(d_x(\mathbf{x},t),d_y(\mathbf{x},t))$ denotes the spatial partial derivatives of the data observations, and where $W=\partial\mathbf{w}/\partial\mathbf{c}_t$ denotes the 2×4 Jacobian of the warp map at $\mathbf{c}_t$. Using (19), we can rewrite the nonlinear differences involving $d_c(\mathbf{x},t)$ as follows:

$$\begin{aligned} d_c(\mathbf{x},t)-d(\mathbf{x},t-1) &= \delta d_w(\mathbf{x},t)+\nabla d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)\,W\,\delta_{\mathbf{c}_t}+ \\ &\quad O(||\delta_{\mathbf{c}_t}||^2), \\ d_c(\mathbf{x},t)-\mu_s(\mathbf{x},t-1) &= \delta d_s(\mathbf{x},t)+\nabla d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)\,W\,\delta_{\mathbf{c}_t}+ \\ &\quad O(||\delta_{\mathbf{c}_t}||^2), \end{aligned}$$

where $\delta d_w(\mathbf{x},t)=d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)-d(\mathbf{x},t-1)$ and $\delta d_s(\mathbf{x},t)=d(\mathbf{w}(\mathbf{x};\mathbf{c}_t),t)-\mu_s(\mathbf{x},t-1)$ are the temporal differences between the data and the model at the corresponding locations dictated by the current guess for the warp parameters $\mathbf{c}_t$.

Finally, substituting these expressions into (18) and isolating second order terms in $\delta_{\mathbf{c}_t}$, we find

$$E(\delta_{\mathbf{c}_t})=\tilde{E}(\delta_{\mathbf{c}_t})+O(||\delta_{\mathbf{c}_t}||^2), \tag{20}$$

where

$$\begin{aligned} \tilde{E}(\delta\mathbf{c};\,\mathbf{c}_t) =& \kappa+\sum_{\mathbf{x}\in\mathcal{N}_{t-1}}\Bigg[\epsilon\,\frac{o_w(\hat{\boldsymbol{d}}_{\mathbf{x},t})}{2\sigma_w^2}[\delta d_w+\nabla d^T\,W\,\delta_{\mathbf{c}_t}]^2+ \\ &\frac{o_s(\hat{\boldsymbol{d}}_{\mathbf{x},t})}{2\sigma_s^2}[\delta d_s+\nabla d^T\,W\,\delta_{\mathbf{c}_t}]^2\Bigg]+ \\ &\frac{1}{2}(\mathbf{c}_t+\delta_{\mathbf{c}_t}-\vec{\xi})^T\mathbf{V}_1^{-1}(\mathbf{c}_t+\delta_{\mathbf{c}_t}-\vec{\xi})+ \\ &\frac{1}{2}(\mathbf{c}_t+\delta_{\mathbf{c}_t}-\mathbf{c}_{t-1})^T\mathbf{V}_2^{-1}(\mathbf{c}_t+\delta_{\mathbf{c}_t}-\mathbf{c}_{t-1}). \end{aligned} \tag{21}$$

Notice that $\tilde{E}(\delta_{\mathbf{c}_t})$ is a quadratic function of $\delta_{\mathbf{c}_t}$. Moreover, notice that $\tilde{E}$ provides a second order approximation of E in the neighborhood of $\delta_{\mathbf{c}_t}=0$, so $\tilde{E}(0)=E(0)$ and $\nabla\tilde{E}(0)=\nabla E(0)$. Therefore, $\delta_{\mathbf{c}_t}=0$ is a local minimum of $E(\delta_{\mathbf{c}_t})$ if and only if it is also a local minimum of the quadratic objective function $\tilde{E}(\delta_{\mathbf{c}_t})$.

