*IEEE Conference on Computer Vision and and Pattern Recognition, Kauai, 2001, Vol. I, pp. 415–422*

# Robust Online Appearance Models for Visual Tracking

**Allan D. Jepson*** **David J. Fleet†** **Thomas F. El-Maraghi*†**

* Department of Computer Science, University of Toronto, Toronto, M5S 1A4
† Xerox Palo Alto Research Center, 3333 Coyote Hill Rd, Palo Alto, CA 94304

## Abstract

*We propose a framework for learning robust, adaptive, appearance models to be used for motion-based tracking of natural objects. The approach involves a mixture of stable image structure, learned over long time courses, along with 2-frame motion information and an outlier process. An on-line 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. It is also provides the ability to adapt to natural changes in appearance, such as those due to facial expressions or variations in 3D pose. We show experimental results on a variety of natural image sequences of people moving within cluttered environments.*

## 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 unstable 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 2-frame transient component, and an outlier process. The stable model identifies the most reliable structure for motion estimation, while the two-frame constraints provide additional information when the appearance model is being initialized or provides relatively few constraints. The parameters of the model are learned efficiently with an on-line version of the EM algorithm.

The appearance model and the tracker can be used with different types of image properties. Here we consider a class of models that express image appearance in terms of the complex-valued coefficients of a steerable pyramid. This wavelet-based model allows for the stability at different scales or in different spatial neighborhoods to be assessed independently.

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 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 appearance models include templates [8, 13, 14, 15], view-based subspace models [2, 9], the most recent frame in 2-frame flow estimation [16, 17], temporally filtered, motion-compensated images [10, 18, 20], and global statistics [1, 3].

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. Reliability can be improved with the use of subspace models of appearance [2, 9], but these are object specific and often require training prior to tracking. Frey [8] proposed a tracker with image templates that model the mean and the variance of each pixel during tracking. The method we propose below bears some similarity to this; however, we use wavelets, online learning, and a robust mixture model instead of a Gaussian density at each pixel.

The use of global statistics, such as color histograms have been popular for tracking [1, 3], but they will not accurately register the model to the image in many cases. These methods also fail to accurately track regions that share similar statistics with nearby regions.

Figure 1. Cropped images from a 1200 frame sequence taken with handheld video camera. The ellipse shows the region in which the motion and appearance are estimated.

Motion-based trackers integrate motion estimates through time. With 2-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 by accumulating an appearance model through time. Indeed, optimal motion estimation can be formulated as the estimation of both motion and appearance simultaneously [20]. For example, one could filter the stabilised images with linear IIR filters [10, 18]. But linear filtering does not provide robustness to outliers or measures of stability.

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 structure and naturally combines both stable structure and transient image information; 2) an on-line version of EM for learning model parameters; and 3) a tracking algorithm which simultaneously 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 2-frame motion information and stable appearance that is learned over many frames.

## 3 $\mathcal{WSL}$ Appearance Model Framework

We first introduce the model with a single real-valued data observation, say $d_t$ at each frame $t$. As a motivational example, consider tracking a region, such as the face in Fig. 1 (see also [21]), 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 glasses are removed.

These phenomena motivate the components of our appearance model. The first component is the stable model, $\mathcal{S}$, which is intended to capture the behaviour of temporally stable image observations when and where they occur. In particular, given that the stable component generated the observation $d_t$, we model the probability density for $d_t$ by the Gaussian density $p_s(d_t \mid \mu_{s,t}, \sigma^2_{s,t})$. Here $\mu_{s,t}$ and $\sigma^2_{s,t}$ are piecewise, slowly varying functions 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, or occlusion. 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 Figure 2(top) provides an idealized example of these generative processes. The smooth (dashed blue) curve represents the piecewise slowly varying appearance signal. The observed data (red) 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 Figure 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 model is motivated by the the desire to integrate the appearance model with an image-based tracking algorithm. That is, for a selected image region we wish to learn a model for the dominant stable image structure within the region and to simultaneously track it. This is difficult because we do not expect to have an initial stable appearance model, nor a good idea for how the object moves. The third component, called the wandering model $\mathcal{W}$, determines what should be tracked in such a situation. In effect, this wandering component permits the tracker described in Section 6 to gracefully degrade to a 2-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. As such, we choose the probability density for $d_t$, given that it is generated by $\mathcal{W}$, to be the Gaussian density $p_w(d_t \mid d_{t-1})$. Here the mean is simply the observation from the previous frame, $d_{t-1}$, and the variance is fixed at $\sigma^2_w$.

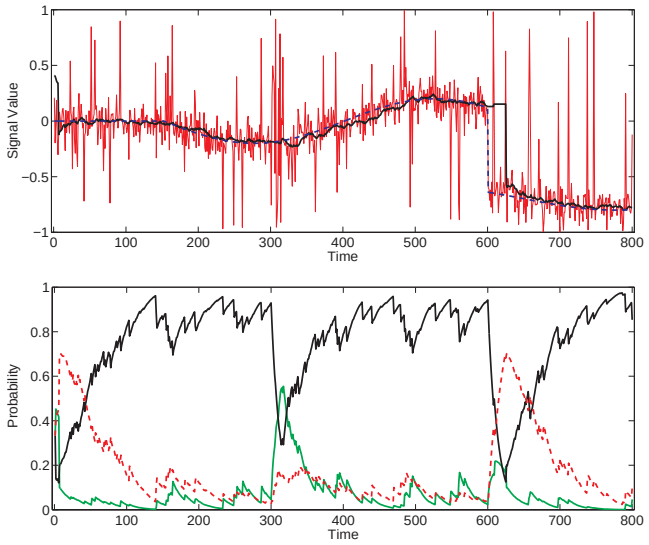


Figure 2. Estimation using on-line EM. (top) The original data (thin red) with true state (dashed blue) and the estimated mean of the stable process (thick black). The noise is a mixture of Gaussian and uniform densities, with mixing probabilities (0.9, 0.1), except for 15 frames at 300 which are pure outliers. (bottom) Mixing probabilities for $\mathcal{S}$ (black), $\mathcal{W}$ (dashed red), and the $\mathcal{L}$ (light green).

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

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

where $\mathbf{m} = (m_w, m_s, m_l)$ are the mixing probabilities, and $\mathbf{q}_t = (\mu_{s,t}, \sigma^2_{s,t})$ contains the mean and variance parameters of the stable component of the model.

## 4 Parameter Estimation with On-line EM

Our goal is to estimate the parameters of the generative model in (1), namely, the mean and variance of the prediction of the data, $d_t$, by the stable process, $\mathbf{q} = (\mu_s, \sigma_s^2)$, and the mixing probabilities $\mathbf{m} = (m_w, m_s, m_l)$. Moreover, since we plan to apply the estimation scheme to filter responses, we seek a simple computational algorithm which requires a small amount of memory for each observation.

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 density in (1):

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

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 [4] for estimating $\mathbf{m}_t$ and $\mathbf{q}_t$ that assumes they are constant under the temporal window. The form of these EM-updates provides the basis for our on-line method.

Given a current guess for the state variables $\mathbf{m}_t$ and $\mathbf{q}_t$ (constant over the temporal window), the E-step provides the ownership probabilities for each observation $d_k$:

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

for $i \in \{w, s, l\}$ (see [4]). Conditioned on these ownerships, the M-step then computes new maximum likelihood estimates for the parameters $\mathbf{m}_t$ and $\mathbf{q}_t$. First, the updated mixture probabilities, $\mathbf{m}_t$, are given by

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

for $i \in \{w, s, l\}$ (we have reused the notation $m_{i,t}$ to denote the updated values). Here $K$ is a normalization constant to ensure that the mixing probabilities sum to one. Similarly, the M-step for the mean and variance are

$$\mu_{s,t} = \frac{M_{1,t}}{m_{s,t}}\,, \quad \sigma^2_{s,t} = \frac{M_{2,t}}{m_{s,t}} - \mu^2_{s,t}\,, \quad (5)$$

where $M_{j,t}$ are the ownership weighted first- and second-order moments,

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

for $j = 1, 2$. The standard EM-algorithm then consists of iterating the steps outlined in equations (3) – (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 on-line approach. Instead we adopt an approximation to (3) – (6). To this end, we first exploit a recursive expression for the exponential support $S_t(k)$ to obtain,

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

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

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

We also approximate the mixing probabilities the same way:

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

for $i \in \{s, w, l\}$. One further deviation from these equations is used to avoid singular situations; i.e., we impose a non-zero lower bound on the mixing probabilities and $\sigma_{s,t}$.

In this approximation to the batch EM in (3) – (6), as mentioned above, we do not update the data ownerships of the past observations. Therefore when the model parameters change rapidly this on-line 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 $\mathcal{S}$ in any case. Conversely, when the mean and variance drift slowly, the on-line approximation is typically very good (see Fig. 2).