Given this approximation property of $\tilde{E}(\delta_{\mathbf{c}_t})$, we can safely adopt the strategy of computing $\delta\mathbf{c}_t$ to optimize the quadratic functional $\tilde{E}(\delta_{\mathbf{c}_t})$. This update will be zero if and only if the orginal objective function has a local extrema at $\delta_{\mathbf{c}_t}$. Otherwise, the computed update provides a convenient approximation for a value that minimizes the nonquadratic objective function $E(\delta_{\mathbf{c}_t})$.

The optimization of the quadratic functional $\tilde{E}(\delta_{\mathbf{c}_t})$ leads to the linear system for the update $\delta_{\mathbf{c}_t}$,

$$(A_s+\epsilon A_w+A_p)\,\delta\mathbf{c}_t=\mathbf{b}_s+\epsilon\mathbf{b}_w+\mathbf{b}_p, \tag{22}$$

where each A_i is a 4×4 matrix and each $\mathbf{b}_i$ is a 4-vector, for $i=w,s,p$:

$$A_w = \sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \frac{o_w(\hat{\mathbf{d}}_{\mathbf{x},t})}{\sigma_w^2} W^T \nabla d \nabla d^T W, \quad \mathbf{b}_w = -\sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \frac{o_w(\hat{\mathbf{d}}_{\mathbf{x},t})}{\sigma_w^2}\, \delta d_w\, W \nabla d$$

$$A_s = \sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \frac{o_s(\hat{\mathbf{d}}_{\mathbf{x},t})}{\sigma_s^2} W^T \nabla d \nabla d^T W, \quad \mathbf{b}_s = -\sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \frac{o_s(\hat{\mathbf{d}}_{\mathbf{x},t})}{\sigma_s^2}\, \delta d_s\, W \nabla d$$

$$A_p = \mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}, \quad \mathbf{b}_p = -\mathbf{V}_1^{-1}(\mathbf{c}_t - \vec{\boldsymbol{\xi}}) - \mathbf{V}_2^{-1}(\mathbf{c}_t - \mathbf{c}_{t-1}).$$

Each term in these sums over $\mathbf{x}$ are formed from a different motion constraint, weighted by the ownership probabilities for the $\mathcal{W}$ and $\mathcal{S}$ processes, respectively.

Notice that the usefulness of the approximate objective function $\tilde{E}(\delta_{\mathbf{c}_t})$ depends on the range over which the linear approximation in (19) is a good approximation for the data. The fact that the bandpass phase is expected to be roughly linear over a signficant range [8] provides another motivation for using phase properties. In addition, as mentioned in Section 5, we use the method described by [9] to detect probable outliers in the phase-based warp constraints due to phase instability. That is, when a local phase observation, $d_{\mathbf{x},t}$, is deemed unstable, the corresponding gradient constraints are undefined and not included in (22). When an unstable observation at time $t-1$ maps to a good observation at time t under the current warp, then the likelihood $p_w(\hat{\mathbf{d}}_{\mathbf{x},t} \mid d_{\mathbf{x},t-1})$ is undefined. Instead, we use $p_w = 0.05$ when the previous observation was deemed unstable. We also remove the corresponding $\mathcal{W}$ constraints from the linear system (22).

In practice, as with most nonconvex optimizations to estimate mixture model parameters, we find it useful to apply this fitting procedure within a coarse-to-fine strategy, and use deterministic annealing in estimating the motion parameters to help avoid becoming stuck in local minima (e.g., see [16], [19]). The initial guess for the warp parameters is based on a simple, constant velocity model, so the initial guess is simply equal to the estimated warp parameters from the previous frame. By way of annealing, instead of using the variances $\sigma_{s,t}^2$ and σ_w^2 in computing the ownerships and gradients (23) for the $\mathcal{S}$ and the $\mathcal{W}$ processes, we use parameters $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{W}}$. After each iteration in which (22) is solved, these values are decreased according to the schedule