Given sudden changes in appearance, or unstable data, the $\mathcal{S}$ process often loses track of the mean, and is given a small mixing probability (see Fig. 2). Thus it is necessary to occasionally restart the appearance model. Here we restart the model whenever the $m_{s,t}$ falls below a fixed threshold (we used 0.1). This is done by simply resetting the values of all state variables. The new values for the mixing probabilities $m_{i,t}$ are 0.4, 0.15, and 0.45 for $i = w, s, l$. The smaller value for $m_{s,t}$ reflects an initial uncertainty for the $\mathcal{S}$ model. The new values for the moments $M_{j,t}$ for $j = 1, 2$ are taken to be $d_t m_{s,t}$ and $\sigma_{s,0}^2 m_{s,t}$, respectively. In effect this restarts 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 $\sigma_{s,0} = \sigma_w/1.5$. These same values are used for initialization in the first frame.

Figure 2 illustrates the EM procedure on our 1D example with half-life $n_s = 8$. Initially the $\mathcal{W}$ model owns most of the data until the stable $\mathcal{S}$ model gains confidence. During the outlier burst at frame 300 the outlier $\mathcal{L}$ model owns a greater share of the data. At the jump at frame 600 the $\mathcal{S}$ component is a poor predictor for the data, and its mixing probability drops quickly. Accordingly, because the $\mathcal{W}$ component can explain the data, its mixing probability increases. At frame 625 the $\mathcal{S}$ model mixing probability drops sufficiently low that the procedure restarts, after which the $\mathcal{S}$ model locks back onto the true state.

## 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. In this work, we applied the on-line EM procedure to responses of a steerable pyramid (based on the $G_2$ and $H_2$ filters of [7]). Steerable pyramids provide a description of the image at different scales and orientations which is useful for coarse-to-fine differential motion estimation, and for isolating stability at different scales. Here we use $G_2$ and $H_2$ filters at two scales, tuned to wavelengths of 8 and 16 pixels (subsampled by factors of 2 and 4), with 4 orientations at each scale.

From the filter outputs, we chose to maintain a representation of the phase structure as our appearance model. This gives us a natural degree of amplitude and illumination independence, and it provides the fidelity for accurate image alignment afforded by phase-based methods [5, 6]. Phase responses associated with small filter amplitudes, or those deemed unstable according to the technique described in [5], were treated as outliers.

In what follows, 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 from all filters at time $t$ in the region. Let $\mathcal{A}_t = \{(\mathbf{m}(\mathbf{x}, t),\, \mathbf{q}(\mathbf{x}, t)\}_{\mathbf{x}\in\mathcal{N}_t}$ denote the entire appearance model of the phase at each orientation, scale, and spatial location in $\mathcal{N}_t$. The half-life of the exponential temporal support, $S_t(k)$, was set to $n_s = 20$ frames. The other parameters of the on-line EM estimator are: 1) the outlier probability, which is uniform on $[-\pi,\, \pi)$; 2) the standard deviation of the $\mathcal{W}$ process on phase differences, which we take to be mean-zero Gaussian with $\sigma_w = 0.35\pi$; and 3) the minimum standard deviation of the stable process, $\sigma_{s,0} = 0.1\pi$. These latter parameters are specific to the use of phase.

## 6 Motion-Based Tracking

We demonstrate the behaviour of the adaptive, phase-based appearance model in the context of tracking nonrigid objects. For this demonstration we manually specify an elliptical region $\mathcal{N}_0$ at time 0. The tracking algorithm then estimates the image motion and the appearance model as it tracks the dominant image structure in $\mathcal{N}_t$ over time.

The motion is represented in terms of frame-to-frame parameterized image warps. In particular, given the warp parameters $\mathbf{a}_t$, a pixel $\mathbf{x}$ at frame $t-1$ corresponds to the image location $\mathbf{x}_t = \mathbf{w}(\mathbf{x}; \mathbf{a}_t)$ at time $t$, where $\mathbf{w}(\mathbf{x}; \mathbf{a}_t)$ is the warp function. We use similarity transforms here, so $\mathbf{a}_t = (\mathbf{u}_t, \theta_t, \rho_t)$ is a 4-vector describing translation, rotation, and scale changes, respectively. Given the parameter vector $\mathbf{a}_t$, $\mathcal{N}_t$ is just the elliptical region provided by warping $\mathcal{N}_{t-1}$ by $\mathbf{w}(\mathbf{x}; \mathbf{a}_t)$. Other parameterized image warps and other forms of image regions could also be used.