$$\sigma_{\mathcal{S}} \leftarrow \min(0.95\sigma_{\mathcal{S}}, \hat{\sigma}_{\mathcal{S}})$$
$$\sigma_{\mathcal{W}} \leftarrow \min(0.95\sigma_{\mathcal{W}}, \hat{\sigma}_{\mathcal{W}}),$$

where $\hat{\sigma}_{\mathcal{S}}$ and $\hat{\sigma}_{\mathcal{S}}$ are the maximum likelihood variance estimates of the $\mathcal{S}$-process and $\mathcal{W}$-process phase differences, over the entire neighborhood, $\mathcal{N}_t$, given the motion estimate obtained in the current iteration. Once the variances reach a minimal value the annealling is turned off and they are allowed to fluctuate according to the current motion parameters. Moreover, as the variance of the $\mathcal{S}$ process decreases according to the spatial ensemble of data observations at each iteration, the variances used for each individual observation in computing ownerships and likelihood gradients are never allowed to be lower than the corresponding variance of $\sigma_{s,t}^2$.

ACKNOWLEDGMENTS

The authors would like to thank the Xerox Foundation and the Palo Alto Research Center for their financial support.

REFERENCES

[1] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 232-237, 1998.
[2] M.J. Black and A.D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation," *Int'l J. Computer Vision*, vol. 26, no. 1, pp. 63-84, 1998.
[3] T. Cham and J.M. Rehg, "A Multiple Hypothesis Approach to Figure Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239-245, 1998.
[4] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects Using Mean Shift," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 142-149, 2000.
[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. Series B*, vol. 39, pp. 1-38, 1977.
[6] G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models," *Proc. European Conf. Computer Vision*, pp. 581-595, 1998.
[7] T.F. El-Maraghi, "Robust On-Line Appearance Models for Visual Tracking," PhD thesis, Dept. of Computer Science, Univ. of Toronto, 2002.
[8] D.J. Fleet, *Measurement of Image Velocity*. Norwell, Mass.: Kluwer, 1992.
[9] D.J. Fleet and A.D. Jepson, "Stability of Phase Information," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1253-1268, Dec. 1993.
[10] D.J. Fleet, A.D. Jepson, and M. Jenkin, "Phase-Based Disparity Measurement," *Computer Vision and Image Understanding*, vol. 53, no. 2, pp. 198-210, 1991.
[11] W. Freeman and E.H. Adelson, "The Design and Use of Steerable Filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 891-906, 1991.
[12] B. Frey, "Filling in Scenes by Propagating Probabilities through Layers into Appearance Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 185-192, 2000.
[13] G.D. Hager and P.N. Belhumeur, "Efficient Region Tracking with Parametric Models of Geometry and Illumination," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025-1039, Oct. 1998.
[14] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions," *Int'l J. Computer Vision*, vol. 12, no. 1, pp. 5-16, 1994.
[15] M. Isard and A. Blake, "Condensation—Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision*, vol. 29, no. 1, pp. 2-28, 1998.
[16] A. Jepson and M.J. Black, "Mixture Models for Optical Flow Computation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 760-761, 1993.
[17] A.D. Jepson, D.J. Fleet, and M.J. Black, "A Layered Motion Representation with Occlusion and Compact Spatial Support," *Proc. European Conf. Computer Vision*, vol. 1, pp. 692-706, 2002.
[18] N. Jojic and B.J. Frey, "Learning Flexible Sprites in Video Layers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 199-206, 2001.
[19] S.X. Ju, M.J. Black, and A.D. Jepson, "Skin and Bones: Multi-Layer, Locally Affine, Optical Flow and Regularization with Transparency," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 307-314, 1996.
[20] D. Koller, K. Daniilidis, T. Thorhallson, and H.-H. Nagel, "Model-Based Object Tracking in Traffic Scenes," *Proc. European Conf. Computer Vision*, pp. 437-452, 1992.
[21] D. Kriegman, *Personal Comm.* 2002.
[22] F. Leymarie and M. Levine, "Tracking Deformable Objects in the Plane Using an Active Contour Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 617-634, 1993.
[23] F.G. Meyer and P. Bouthemy, "Region-Based Tracking Using Affine Motion Models in Long Image Sequences," *Computer Vision, Graphics, and Image Processing: Image Understanding*, vol. 60, no. 2, pp. 119-140, 1994.
[24] D. Morris and J. Rehg, "Singularity Analysis for Articulated Object Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 289-296, 1998.
[25] C. Olson, "Maximum-Likelihood Template Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 52-57, 2000.