To find an optimal warp we (locally) maximize the sum of the data log likelihood and a log prior that provides a preference for slow and smooth motions (cf. [19]). In terms of the motion and appearance models outlined above, the data log-likelihood can be expressed as

$$L(\{d(\mathbf{w}(\mathbf{x}; \mathbf{a}_t), t)\} \,|\, \mathcal{A}_{t-1}) \;= \qquad (10)$$
$$\sum_{\mathbf{x}\in\mathcal{N}_{t-1}} \log p(d(\mathbf{w}(\mathbf{x}; \mathbf{a}_t), t) \,|\, \mathbf{m}_{\mathbf{x},t-1}, \mathbf{q}_{\mathbf{x},t-1}, d(\mathbf{x}, t-1))\;.$$

Intuitively, this can be understood as follows: data at the current frame $t$ is warped back to the coordinates of frame $t-1$ according to the parameters $\mathbf{a}_t$. The log likelihood of this warped data $\{d(\mathbf{w}(\mathbf{x}; \mathbf{a}_t), t)\}$ is then computed with respect to the appearance model $\mathcal{A}_{t-1}$.

Figure 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 $o_s(\mathbf{x}, t)$. The rows correspond to frames 244, 259, 274, and 289, top to bottom. Note the model persistence and the drop in data ownership within the occluded region.

The prior is introduced mainly to cope with occlusions, and to exploit the persistence of the stable component $\mathcal{S}$. We take the prior density over the motion parameters $\mathbf{a}_t = (\mathbf{u}_t, \theta_t, \rho_t)$, conditioned on the motion at time $t-1$, $\hat{\mathbf{a}}_{t-1}$, to be a product of two 4D Gaussians:

$$p(\mathbf{a}_t \mid \hat{\mathbf{a}}_{t-1}) \;=\; G(\mathbf{a}_t;\, \vec{\eta},\, \mathbf{C}_1)\, G(\mathbf{a}_t;\, \hat{\mathbf{a}}_{t-1},\, \mathbf{C}_2)\,. \qquad (11)$$

The first Gaussian prefers slow motions, with mean $\vec{\eta} \equiv (0, 0, 0, 1)$ and covariance $\mathbf{C}_1 \equiv \mathrm{diag}(8^2, 8^2, 0.05^2, 0.01^2)$. Here, translations are measured in pixels, rotational velocities in radians, and the scale parameter is a multiplicative factor so $\vec{\eta}$ specifies the identity warp. The second Gaussian prefers slow changes in motion, with $\mathbf{C}_2 \equiv \mathrm{diag}(1, 1, 0.02^2, 0.01^2)$.

In order to estimate $\mathbf{a}_t$ we can almost directly apply the EM-algorithm described in [11]. We omit the details due to space limitations, and instead just sketch the E-step and M-step. The E-step determines the ownership probabilities for the backwards warped data $\{d(\mathbf{w}(\mathbf{x}; \mathbf{a}_t), t)\}$, as in (3) above. The M-step uses these ownerships to form a linear system for the update $\delta\mathbf{a}_t$:

$$(A_s + \epsilon A_w)\delta\mathbf{a}_t \;=\; \mathbf{b}_s + \epsilon\mathbf{b}_w\,. \qquad (12)$$

Here, $A_i$ is a $4 \times 4$ matrix and $\mathbf{b}_i$ a 4-vector, for $i = w, s$. These quantities are formed from the motion constraints weighted by the ownership probabilities for the $\mathcal{W}$ and $\mathcal{S}$ processes, respectively (see [11]). Also, $\epsilon$ is a weighting factor for the wandering constraints. A proper M-step for maximizing the likelihood in (10) would use the weight $\epsilon = 1$. We have found it useful to downweight the constraints owned by the wandering model by a factor of $\epsilon = 1/n_s$, where $n_s$ is the half-life of the exponential temporal window used in the appearance model. We use coarse-to-fine matching and deterministic annealing (see [11], [12]) in fitting the warp parameters.

Once the warp parameters $\mathbf{a}_t$ have been determined, we convect the appearance model $\mathcal{A}_{t-1}$ forward to the current time $t$ using the warp specified by $\mathbf{a}_t$. To perform this warp

Figure 4. The adaptation of the model during tracking. (top) The target region in selected frames 200, 300, 480. (bottom) The stable component's mixing probability (left) and mean (right) for the selected frames.