[26] N. Paragios and R. Deriche, "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 266-280, 2000.
[27] W. Rucklidge, "Efficient Guaranteed Search for Gray-Level Patterns," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 717-723, 1997.
[28] J. Shi and C. Tomasi, "Good Features to Track," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 593-600, 1994.
[29] H. Sidenbladh, M.J. Black, and D.J. Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion," *Proc. European Conf. Computer Vision,* vol. 2, pp. 702-718, 2000.
[30] E.P. Simoncelli and W.T. Freeman, "The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation," *Proc. IEEE Int'l Conf. Image Processing,* pp. 444-447, 1995.
[31] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, "Shiftable Multi-Scale Transforms," *IEEE Trans. Information Theory,* vol. 38, no. 2, pp. 587-607, 1992.
[32] C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 246-252, 1999.
[33] H. Tao, H.S. Sawhney, and R. Kumar, "Dynamic Layer Representation with Applications to Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 134-141, 2000.
[34] Y. Weiss and E. Adelson, "Slow and Smooth," *MIT AI Memo 1624,* 1996.
[35] Y. Weiss and D.J. Fleet, "Velocity Likelihoods in Biological and Machine Vision," *Probabilistic Models of the Brain: Perception and Neural Function,* R.P.N. Rao, B.A. Olshausen, and M.S. Lewicki, eds., pp. 81-100, Cambridge: MIT Press, 2001.
[36] www.cs.toronto.edu/vis/projects/adaptiveAppearance.html, 2003.

Allan D. Jepson received the BSc degree in mathematics from the University of British Columbia in 1976, and the PhD degree in applied mathematics from the California Institute of Technology in 1980. He spent two years as a postdoctoral fellow at Stanford University in the Mathematics Department, and then joined the faculty of the Department of Computer Science at the University of Toronto in 1982. From 1989 to 1995, he was a scholar of the Canadian Institute of Advanced Research. His current research interests include image motion estimation, image understanding, and perceptual inference. He is a member of the IEEE Computer Society.

David J. Fleet received the PhD degree in computer science from the University of Toronto in 1991. From 1991 to 2000, he was a faculty member at Queen's University in the Departments of Computer Science, Psychology, and Electrical Engineering. In 1999, he joined the Palo Alto Research Center (PARC), where he currently manages the Digital Video Analysis Group and the Perceptual Document Analysis Group. His research interests include computer vision, image processing, visual perception, and visual neuroscience. He has published several research articles and one book on various topics including the estimation of optical flow and stereoscopic disparity, probabilistic methods in motion analysis, modeling appearance in image sequences, as well as the motion perception and human stereopsis. In 1996, Dr. Fleet was awarded an Alfred P. Sloan Research Fellowship. He has won paper awards at ICCV 1999 and at CVPR 2001. He was program cochair for the 2003 IEEE Conference on Computer Vision and Pattern Recognition, and he currently serves as an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is a member of the IEEE Computer Society.

Thomas F. El-Maraghi received the BSc degree in electrical and computer engineering, with first class honors, from Queen's University in Kingston, Ontario, in 1994. He remained at Queen's University for the next two years and received the MSc degree in computer science in 1996. He studied computer vision at the University of Toronto from 1996 to 2003, where he received the PhD degree in computer science. His research interests are in computer vision, with a particular emphasis on visual tracking and segmentation. He is a member of the IEEE Computer Society.