Figure 5. Adaptation to changes of expression. (top) The target region in selected frames 420, 455, 490. (bottom) The stable component's mixing probability (left) and mean (right) for the selected frames (time increases left to right in each set). Note how the regions around the mouth and eyebrows adapt, while others remain stable.

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 [6].

## 7 Experiments

The behaviour 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 $m_s(x, t)$ 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 one shown on the top row [21]. The significant responses for $m_{s,t}$ and $o_{s,t}$ demonstrate that the appearance model successfully identified stable structure, typically inside the object boundary. On the second and third rows of Fig. 3, where the person is 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 $m_s$ falls below the plotting threshold. This clearly demonstrates the persistence of the appearance model. The third row depicts the model after

Figure 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).

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 [21]. 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. 5(bottom left), 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 behaviour. 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 behaviour during a brief occlusion event is shown in Fig. 6, where the person's hand reaches up to brush their hair back. The model persists, with $m_s$ and $\mu_s$ remaining essentially constant despite the occlusion. By contrast, notice that the data ownerships $o_{s,t}$ clearly 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 $\sigma_s$ is $0.1\pi$, 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


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

a pointer to this), but that the appearance model is subsequently able to lock back on. Such a misalignment would clearly persist in any 2-frame tracking algorithm.

Finally, Fig. 7 shows the stability of the joint estimation of motion and appearance, despite significant changes in size and lighting conditions. Even more challenging for the current method are the (at times) small target regions, and the small separation of the object motion from the background motion (about a pixel per frame). Also, roughly half the target region is occluded by the bushes at times. The two runs depicted in Fig. 7 are close to the limit of our approach in terms of these latter sources of difficulty.

In our tests we have identified two failure modes [21]. The first is caused by acceptance of non-target 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. Tracking can fail if the object then moves independently.

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

## References

[1] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. *Proc. CVPR*, 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. *IJCV*, 26(1):63–84, 1998

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Proc. IEEE CVPR*, v. 2, pp. 142–149, Hilton Head, 2000

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B*, 39:1–38, 1977.

[5] D.J. Fleet and A.D. Jepson. Stability of phase information. *IEEE Trans. PAMI*, 15(12):1253–1268, 1993

[6] D.J. Fleet, A.D. Jepson, and M. Jenkin. Phase-based disparity measurement. *CVIU*, 53(2):198–210, 1991

[7] W. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. PAMI*, 13:891–906, 1991

[8] B. Frey. Filling in scenes by propagating probabilities through layers into appearance models. *Proc. IEEE CVPR*, v. I, pp. 185–192, Hilton Head, 2000

[9] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. PAMI*, 27(10):1025–1039, 1998

[10] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12(1):5–16, 1994

[11] A. Jepson and M. J. Black. Mixture models for optical flow computation. *Proc. IEEE CVPR*, pp. 760–761, 1993

[12] 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 CVPR*, pp. 307–314, 1996

[13] D. Morris and J. Rehg. Singularity analysis for articulated object tracking. *Proc IEEE CVPR*, pp. 289–296, 1998

[14] C. Olson. Maximum-likelihood template tracking. *Proc IEEE CVPR*, v. II, pp. 52–57, Hilton Head, 2000

[15] W. Rucklidge. Efficient guaranteed search for gray-level patterns. *Proc CVPR*, pp. 717–723, Puerto Rico, 1997

[16] J. Shi and C. Tomasi. Good features to track. *Proc IEEE CVPR*, pp. 593–600, 1994

[17] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *Proc. ECCV*, v. II, pp. 702–718. Springer-Verlag, Dublin, 2000

[18] H. Tao, H.S. Sawhney, and R. Kumar. Dynamic layer representation with applications to tracking. *Proc. IEEE CVPR*, v. 2, pp. 134–141, Hilton Head, 2000

[19] Y. Weiss and E. Adelson. Slow and smooth. *MIT AI Memo 1624*, 1998

[20] Y. Weiss and D.J. Fleet. Velocity likelihoods in biological and machine vision. In Rao et al (eds), *Probabilistic Models of the Brain: Perception and Neural Function*, pp. 81–100, Cambridge, 2001. MIT Press

[21] See www.cs.toronto.edu/vis/projects/adaptiveAppearance.html for mpeg videos of tracking results.